АКТУАЛЬНЫЕ ТЕНДЕНЦИИ И РЕЗУЛЬТАТЫ В ОБЛАСТИ КОМПЬЮТЕРНОГО ЗРЕНИЯ, ГЕНЕРАЦИИ ДАННЫХ И ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ (2020-2024)

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН, директор по направлению – руководитель научного комплекса «Искусственный интеллект и техническое зрение» ФАУ «ГосНИИАС», *viz@gosniias.ru*

Расширенная версия доклада

Семинар НИУ ВШЭ по высокопроизводительным вычислениям

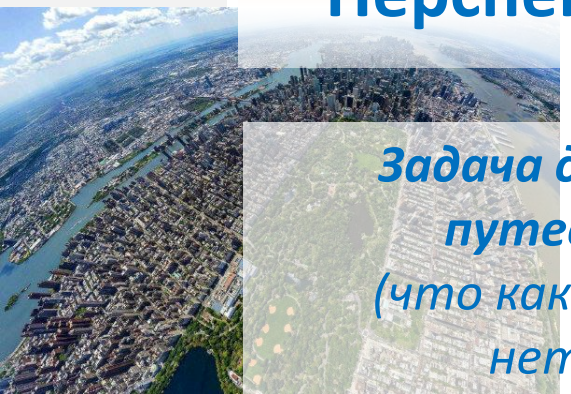
Москва, 04.06.2024

Тенденции и результаты в ML и AI (2020-2024)

- Результаты, проблемы и вызовы 2020 (CV, NLP, GM, RL)
- Тенденции и результаты 2020-2024:
  - компьютерное зрение
  - большие языковые модели
  - генерация данных
  - обучение с подкреплением
  - LLM-агенты, общий ИИ (AGI)
- Перспективы, проблемы и вызовы 2024

4 трека из 2020

Новый трек 2020+



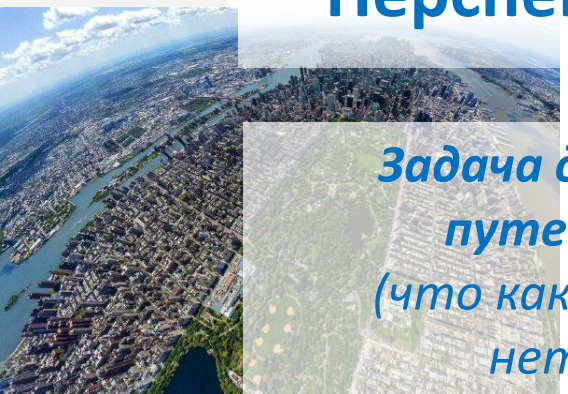
Задача доклада – дать общую картину ландшафта ИИ. *Это* ***набор кратких путеводителей по областям, результатам и идеям современного ИИ,*** *(что как развивается, ключевые слова, где про это почитать) поэтому в нем нет глубоких объяснений, зато дан ряд примеров, схем и обобщений*

Тенденции и результаты в ML и AI (2020-2024)

- Результаты, проблемы и вызовы 2020 (CV, NLP, GM, RL)
- Тенденции и результаты 2020-2024:
  - компьютерное зрение
  - ~~большие языковые модели~~ **не сегодня**
  - генерация данных
  - обучение с подкреплением
  - ~~LLM-агенты, общий ИИ (AGI)~~ **не сегодня**
- Перспективы, проблемы и вызовы 2024

4 трека из 2020

Новый трек 2020+



Задача доклада – дать общую картину ландшафта ИИ. *Это* ***набор кратких путеводителей по областям, результатам и идеям современного ИИ,*** *(что как развивается, ключевые слова, где про это почитать) поэтому в нем нет глубоких объяснений, зато дан ряд примеров, схем и обобщений*

Тенденции и результаты 2021-2024 в области компьютерного зрения

Трансформеры в CV

Фундаментальные модели в CV

Multimodal LLM

CNN vs. Transformers

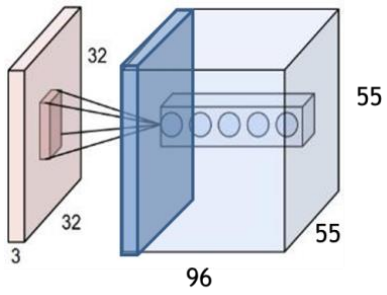
Исходный инструмент:
Персептрон (MLP)

Сверточные нейронные сети в компьютерном зрении

CV до 2020

Visualizing and Understanding Convolutional Networks,, ECCV'14

Ключевое изобретение: сверточный нейрон Работает локально в небольшой окрестности



+ Иерархическая обработка с повышением абстракции данных от уровня к уровню

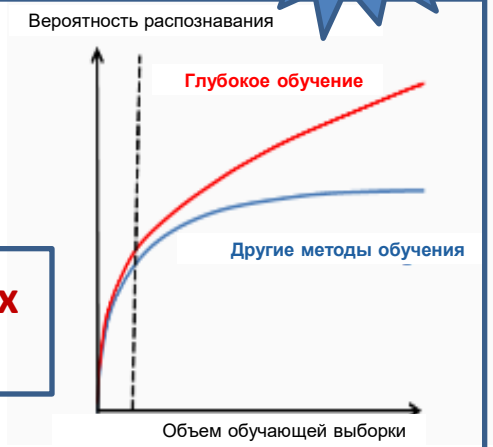
Какие элементы изображения распознают нейроны разных уровней: **чем выше слой сети, тем выше уровень абстракции**

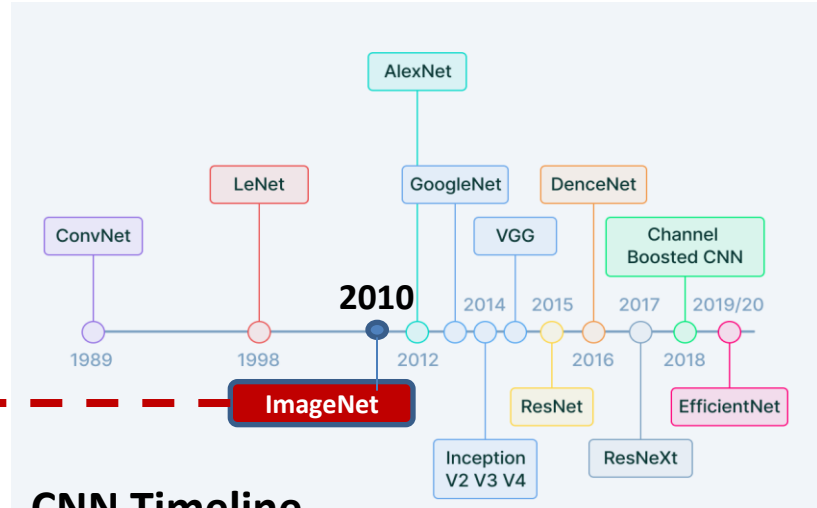
AlexNet (2011)

ХИТ ИИ 2011!

+ С 2011 г. - распознавание образов на уровне человека или выше (superhuman)

+ Обучение на сверхбольших объемах данных



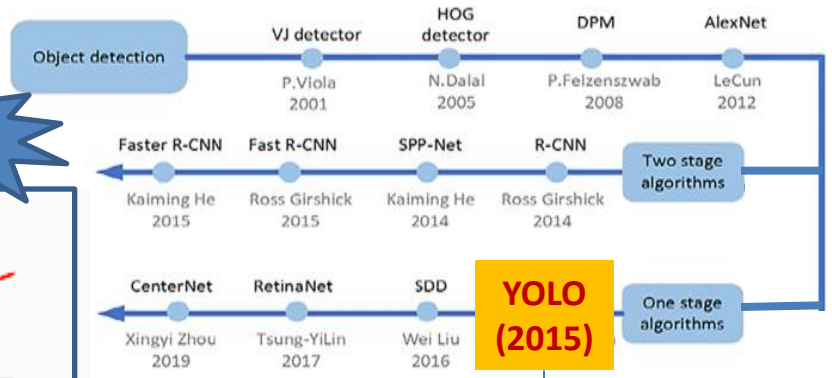


CNN Timeline

Evolution of Convolutional Neural Network, 2023

V7 Labs

Object Detection with CNN Timeline



ХИТ ИИ 2015!

2015-16: CNN решают все задачи компьютерного зрения и царят до 2020...

CNN Architectures

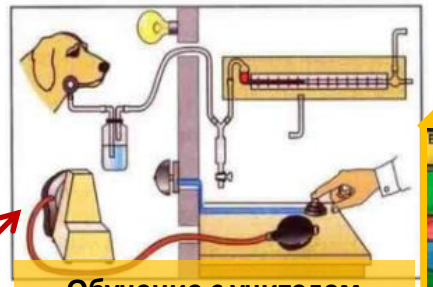
AlexNet
VGG16
VGG19
Inception V1
Inception V2
GoogleNet
ResNet18
ResNet34
ResNet50
ResNet101
ResNet152
DenseNet
SqueezeNet
MobileNet V2
ShuffleNet V2
EfficientNet

Проблемы и вызовы в CV (2020)

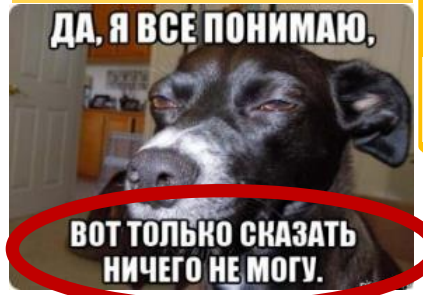
Почему распознавания образов не хватило?

CV до 2020

CNN учится распознаванию образов на примерах: стимул-реакция без какого-либо символьного представления



Обучение с учителем



Модель сцены

Память + сообразительность ≠ Интеллект!

Основное ограничение: невозможность использования структурных моделей, логик и онтологий

Важная проблема: необходимость обучения на больших выборках (**желательно Zero-Shot или Few-Shot**)

В 2016-2019 гг. мы считали, что основным путем преодоления ограничений ГНС станут структурные ГНС на графах

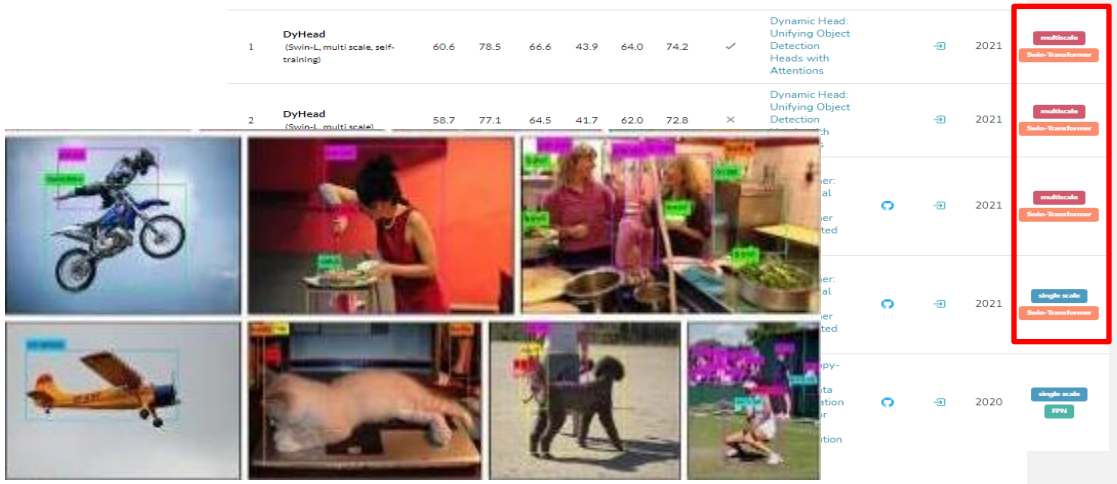
Но мы ошиблись!
Сегодня эту роль играет совсем другой тип ГНС, возникший в области NLP, *к которой мы и переходим…*

Трансформеры в CV

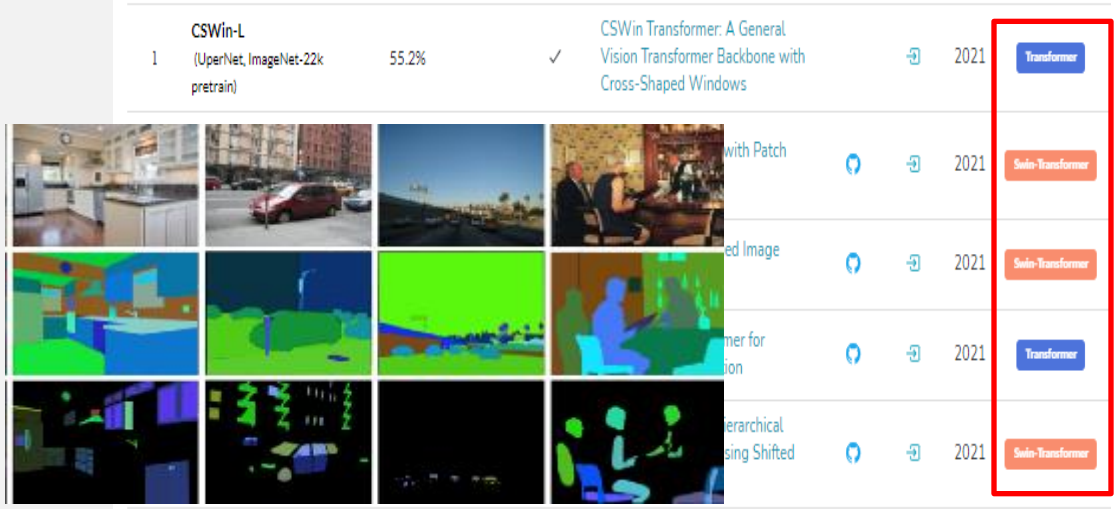
Трансформеры идут в CV! (2020+)

Классификация ImageNet

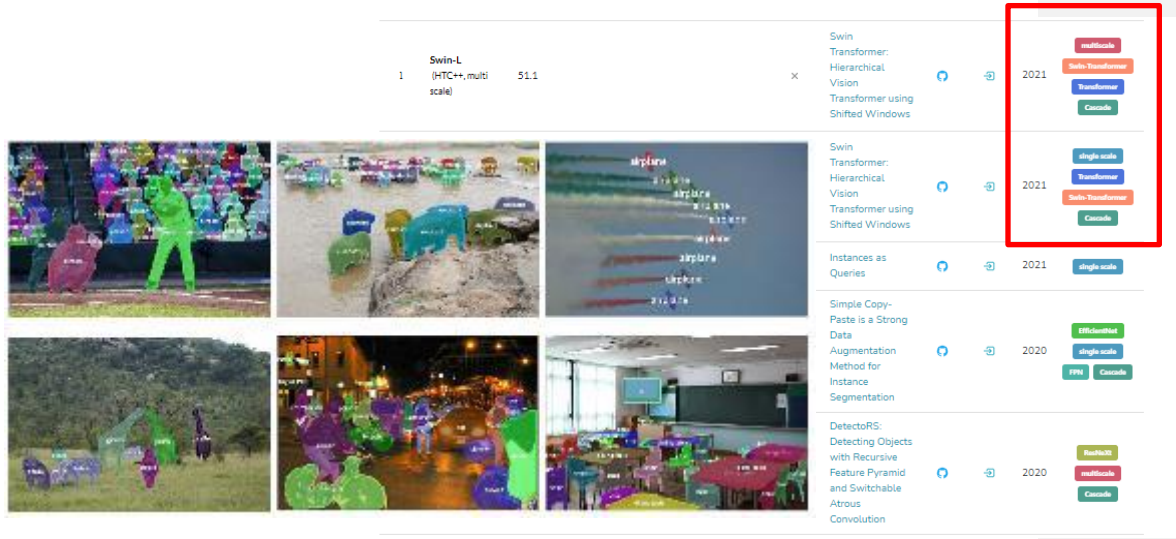
Обнаружение MS COCO



Сегментация ADE20K



Сегментация(instance) MS COCO

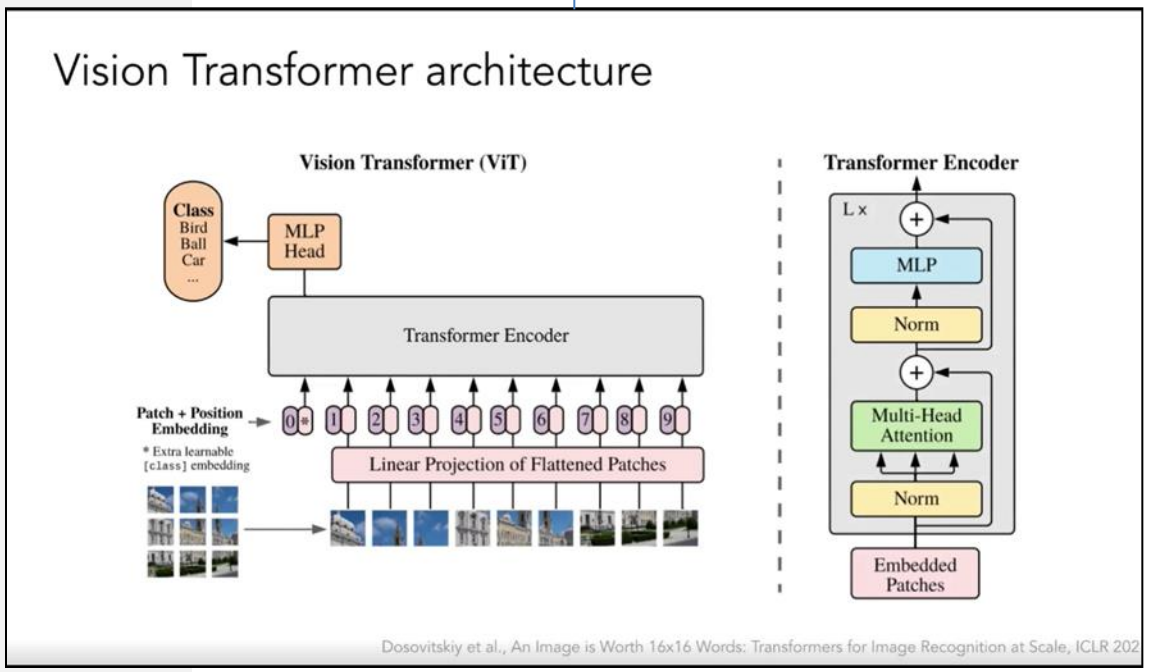


По данным https://paperswithcode.com/

Трансформеры идут в CV! (2020+)

Базовая идея применения трансформеров к задачам зрения: вместо текстовых токенов - фрагменты изображений

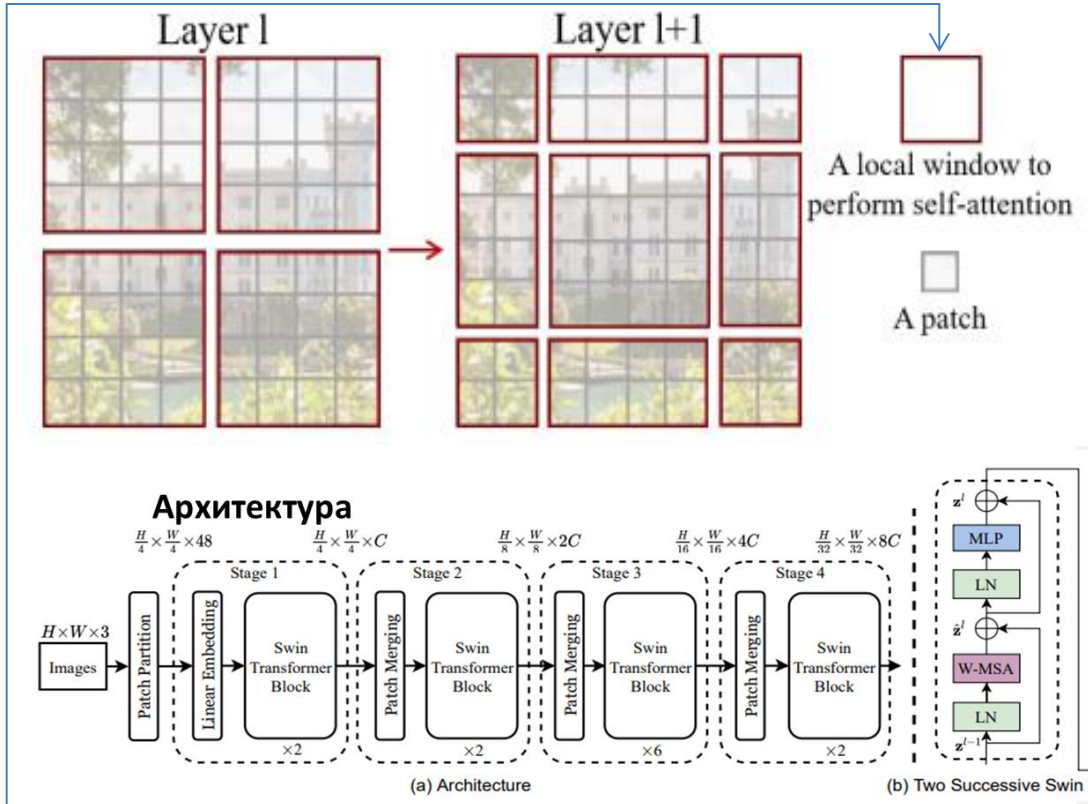
Vision Transformer (2020)



Изображение как «текст»: все токены сравниваются попарно, отсюда квадратичные вычисления

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *Alexey Dosovitskiy*, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

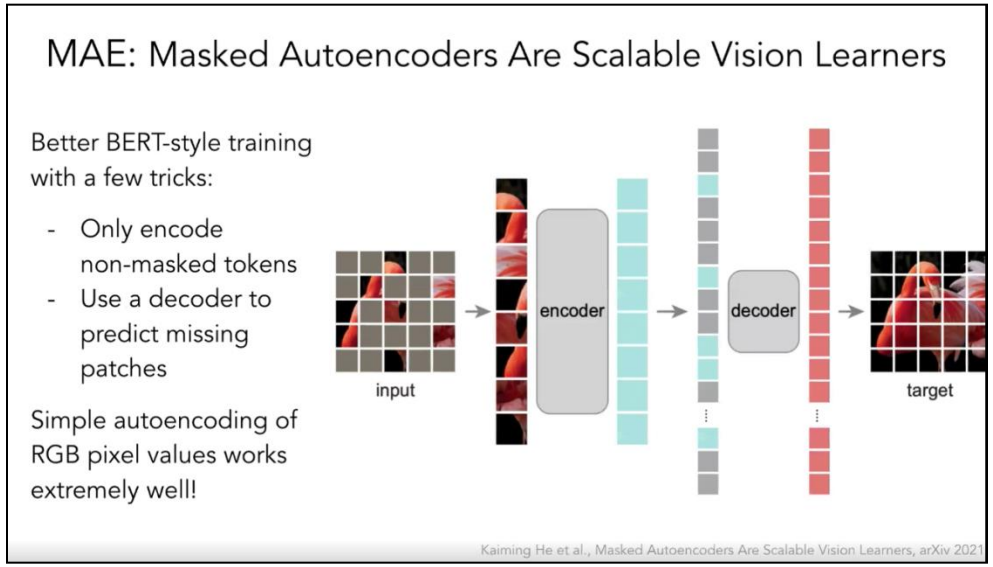
Shifted Windows *Swin Transformer (2021)*

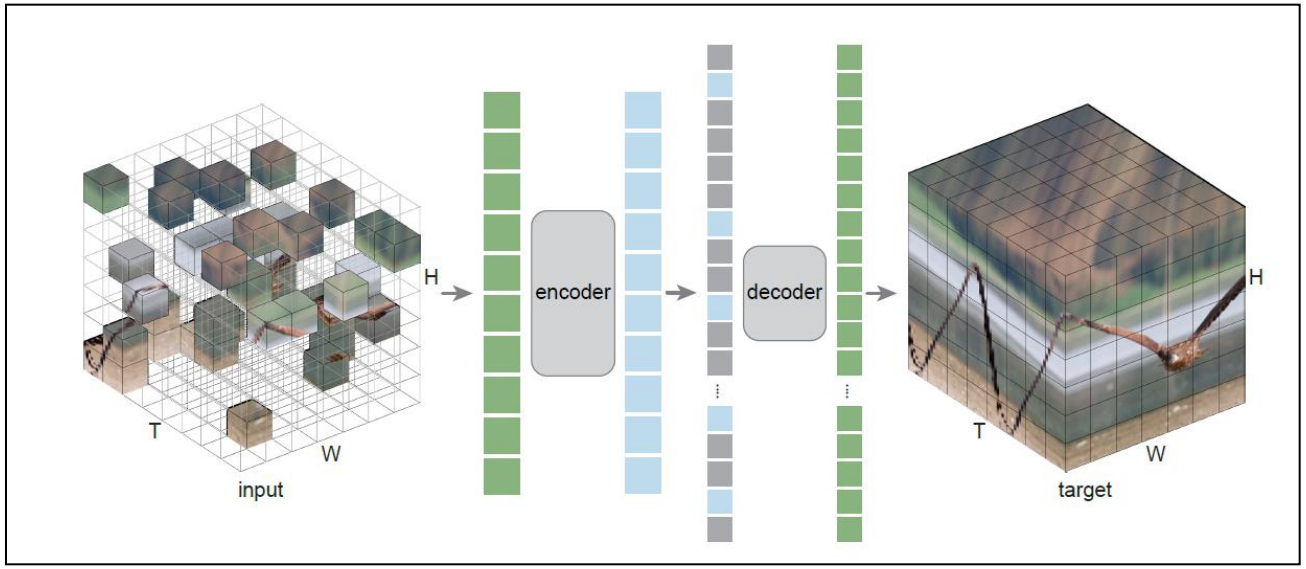


Трансформеры горзадо медленнее сверточных сетей из-за квадратичных вычислений, но если их блоки применять локально, то скорость существенно улучшается!

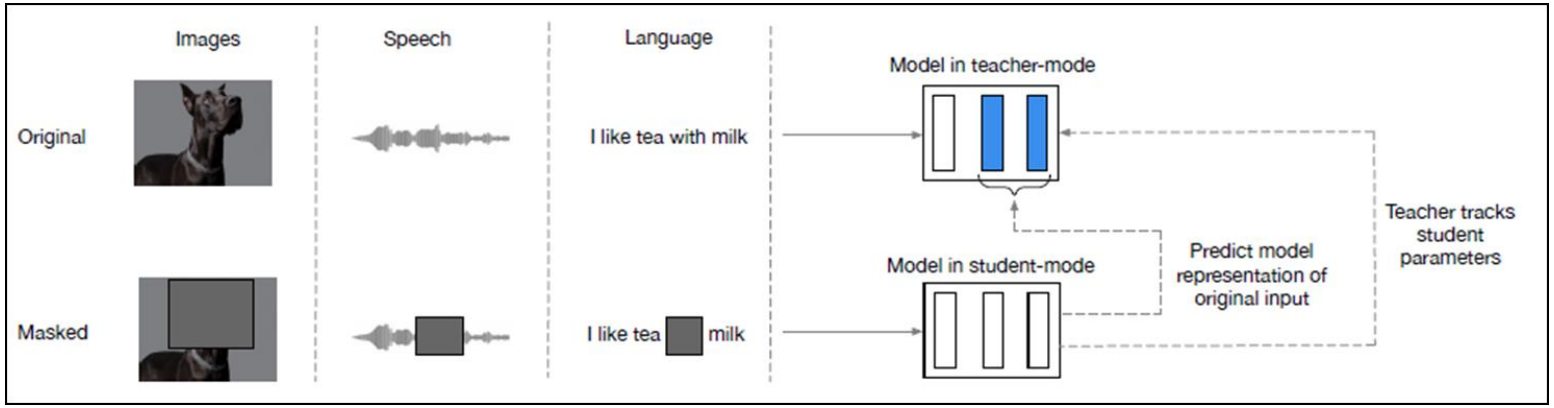
MAE: Masked Autoencoders As Scalable Vision Learners (2021)

Self-Supervised Learning

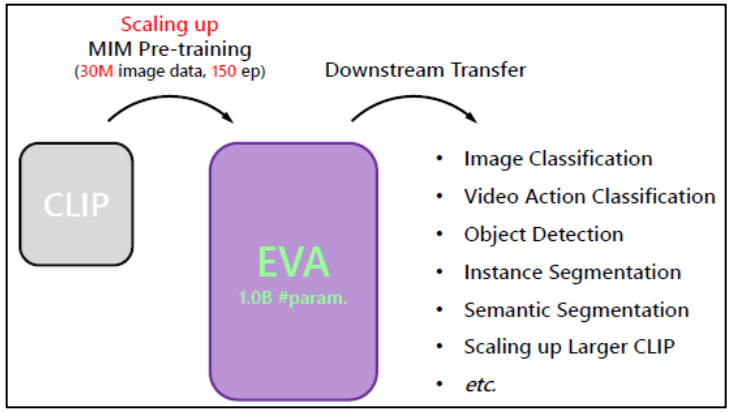




MAE: Masked Autoencoders As Spatiotemporal Learners, Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, Kaiming He, Meta AI, FAIR, **2022**



data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language, MetaAI, 2022



EVA: Exploring the Limits of Masked Visual Representation Learning at Scale, Fang et al., 2022

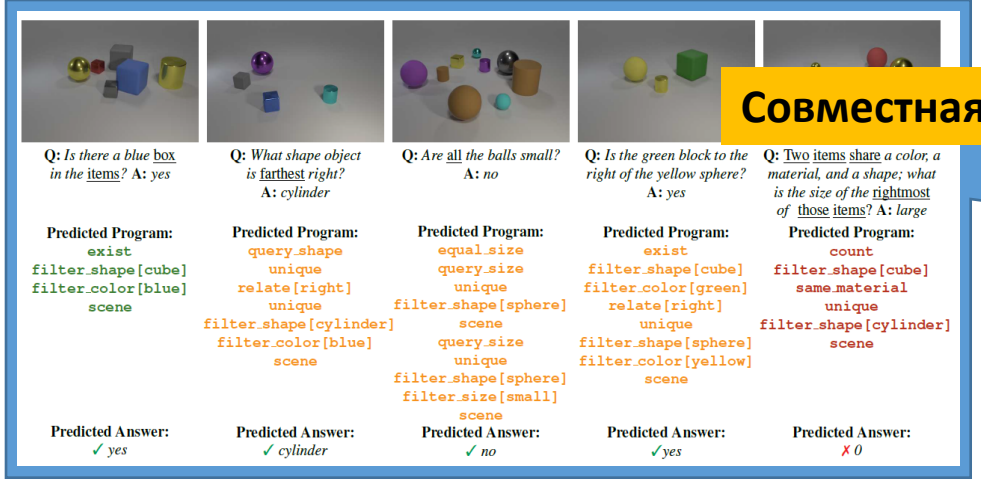
Сети GPT и BERT обучают на восстановление «замаскированных» фрагментов текста.
Оказалось для изображений, видео и других типов данных это тоже работает!

Multi-modal Learning

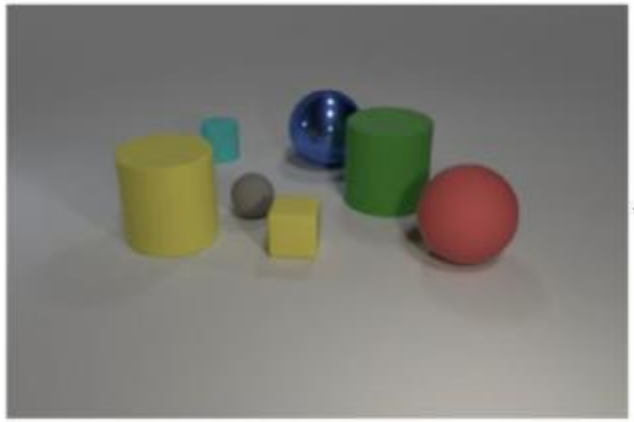
Совместная обработка сигнальной и символьной информации

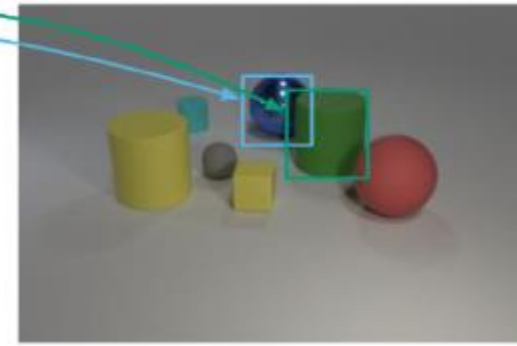
2017: ИИ-1 встречается с ИИ-2!!!

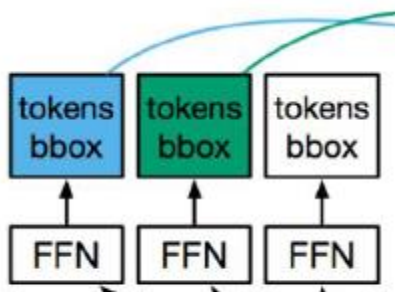
M-DETR



Ответы на визуальные вопросы, требующие рассуждений: «визуальный тест Тьюринга»







2021: Раньше казалось, что здесь нужны базы знаний и логическое программирование, теперь это делают трансформеры!

Kamath et al 2021

Self-Attention for Vision

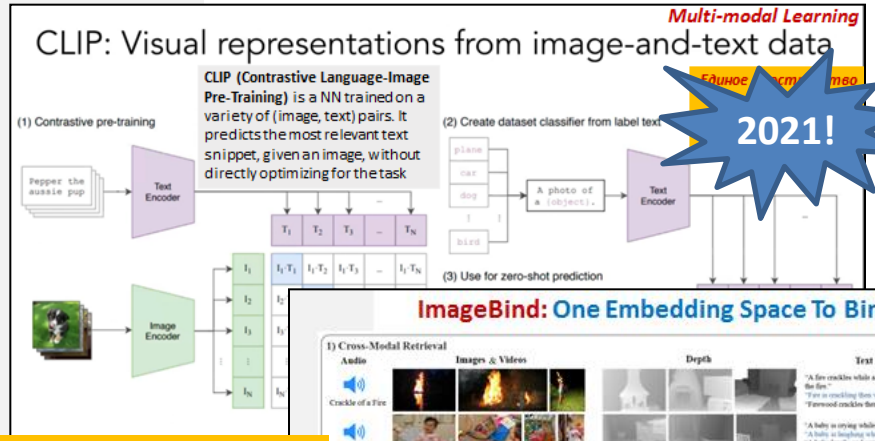
Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley

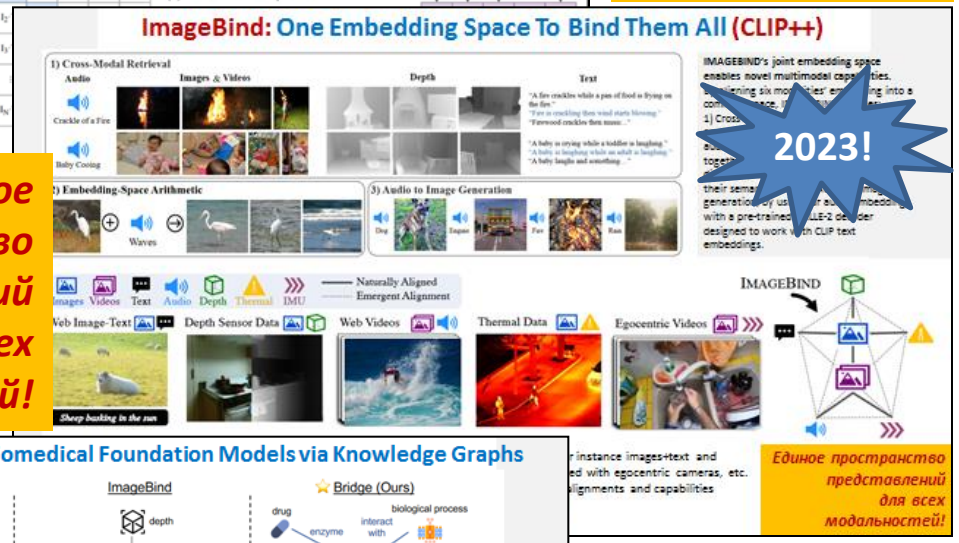
Multimodal Learning

Overvew

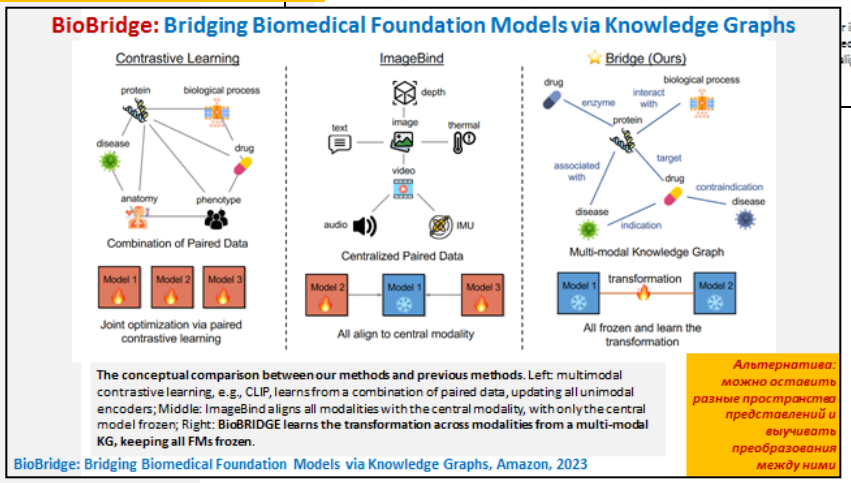
Data Fusion at the feature level (на уровень ниже)



Единое пространство представлений для текста и изображений!



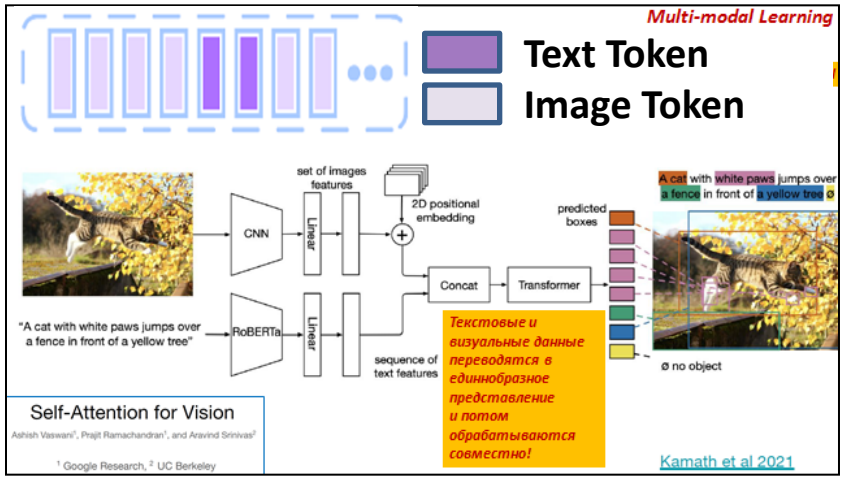
Единое пространство представлений для всех модальностей!



Альтернатива: можно оставить разные пространства представлений и выучивать преобразования между ними

Разные пространства представлений и выученные преобразования между ними

Data Fusion at the token level



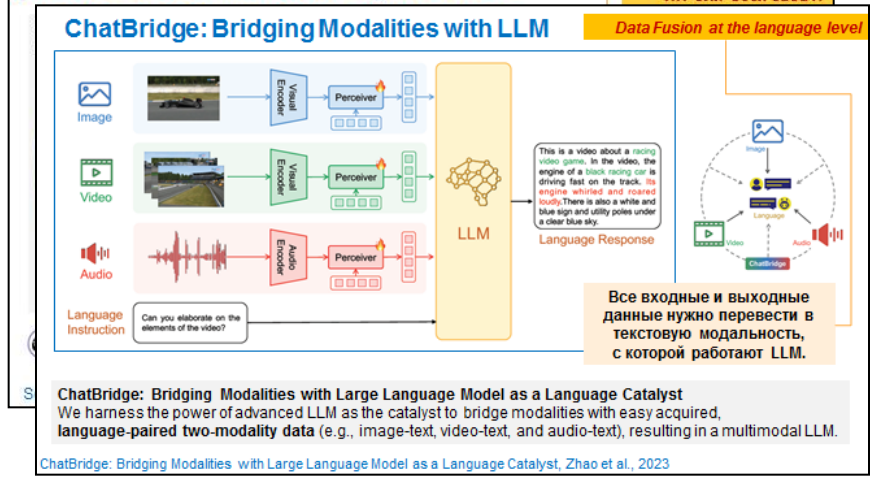
Текстовые и визуальные данные переводятся в единнобразное представление и потом обрабатываются совместно!

Текстовые и визуальные данные переводятся в токены, единая лента которых идет на вход трансформера

Data Fusion at the language level (на уровень выше)

Socratic Models: как заставить разные отделы «мозга» работать вместе. От GhatGPT к чату ИИ-агентов

Возможна ли универсальная схема ИИ для всех задач?



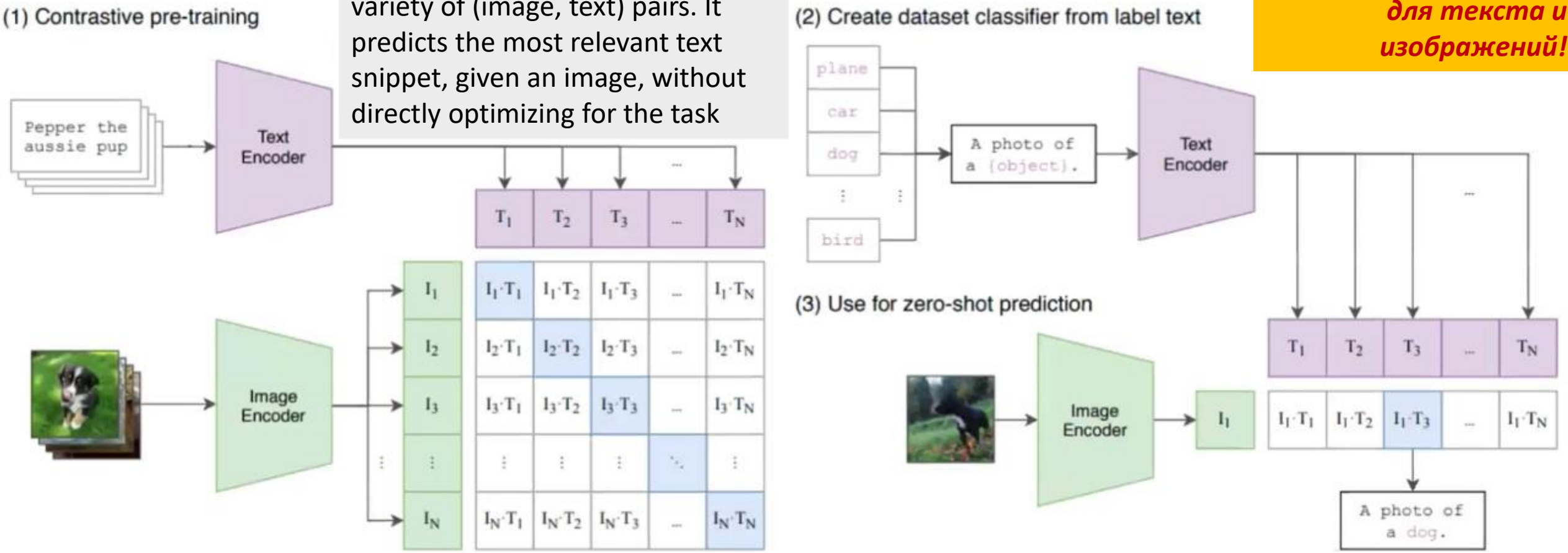
Все входные и выходные переводим в текстовую модальность, с которой работают LLM в чате

Multimodal Learning
Vision-Language Models (VLM)

CLIP: Visual representations from image-and-text data

CLIP (Contrastive Language-Image Pre-Training) is a NN trained on a variety of (image, text) pairs. It predicts the most relevant text snippet, given an image, without directly optimizing for the task

Единое пространство представлений для текста и изображений!

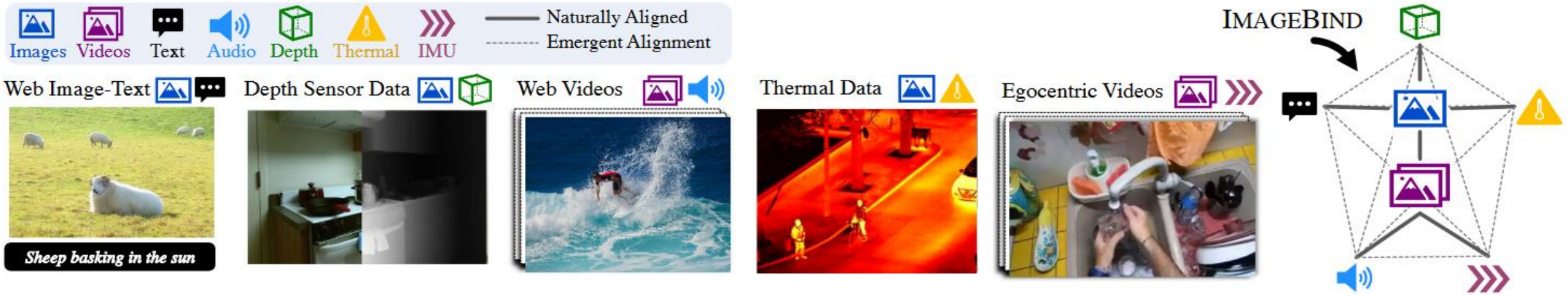


Learning Transferable Visual Models From Natural Language Supervision, 2021

Data Fusion at the feature level

ImageBind: One Embedding Space To Bind Them All (CLIP++)

IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: 1) Cross-Modal Retrieval, which shows emergent alignment of modalities such as audio, depth or text, that aren't observed together. 2) Adding embeddings from different modalities naturally composes their semantics. And 3) Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 decoder designed to work with CLIP text embeddings.

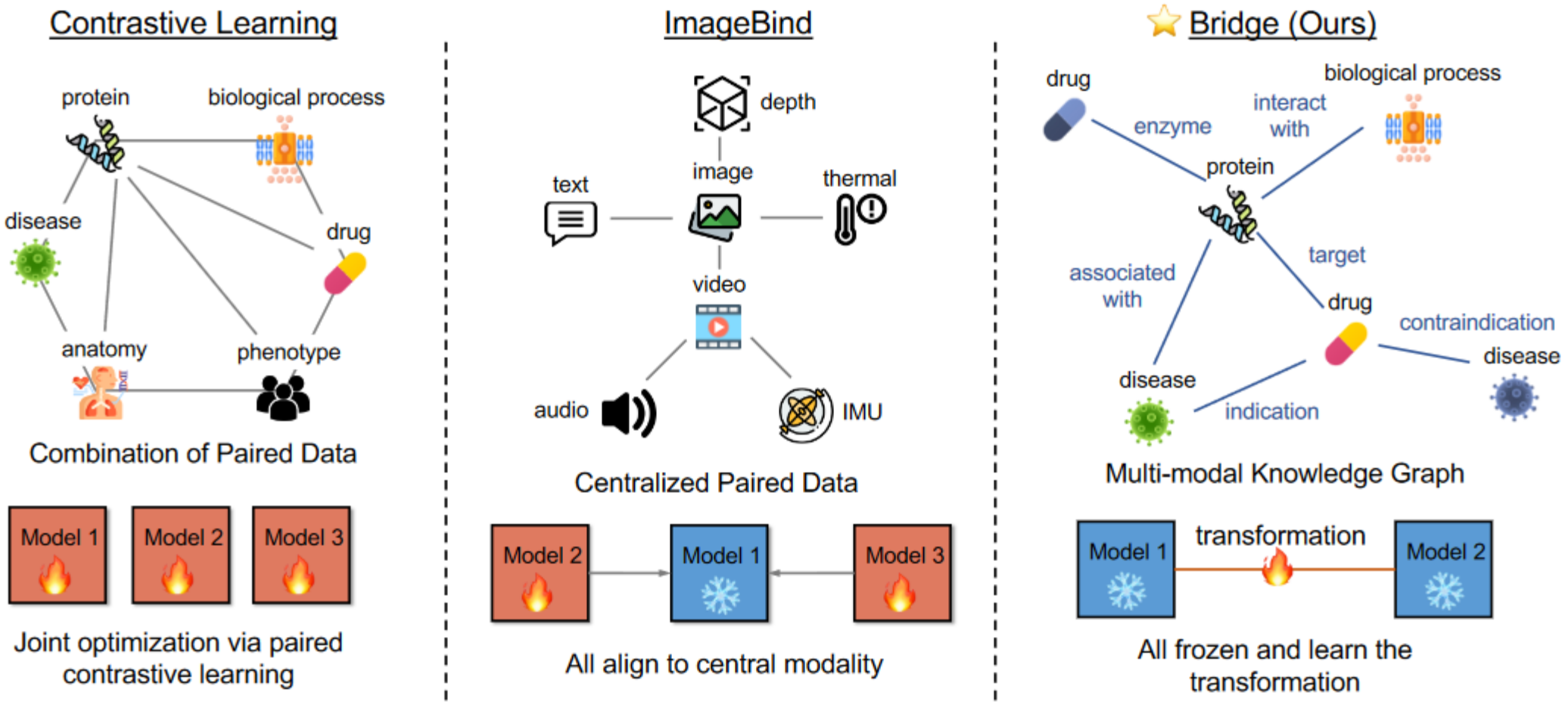


IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, etc. IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities

Единое пространство представлений для всех модальностей!

IMAGEBIND: One Embedding Space To Bind Them All, FAIR, Meta AI, 2023

BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs



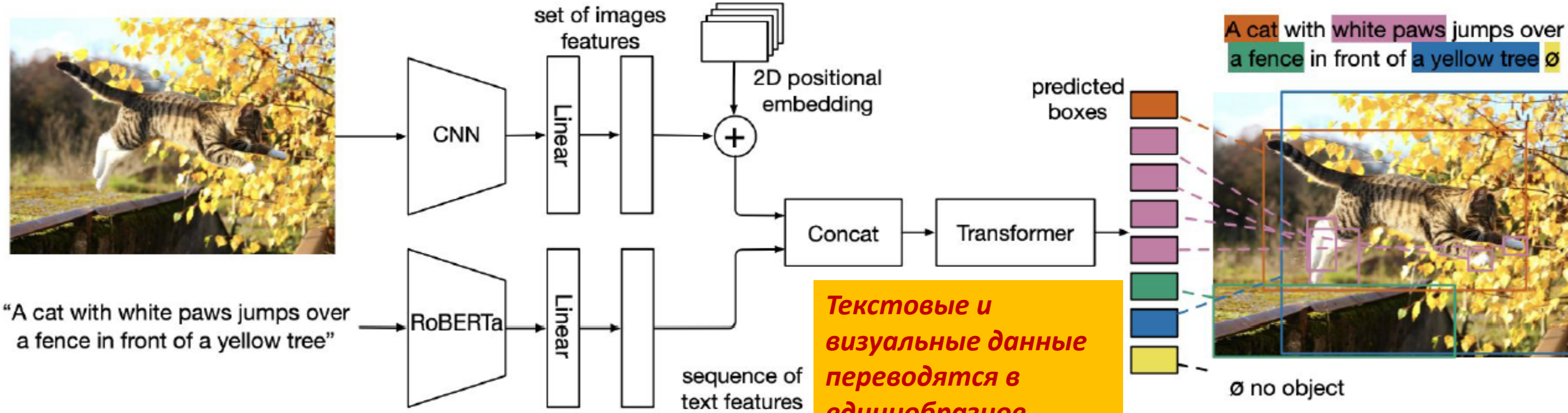
The conceptual comparison between our methods and previous methods. Left: multimodal contrastive learning, e.g., CLIP, learns from a combination of paired data, updating all unimodal encoders; Middle: ImageBind aligns all modalities with the central modality, with only the central model frozen; Right: **BioBRIDGE learns the transformation across modalities from a multi-modal KG, keeping all FMs frozen**.

Альтернатива: можно оставить разные пространства представлений и выучивать преобразования между ними

BioBridge: Bridging Biomedical Foundation Models via Knowledge Graphs, Amazon, 2023

Multimodal DETR (M-DETR)

Data Fusion at the token level



Текстовые и визуальные данные переводятся в единнобразное представление и потом обрабатываются совместно!

Self-Attention for Vision

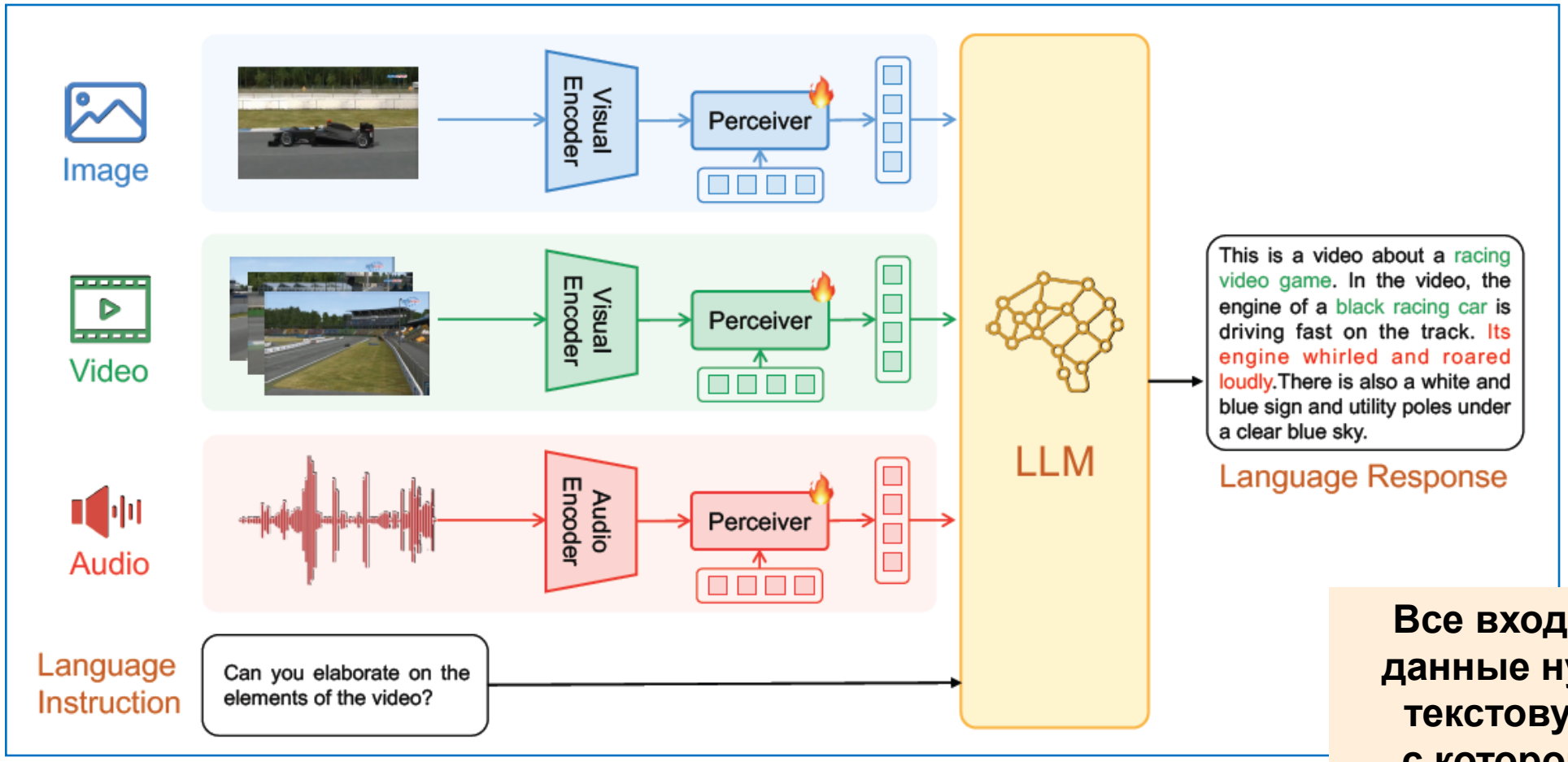
Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

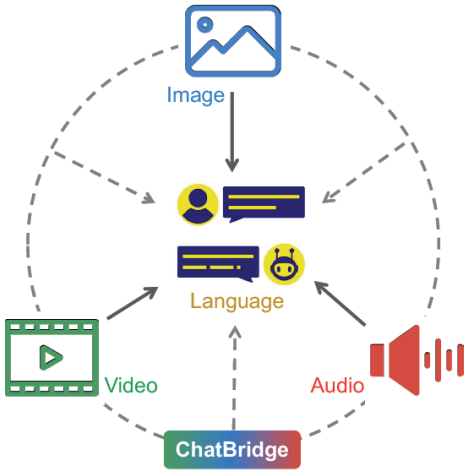
[1] Google Research, [2] UC Berkeley

Kamath et al 2021

ChatBridge: Bridging Modalities with LLM

Data Fusion at the language level



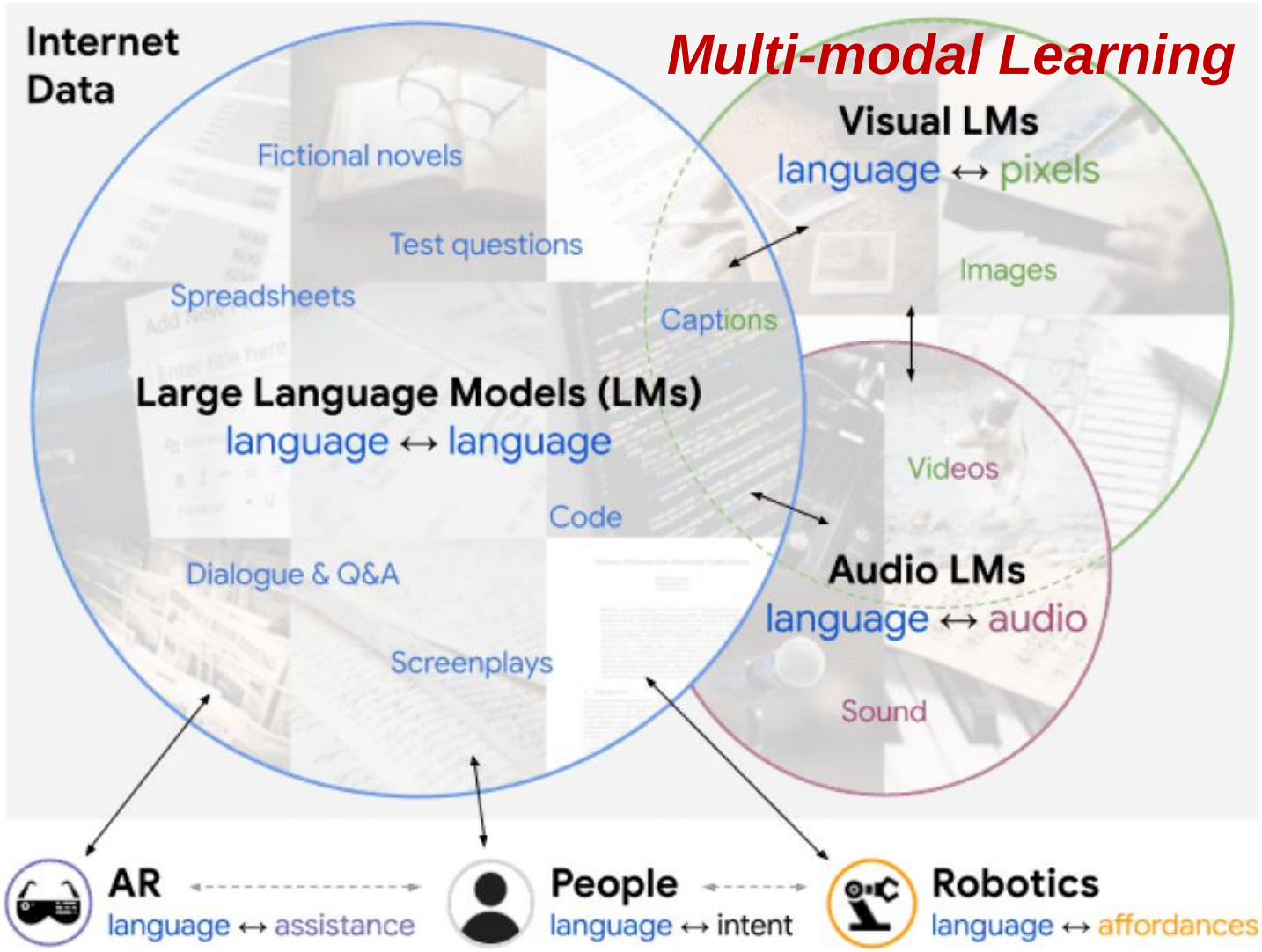


Все входные и выходные данные нужно перевести в текстовую модальность, с которой работают LLM.

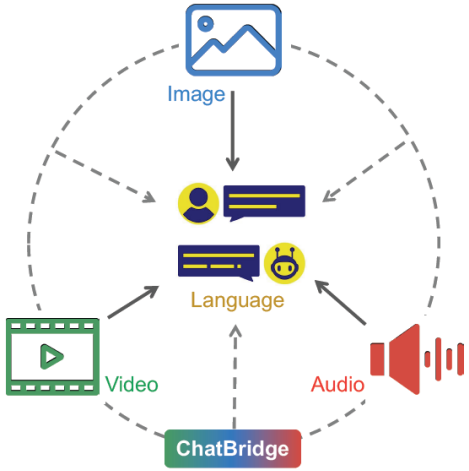
ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst
We harness the power of advanced LLM as the catalyst to bridge modalities with easy acquired, **language-paired two-modality data** (e.g., image-text, video-text, and audio-text), resulting in a multimodal LLM.

ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst, Zhao et al., 2023

Socratic Models: как заставить разные отделы «мозга» работать вместе. От GhatGPT к чату ИИ-агентов



Data Fusion at the agent level



Гипотеза: для этого все входные и выходные данные нужно перевести в текстовую модальность, с которой работают LLM.

Проблема: для обучения нужна огромная мультимодальная база

Решение: научить модели использовать запросы, и пусть модели общаются между собой

«Сократические модели» это способ построения расширяемого коллектива ИИ-агентов с «открытым» списком задач и типов данных, для работы с которыми добавляются новые агенты

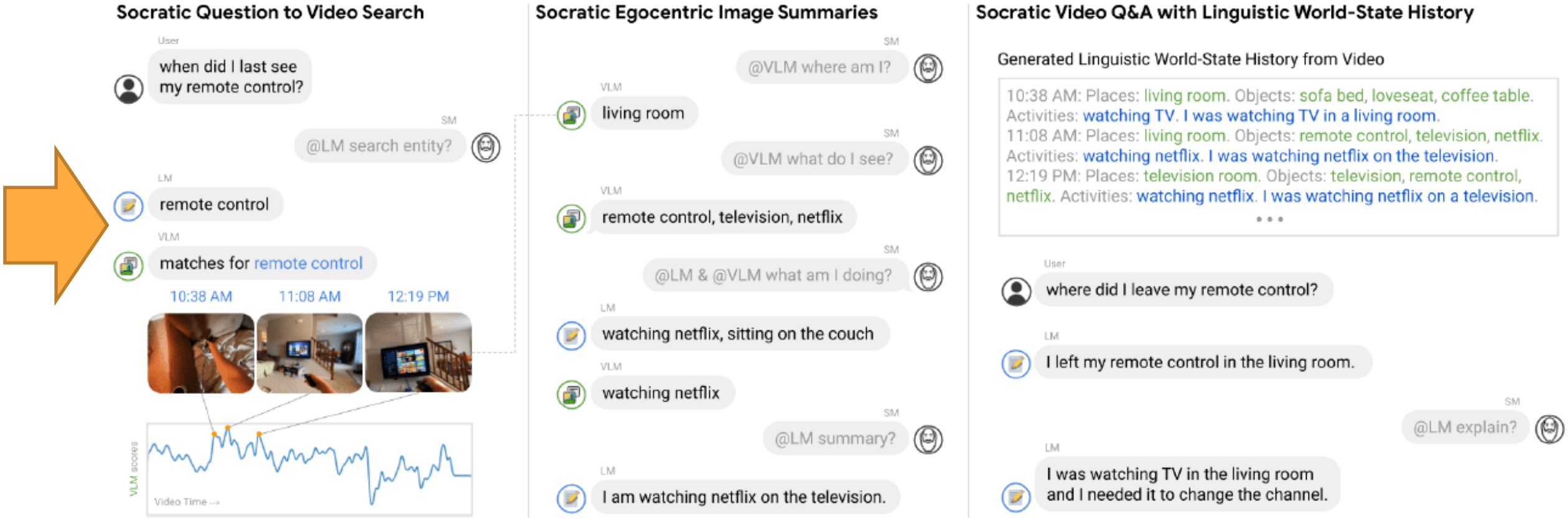
Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Google, 2022

Socratic Models: как заставить разные отделы «мозга» работать вместе. От GhatGPT к чату ИИ-агентов

Data Fusion at the agent level

Метод Сократа – решение задачи путем диалога

Внутри работа «сократической модели» действительно выглядит как чат ИИ-агентов!



from LMs are blue, VLMs green, ALMs purple, prompt text gray, user inputs magenta, VLM-chosen LM outputs green-underlined blue, and ALM-chosen LM outputs purple-underlined blue.

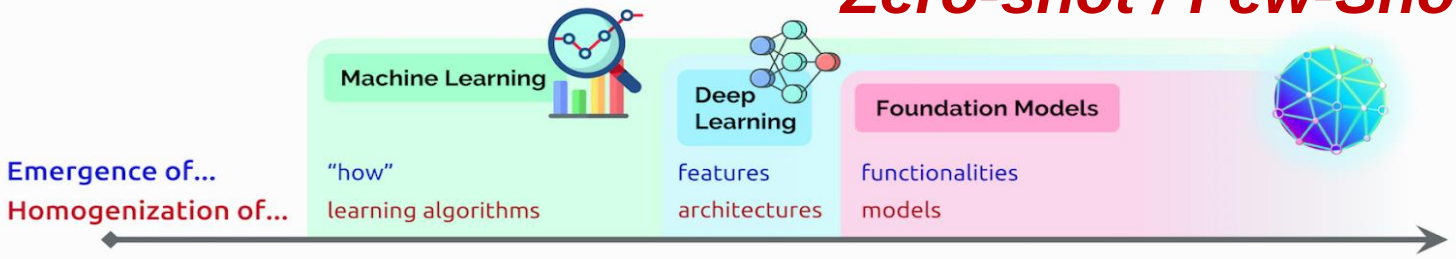
Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Google, 2022

Foundation Models (2020+)

Идея: обучить одну модель на большом объеме данных, а потом сразу применять ее для множества приложений

Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Zero-shot / Few-Shot



Center for Research on Foundation Models

HAI

Stanford University
Human-Centered
Artificial Intelligence

In recent years, a new successful paradigm for building AI systems has emerged: **Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model**.

Foundation models (e.g., GPT-3) have demonstrated impressive behavior, but can fail unexpectedly, harbor biases, and are poorly understood. Nonetheless, they are being deployed at scale.

https://arxiv.org/abs/2108.07258

https://blog.inten.to/foundation-models-b89e7610057

Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

Фундаментальные модели для CV

(Zero-Shot Solving Classical CV Tasks)

Уровни анализа данных в CV

Уровень сцены	**классификация и описание сцен (отношения, действия), VQA**
Уровень объектов	**обнаружение объектов, семантическая сегментация…**
Уровень признаков	**особые точки, линии, углы, области, контуры, текстура…**
Уровень изображения	**пикселы (положение, яркость, цвет)**

Fundamental Models for Vision Tasks

Overview

Supervised *Self-supervised*

На какой задаче предобучить аундамен-тальную модель для CV?

Уровень сцены

Recognize Anything: Image Tagging (Multi-label Image Recognition)

Возможно, любая сложная визуальная задача может стать источником для

Image Captioners Are Scalable Vision Learners Too

Pretraining: Какая прокси-задача лучше всего? (Аналог классификации на ImageNet)

2023!

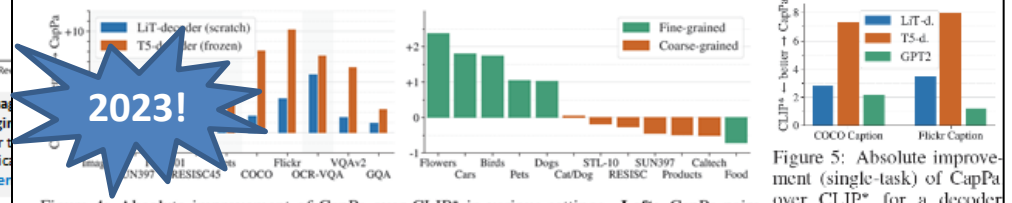
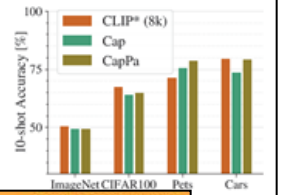


Figure 4: Absolute improvement of CapPa over CLIP* in various settings. **Left:** CapPa pairs significantly better with decoders in image-language tasks, especially when the decoder is a pre-trained and frozen language model. **Right:** CapPa seems to be a noticeably better frozen feature extractor for fine-grained classification tasks (we show L/14 results, see appendix for B/16).

Figure 5: Absolute improvement (single-task) of CapPa over CLIP* for a decoder trained from scratch (LiT-d.), a frozen T5 decoder, and a frozen GPT-2 similar to [44].

We presented an extensive comparison of vision encoders pretrained with a contrastive and generative (captioning) objective and found that the **generatively pretrained encoders obtain better performance when used for captioning, VQA, fine-grained and few-shot classification tasks**, while achieving



Уровень объектов

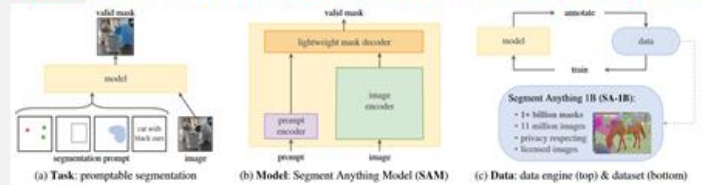
Grounding DINO: Open-Set Object Detection

Задача обнаружения столь же фундаментальна!

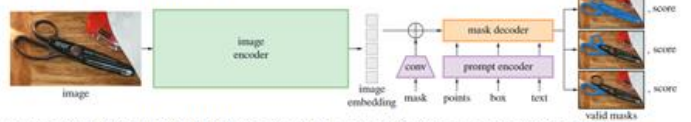
SAM: Segment Anything, Foundation model for segmentation

Задача сегментации столь же фундаментальна!

2023!



Foundation model for segmentation with three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables **zero-shot transfer to a range of tasks via prompt engineering**, and a data engine for collecting SA-1B, our dataset of over 1 billion masks.



Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object

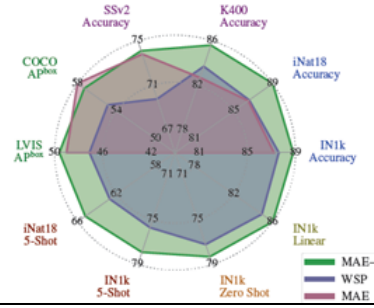
Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

2023

Уровень изображения

MAE *pre*-pretraining for billion-scale pretraining

MAE + Weakly-supervised pretraining (WSP)



Pre-pretraining Masked Autoencoder (MAE) learns visual representations from image datasets without using any

MAE + Weakly-supervised pretraining (WSP)

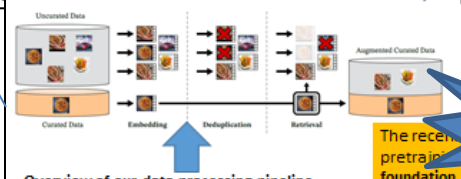
MAE→WSP, or MAWS for short, first trains the encoder using the MAE self-supervised method using only the images. This pre-pretraining stage initializes the model while simultaneously being computationally efficient

DINOv2: Learning Robust Visual Features without Supervision

Discriminative Self-supervised Pre-training

DINO: DIstillation with NO labels. DINO simplifies self-supervised ViT training by directly predicting the output of a teacher network (built with a momentum encoder) by using a standard cross-entropy loss.

Emerging Properties in Self-Supervised Vision Transformers, 2021



Overview of our data processing pipeline. Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting

Discriminative Self-supervised Pre-training

2024!

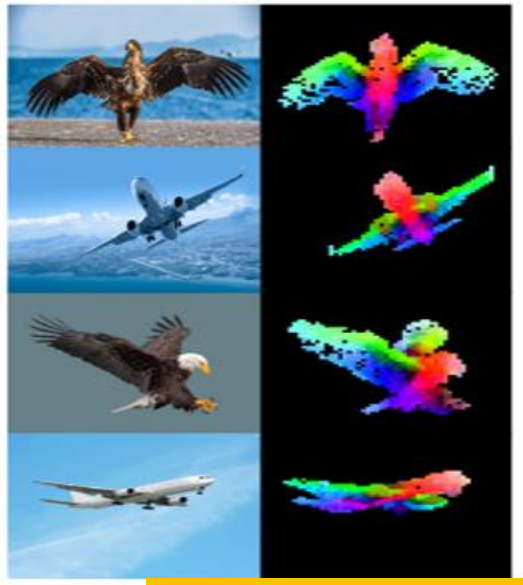
The recen… language processing for model pretrai… data have opened the way for similar **foundation mo… …ion**. These models could greatly simplify the use of …ages in any system by producing general-purpose visual features, i.e., **features that work across image distributions and tasks without finetuning**. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit …bine different techniques to scale our …nd model size.

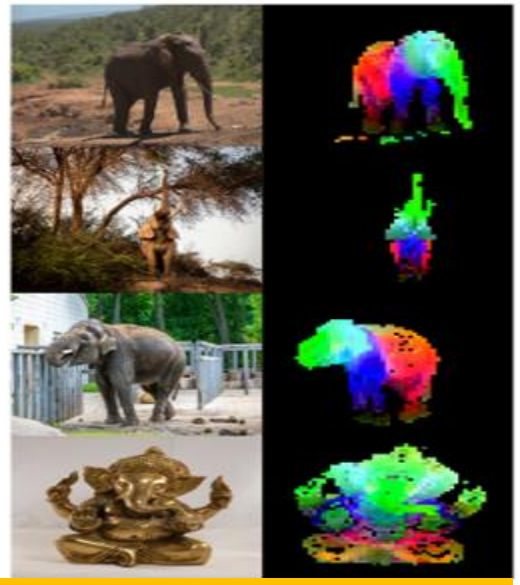
…esearch, 2024.

Видимо, любая сложная визуальная задача может стать источником для получения фундаментальной модели, которую не придется потом дообучать для решения новых задач: важно, чтобы в ней выработались нужные признаки

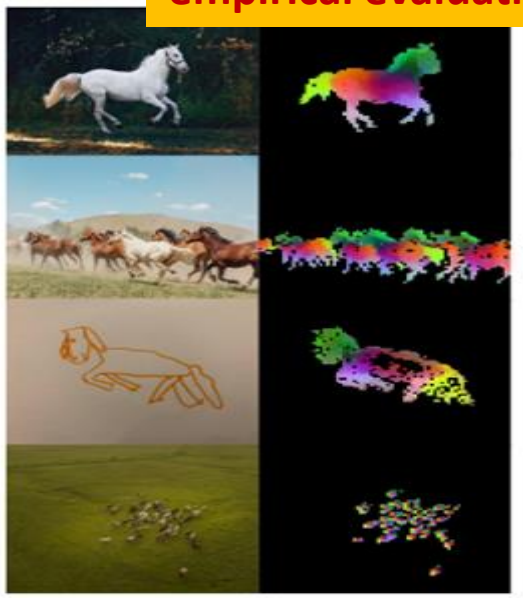
DINOv2: Learning Robust Visual Features without Supervision

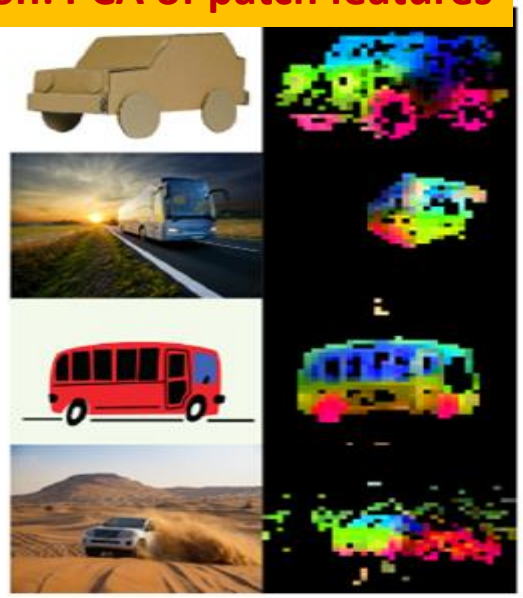
Фундаментальная модель должна иметь универсальные и робастные признаки



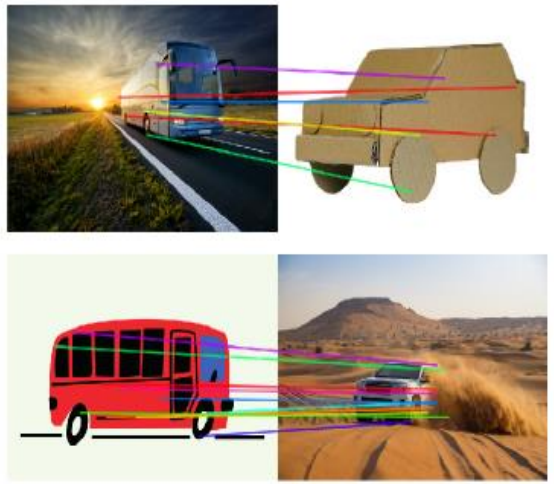


empirical evaluation: PCA of patch features





empirical evaluation: Patch matching



DINOv2: Learning Robust Visual Features without Supervision. Meta AI Research, 2024.

Уровень признаков

Image Captioners Are Scalable Vision Learners Too

Pretraining: Какая прокси-задача лучше всего? (Аналог классификации на ImageNet)

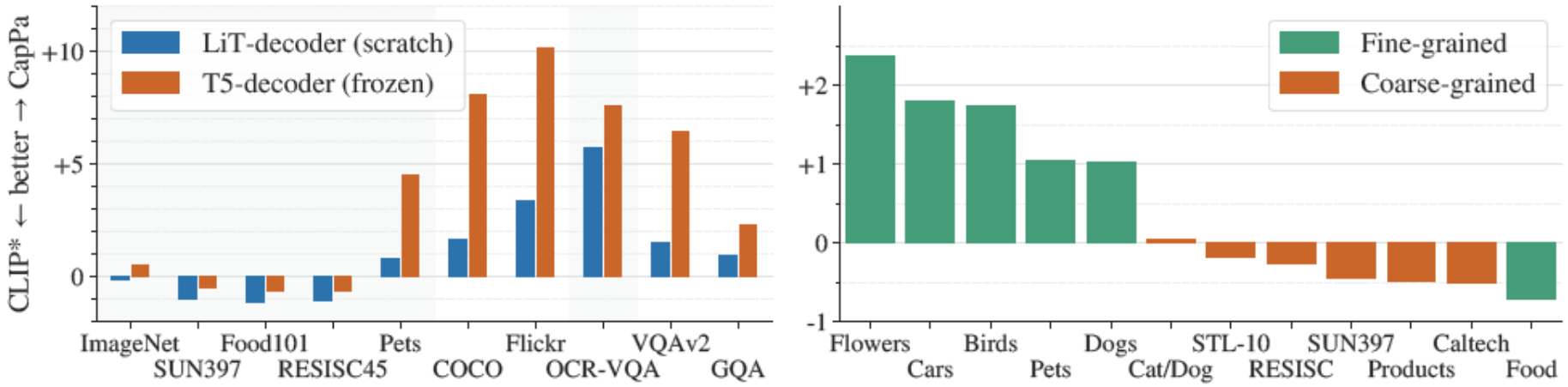


Figure 4: Absolute improvement of CapPa over CLIP* in various settings. **Left:** CapPa pairs significantly better with decoders in image-language tasks, especially when the decoder is a pre-trained and frozen language model. **Right:** CapPa seems to be a noticeably better frozen feature extractor for fine-grained classification tasks (we show L/14 results, see appendix for B/16).

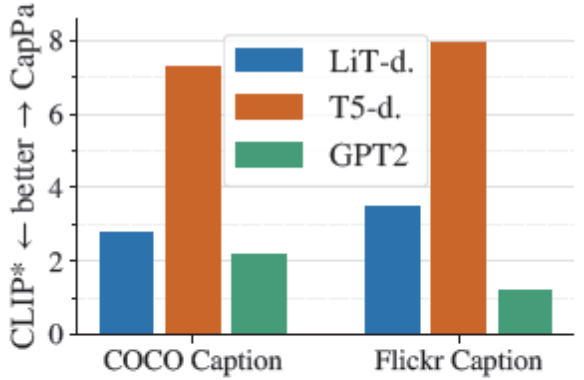


Figure 5: Absolute improvement (single-task) of CapPa over CLIP* for a decoder trained from scratch (LiT-d.), a frozen T5 decoder, and a frozen GPT-2 similar to [44].

We presented an extensive comparison of vision encoders pretrained with a contrastive and generative (captioning) objective and found that the **generatively pretrained encoders obtain better performance when used for captioning, VQA, fine-grained and few-shot classification tasks**, while achieving competitive performance in classification overall. In conclusion, we established **plain image captioning as a competitive pretraining strategy for vision backbones from image-text data**

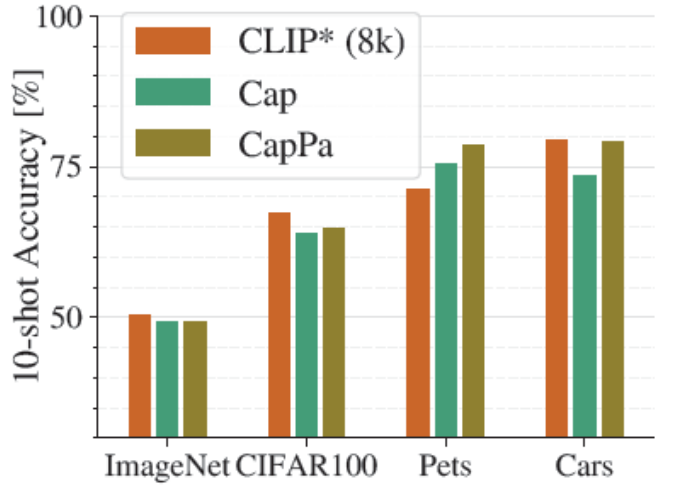


Figure 7: Pretraining on LAION-400M.

Image Captioners Are Scalable Vision Learners Too. Google DeepMind, **2023.**

MAE *pre*-pretraining for billion-scale pretraining

MAE + Weakly-supervised pretraining (WSP)

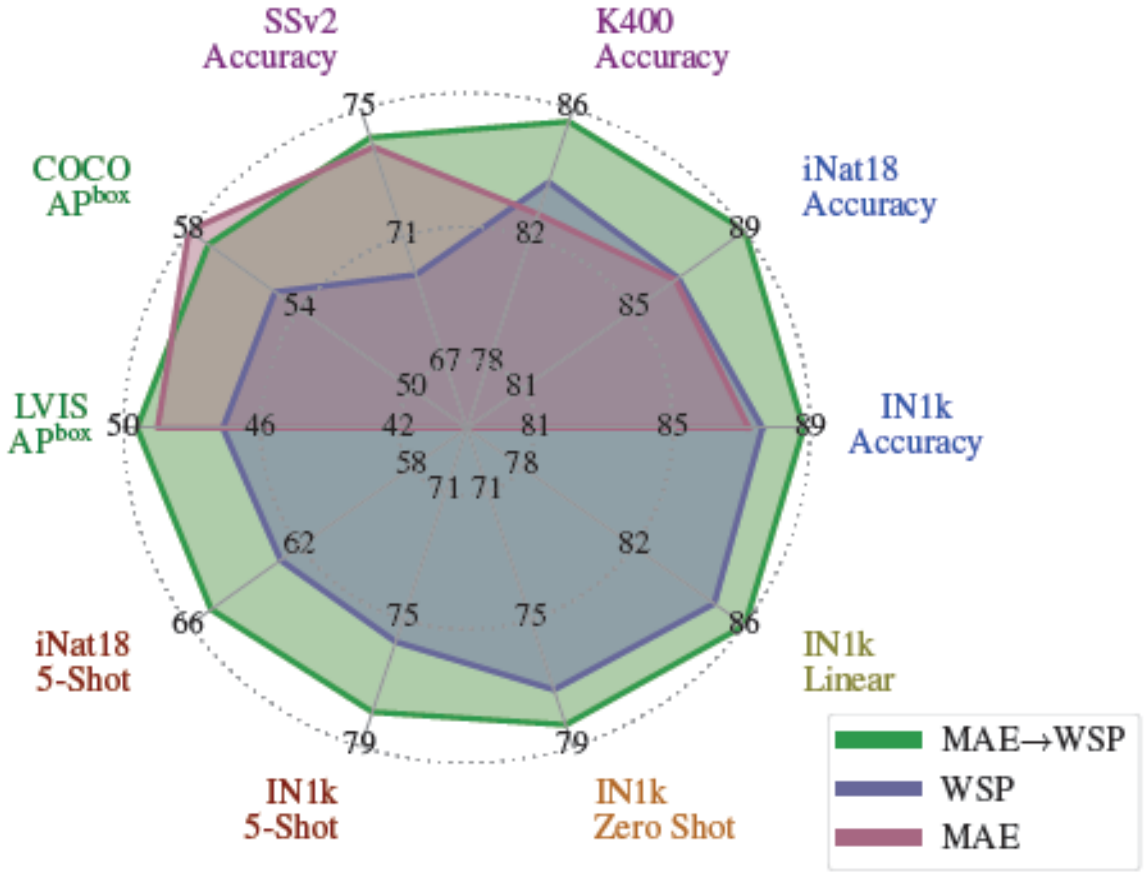


Figure 1: MAE pre-pretraining improves performance. Transfer performance of a ViT-L architecture trained with self-supervised pretraining (MAE), weakly supervised pretraining on billions of images (WSP), and our pre-pretraining (MAE→WSP) that initializes the model with MAE and then pretrains with WSP. Pre-pretraining consistently improves performance.

Pre-pretraining Masked Autoencoder (MAE) learns visual representations from image datasets without using any labels. MAE randomly masks 75% of an image and trains the model to reconstruct the masked input image by minimizing the pixel reconstruction error.

Weakly-supervised pretraining (WSP) leverages images with associated 'weak' supervision for training models. We convert the text into a discrete set of labels, specifically leveraging hash-tag information We then use a multi-label classification loss to train models. We refer to this method as WSP.

MAE→WSP, or MAWS for short, first trains the encoder using the MAE self-supervised method using only the images. This pre-pretraining stage initializes the model while simultaneously being computationally efficient because of the masking used in MAE. In the second stage, we pretrain the encoder using both the image and associated weak supervision. **This combination outperforms using either strategy in isolation, i.e., an MAE model or a weakly supervised model trained from scratch.**

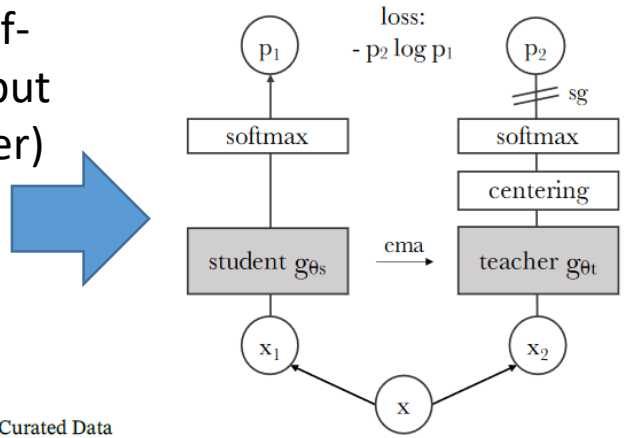
The effectiveness of MAE pre-pretraining for billion-scale pretraining. Meta AI, **2024.**

DINOv2: Learning Robust Visual Features without Supervision

Discriminative Self-supervised Pre-training

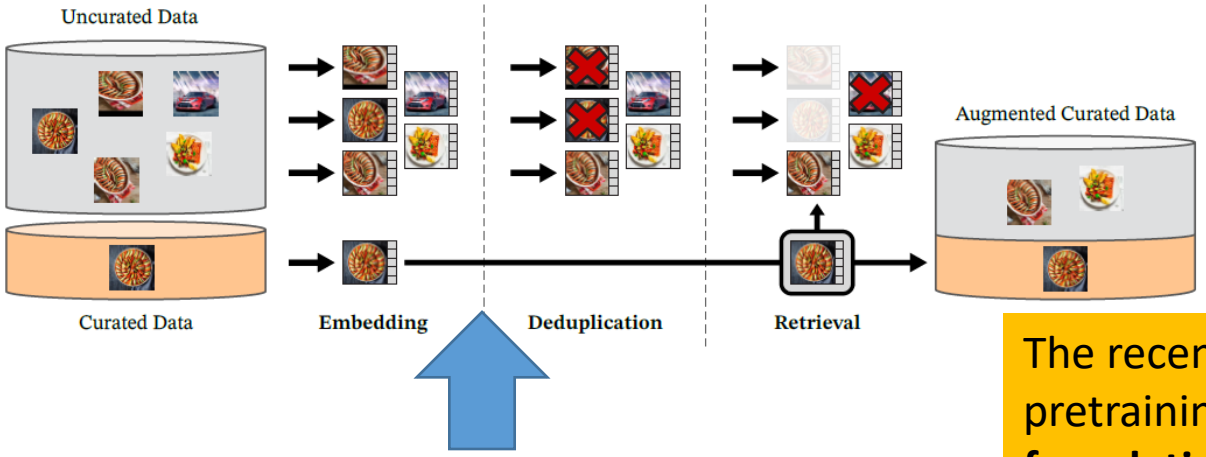
DINO: DIstillation with NO labels. DINO simplifies self-supervised ViT training by directly predicting the output of a teacher network (built with a momentum encoder) by using a standard cross-entropy loss.

Emerging Properties in Self-Supervised Vision Transformers, 2021



Discriminative Self-supervised Pre-training

We learn our features with a discriminative self-supervised method that can be seen as a **combination of DINO and iBOT** (Zhou et al., 2022), **losses with the centering of SwAV** (Caron et al., 2020). We also add a **regularizer** to spread features and a short **high-resolution training** phase.



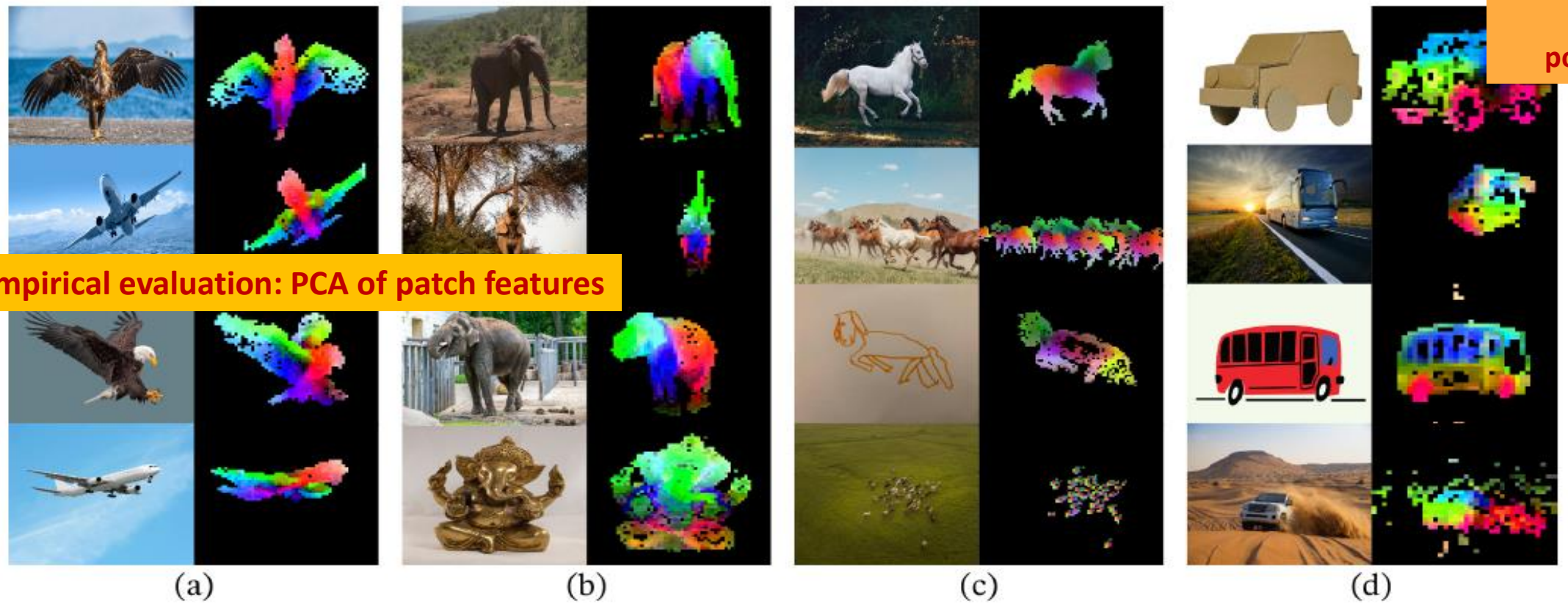
Overview of our data processing pipeline. Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting combination augments the initial dataset through a self-supervised retrieval system.

The recent breakthroughs in natural language processing for model pretraining on large quantities of data have opened the way for similar **foundation models in computer vision**. These models could greatly simplify the use of images in any system by producing generalpurpose visual features, i.e., **features that work across image distributions and tasks without finetuning**. This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit existing approaches and combine different techniques to scale our pretraining in terms of data and model size.

DINOv2: Learning Robust Visual Features without Supervision. Meta AI Research, **2024.**

DINOv2: Learning Robust Visual Features without Supervision

Фундаментальная модель должна иметь универсальные и робастные признаки



empirical evaluation: PCA of patch features

Foundation models in computer vision should have features that work across image distributions and tasks without finetuning!

Qualitative Results. PCA of patch features.
1) our unsupervised foreground / background detector, performs very well and is capable of delineating the boundary of the main object in the picture.
2) other components correspond to "parts" of objects and match well for images of the same category(emerging property – model was not trained to parse parts of objects).

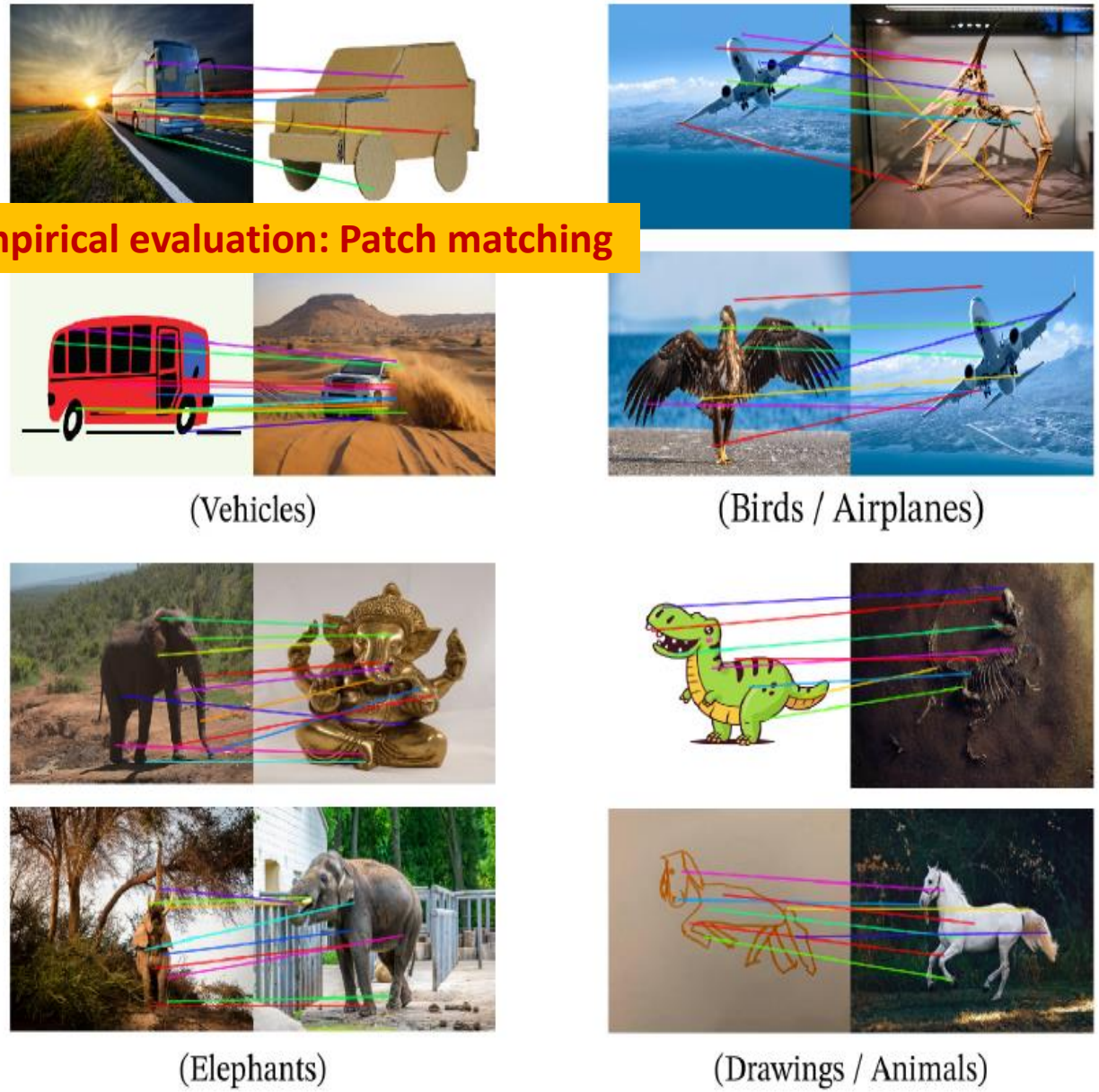
Visualization of the first PCA components. We compute a PCA between the patches of the images from the same column (a, b, c and d) and show their first 3 components. Each component is matched to a different color channel. Same parts are matched between related images despite changes of pose, style or even objects. **Background is removed by thresholding the first PCA component**.

DINOv2: Learning Robust Visual Features without Supervision. Meta AI Research, **2024.**

DINOv2: Learning Robust Visual Features without Supervision

Фундаментальная модель должна иметь универсальные и робастные признаки

empirical evaluation: Patch matching



Qualitative Results. Patch matching. Finally, we explore what type of information our patch-level features contain by matching them across images. We start by detecting the foreground object using the procedure described above. Then, we compute the euclidean distance between patch features extracted from two images and map them by solving an assignment problem. In order to reduce the number of matches, we then apply a non-maximum suppression.

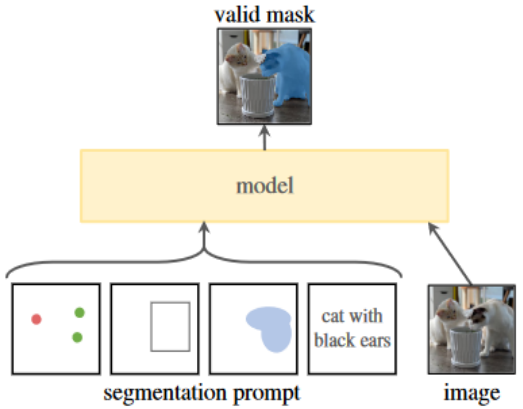
Matching across images. We match patch-level features between images from different domains, poses and even objects that share similar semantic information. **This exhibits the ability of our model to transfer across domains and understand relations between similar parts of different objects**.

Foundation models in computer vision should have features that work across image distributions and tasks without finetuning!

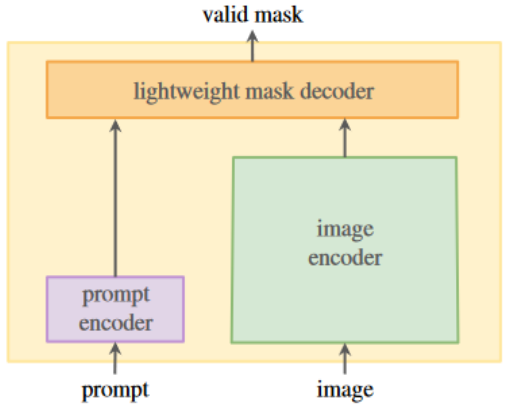
DINOv2: Learning Robust Visual Features without Supervision. Meta AI Research, **2024.**

SAM: Segment Anything, Foundation model for segmentation

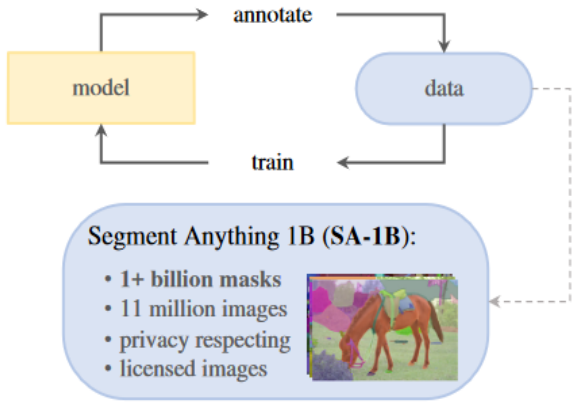
Задача сегментации столь же фундаментальна!



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)

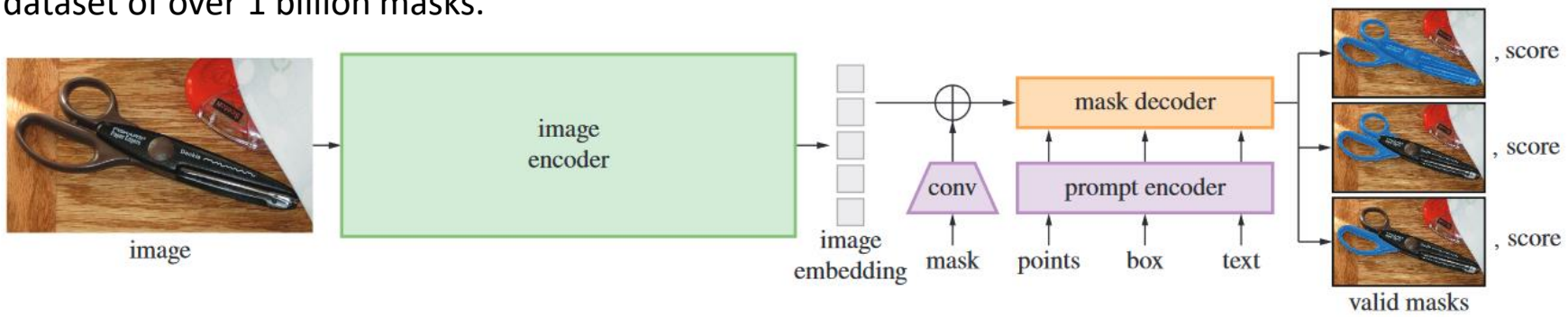


(c) **Data**: data engine (top) & dataset (bottom)

ХИТ ИИ 2023!

Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Foundation model for segmentation with three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables **zero-shot transfer to a range of tasks via prompt engineering**, and a data engine for collecting SA-1B, our dataset of over 1 billion masks.



Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

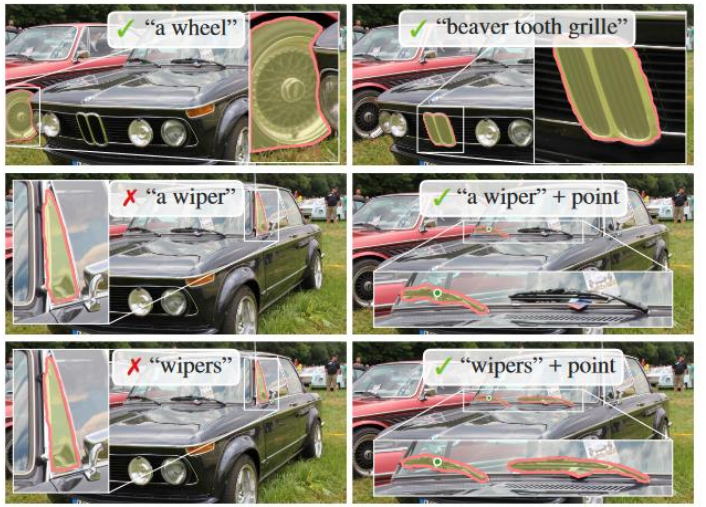
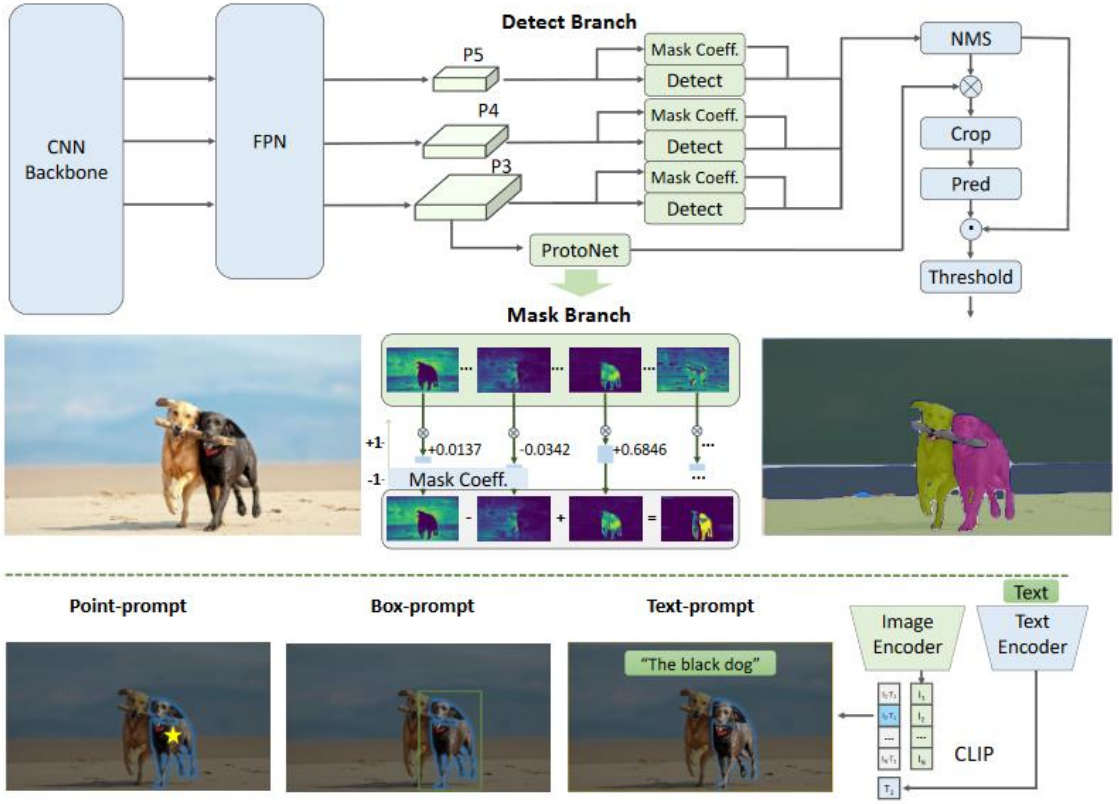


Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Segment Anything, Alexander Kirillov, …, Piotr Doll´ar, Ross Girshick, Meta AI Research, FAIR, 2023

Fast & Faster Segment Anything: SAM еще быстрее и компактнее

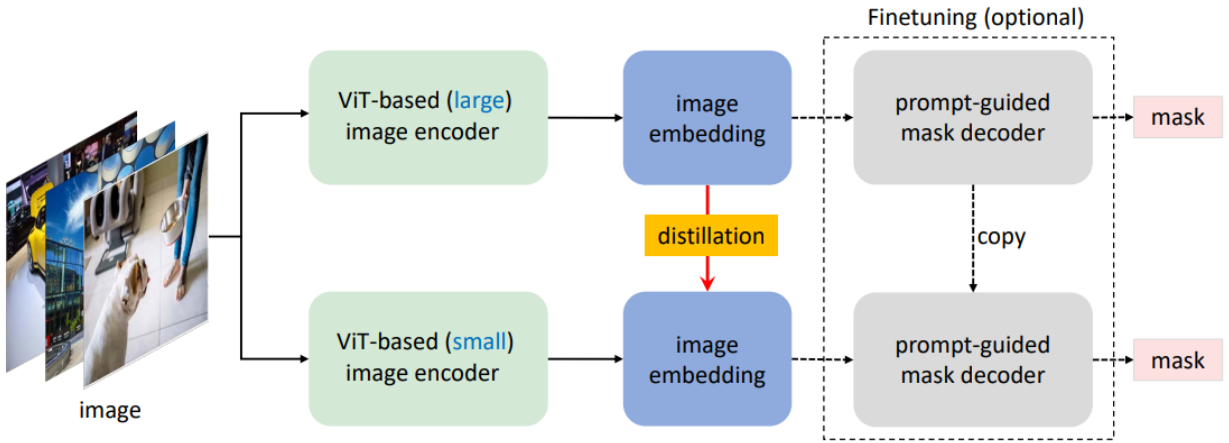


Авторы заменили тяжёлый энкодер ViT в SAM на свёрточный детектор **YOLOv8-seg**. Благодаря этому удалось добиться **ускорения в 50 раз**, сохранив качество предсказаний на сравнимом уровне. Метод FastSAM состоит из двух этапов: **сегментация всех объектов** на изображении и **выбор объекта согласно промпту**.

Fast Segment Anything, 2023

https://habr.com/ru/companies/sberdevices/articles/757606/

В MobileSAM авторы заменили ViT-H, энкодером ViT-S, уменьшенной версией ViT. Для обучения нового SAM авторы предложили **метод раздельной дистилляции в два этапа**: **дистилляция энкодера изображений и дообучение декодера масок**.

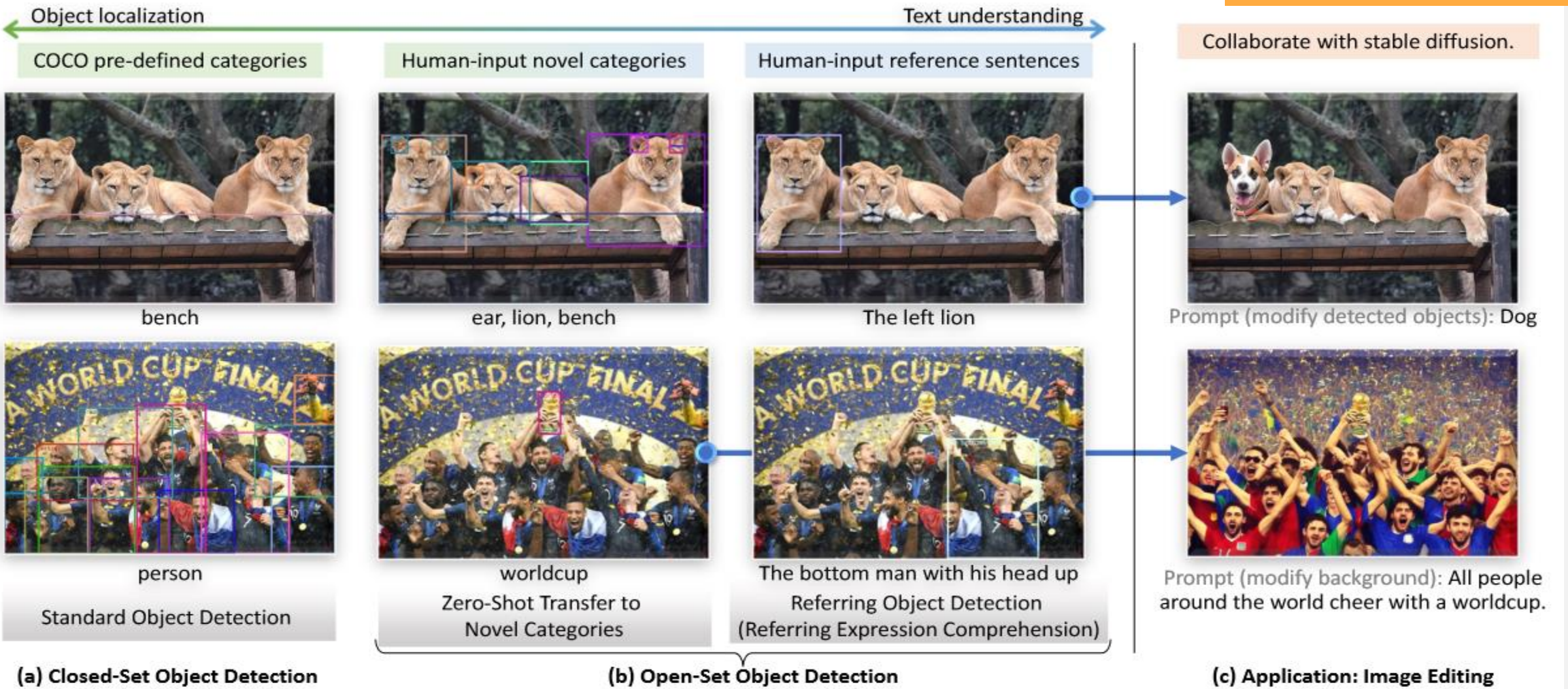


MobileSAM в 60+ раз меньше по размеру и при этом имеет качество, сравнимое с оригинальным SAM. MobileSAM **в 4 раза быстрее и в 7 раз меньше, чем FastSAM**. MobileSAM можно обучить на одном GPU менее чем за день, а инференс одной картинки занимает 10 мс.

Faster Segment Anything: Towards Lightweight SAM for Mobile Applications, 2023

Grounding DINO: Open-Set Object Detection

Задача обнаружения столь же фундаментальна!

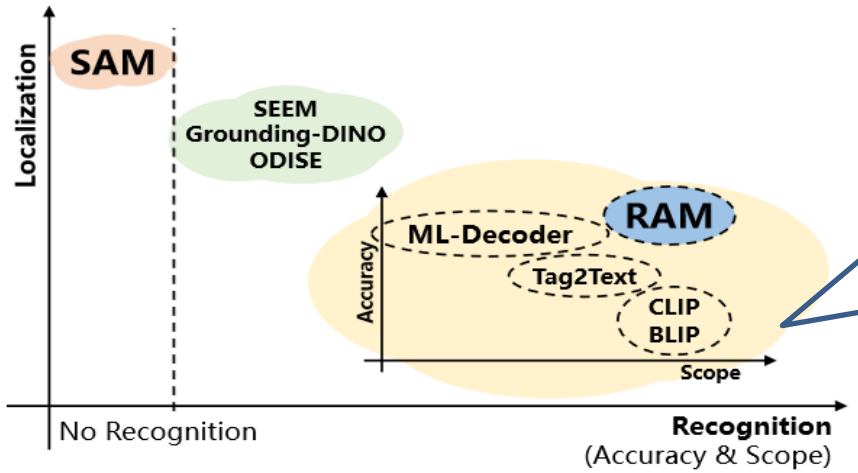


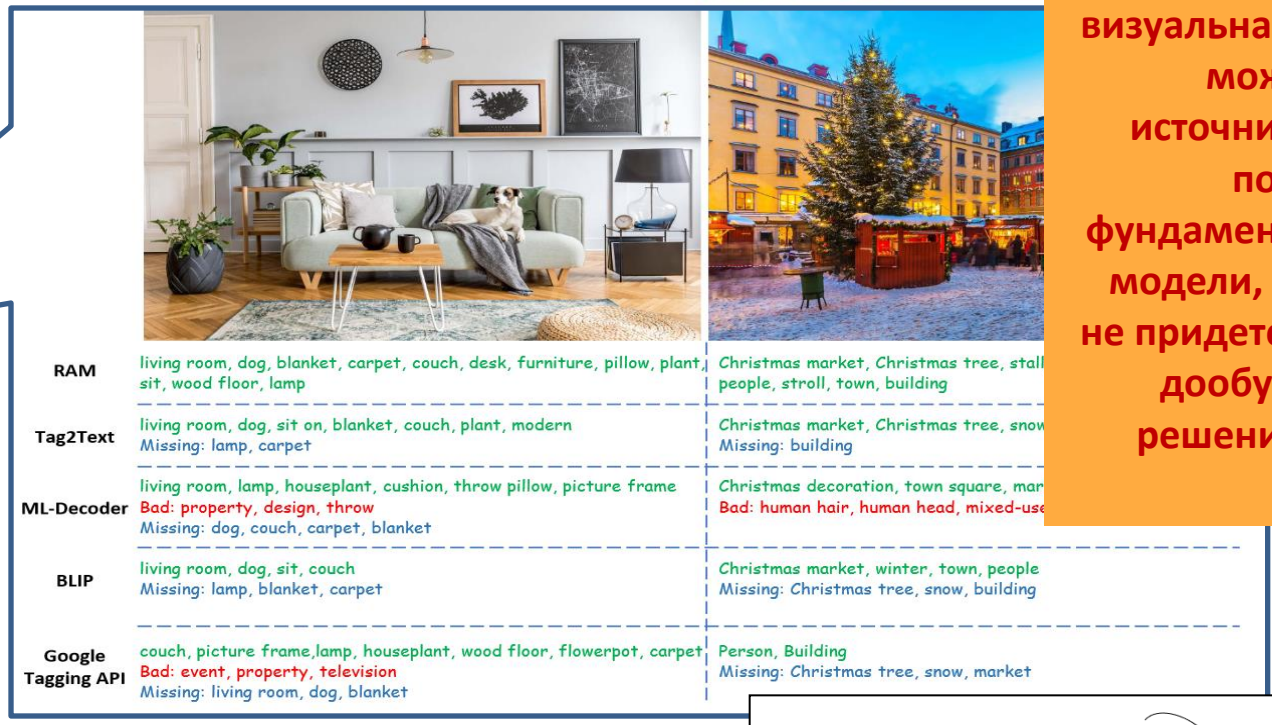
(a) Closed-set object detection requires models to detect objects of pre-defined categories. (b) Previous work zero-shot transfer models to novel categories for model generalization. We propose to add Referring expression comprehension (REC) as another evaluation for model generalizations on novel objects with attributes. (c) We present an image editing application by combining Grounding DINO and Stable Diffusion

Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, 2023

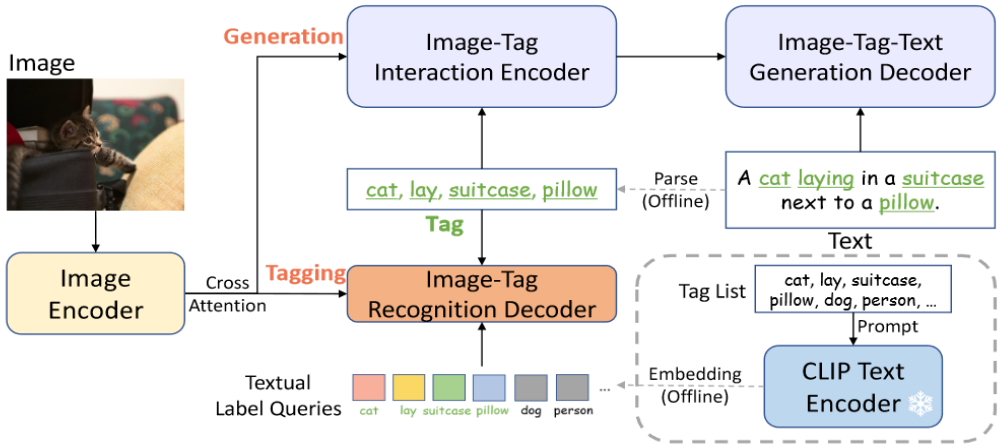
Recognize Anything: Image Tagging (Multi-label Image Recognition)

Возможно, любая сложная визуальная задача может стать источником для получения фундаментальной модели, которую не придется потом дообучать для решения новых задач

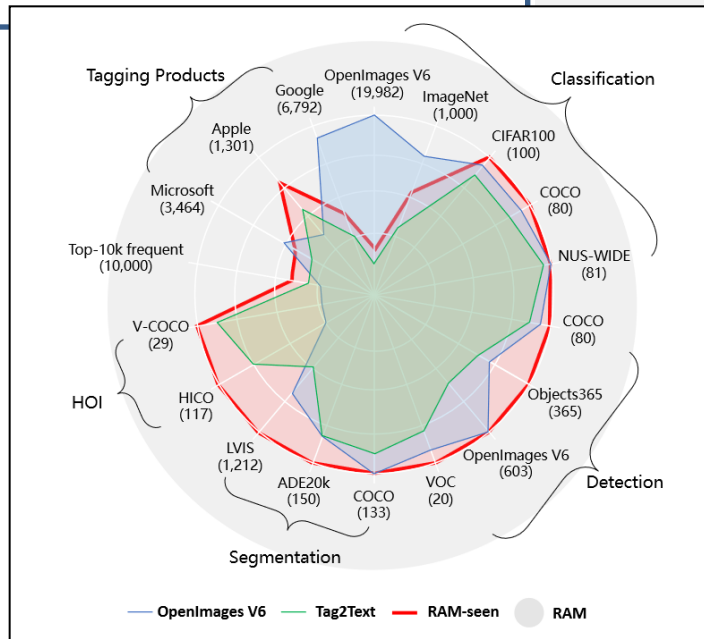




With image-tag-text triplets, RAM unifies the captioning and tagging tasks. RAM introduces an off-the-shelf text encoder to **encoder tags into textual label queries** with semantically-rich context, empowering the **generalization to unseen categories** in training stage.



Recognition Scopes of different tagging models. Tag2Text recognizes 3,400+ fixed tags. RAM upgrades the number to 6,400+, covering more valuable categories than OpenImages V6. With open-set capability, RAM is feasible to recognize any common category.



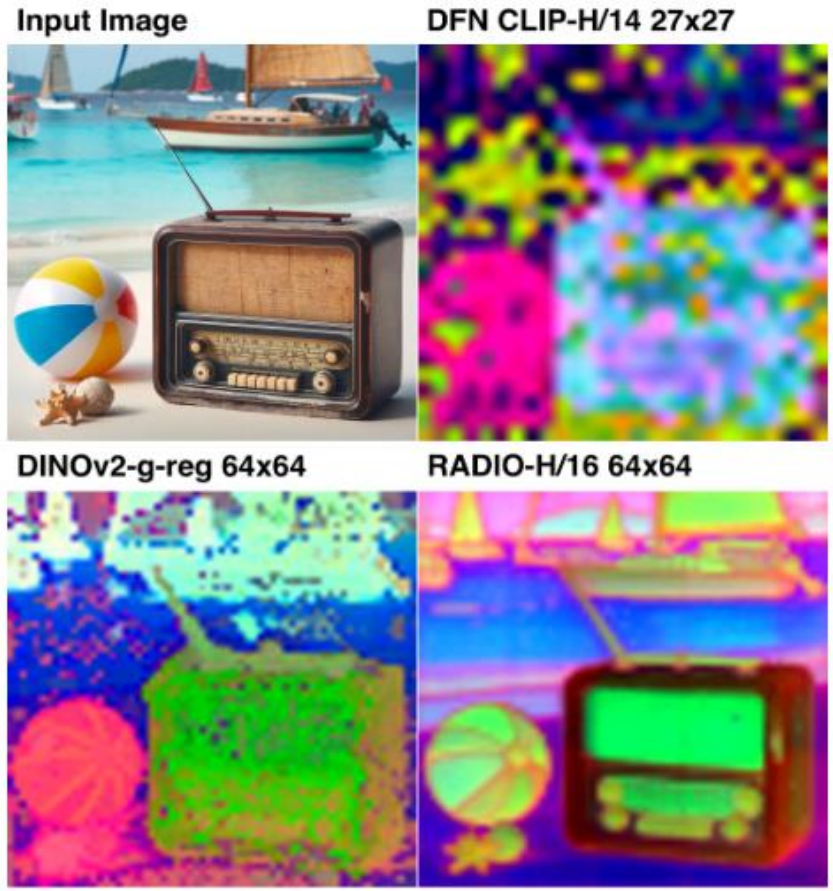
Recognize Anything: A Strong Image Tagging Model, 2023

AM-RADIO (апрель 2024): Vision Foundation Model for All Domains

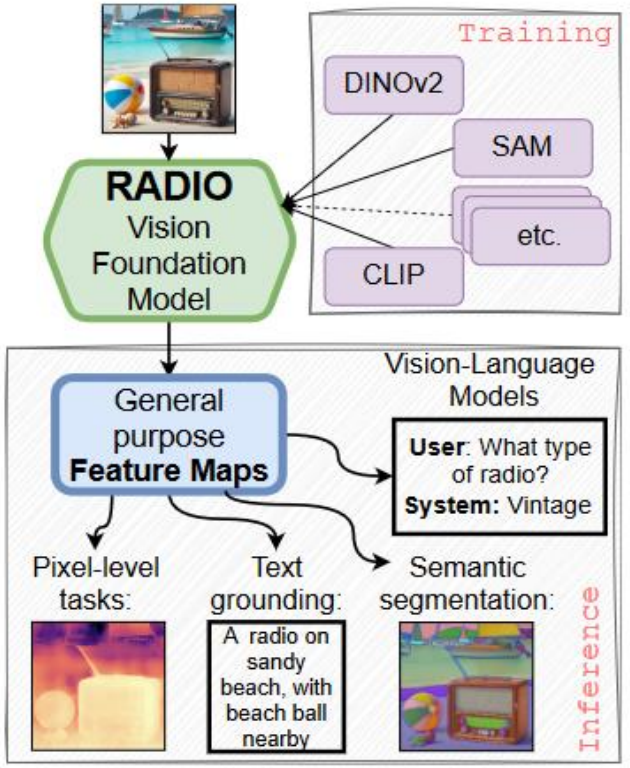
Все уровни, все признаки!

Универсальная фундаментальная модель для задач CV!

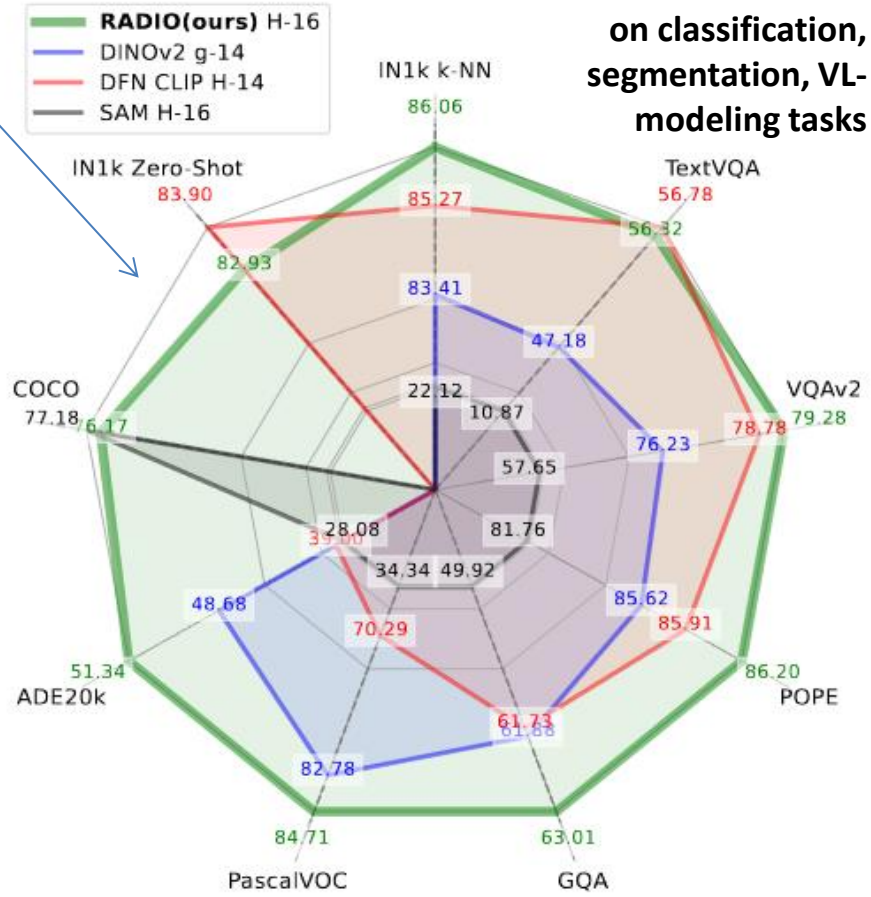
PCA feature visualization of different models



the overview of the AM-RADIO framework



benchmarks on classification, segmentation, VL-modeling tasks

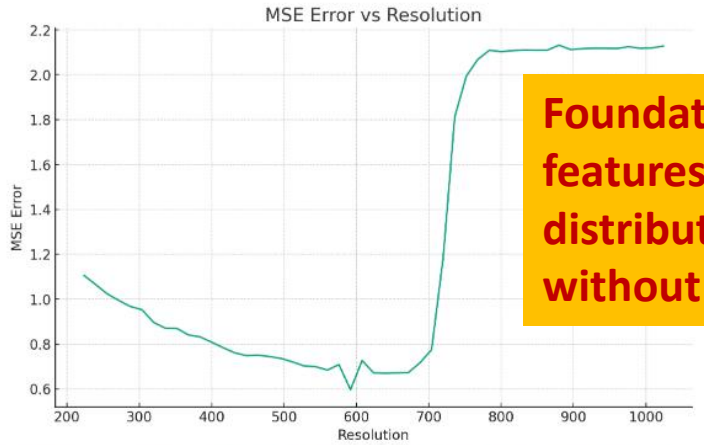


AM-RADIO is a framework to distill multiple pretrained vision foundation models, such as CLIP, DINOv2, SAM, into a single model that we call RADIO. As a result, a single vision foundation model agglomerates unique properties of the original models. This unifying approach obtains state-of-the-art feature representations in a single forward pass while also enabling unique properties such as **zero-shot classification (CLIP) or open set instance segmentation (SAM) at negligible additional cost**. Our proposed RADIO model can **process any resolution and aspect ratio, and produces semantically rich dense encodings**.

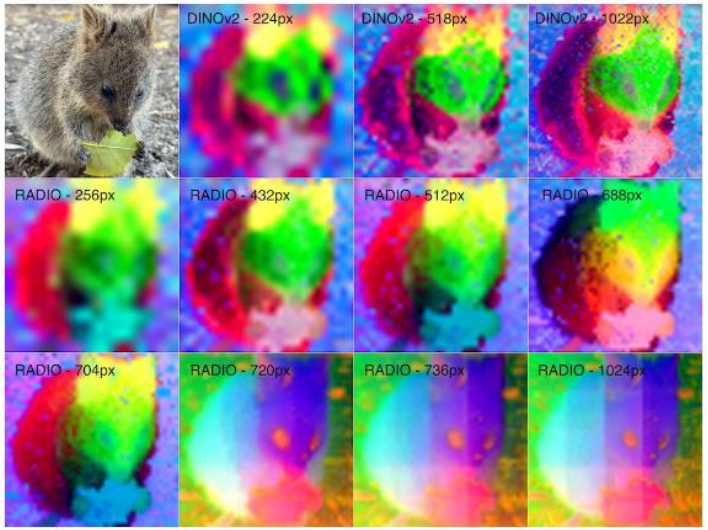
AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One, NVIDIA, 2024

AM-RADIO (апрель 2024): Vision Foundation Model for All Domains

А возможно, лучше все сразу!



Foundation models in CV should have features that work across image distributions, tasks *and resolutions* without fine-tuning!



RADIO "mode switches" when resolution is increased. In the plot, we show the MSE error between the RADIO features coming from its DINOv2 head at different resolutions.

Model	Resolution
DINOv2-g	518
SAM-H	1024
RADIO	512
	1024

Images

AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One, NVIDIA, 2024

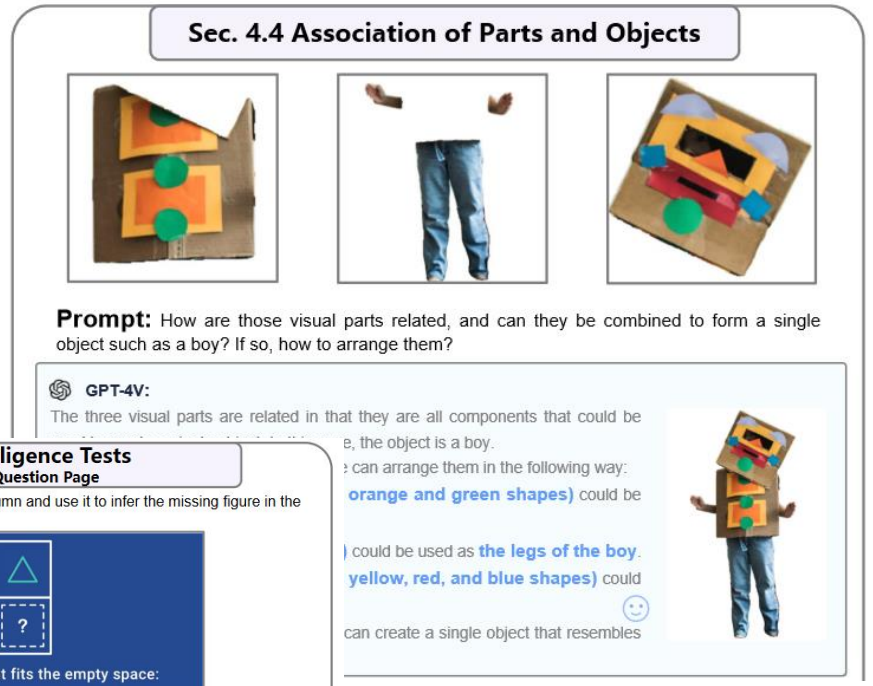
Multimodal Large Language Models

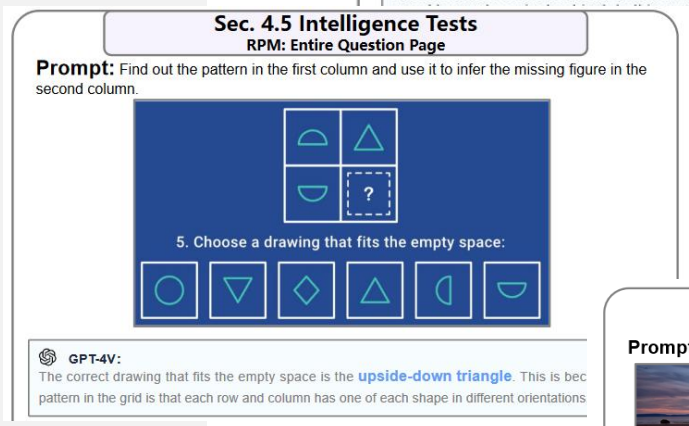
(другой уровень VL-задач)

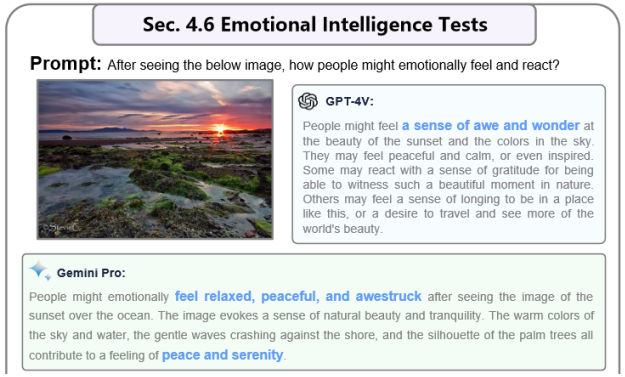
Уровни анализа в VL моделях

Уровень модели мира	**когда нужно не только видеть сцену, но и понимать мир**
Уровень сцены	**классификация и описание сцен (отношения, действия), VQA**
Уровень объектов	**обнаружение объектов, семантическая сегментация…**
Уровень признаков	**особые точки, линии, углы, области, контуры, текстура…**
Уровень изображения	**пикселы (положение, яркость, цвет)**

Gemini vs GPT-4V: Comparison of Vision-Language Models

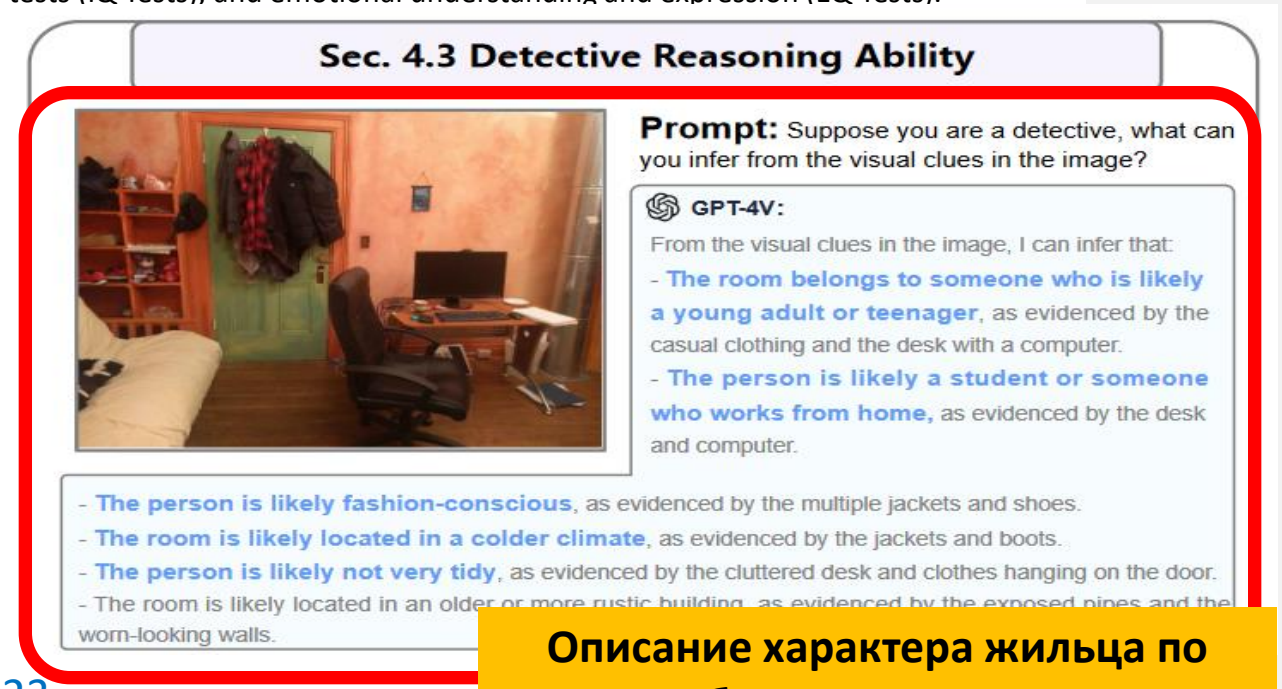






Multimodal evaluation

- **Image Recognition and Understanding:** Sec. 2 addresses the fundamental recognition and comprehension of image content without involving further inference, including tasks such as identifying landmarks, foods, logos, abstract images, autonomous driving scenes, misinformation detection, spotting differences, and object counting.
- **Text Recognition and Understanding in Images:** Sec. 3 concentrates on text recognition (including OCR) within images, such as scene text, mathematical formulas, and chart & table text recognition. Similarly, no further inference of text content is performed here.
- **Image Inference Abilities:** Beyond basic image recognition, Sec. 4 involves more advanced reasoning. This includes understanding humor and scientific concepts, as well as logical reasoning abilities like detective work, image combinations, look for patterns in intelligence tests (IQ Tests), and emotional understanding and expression (EQ Tests).



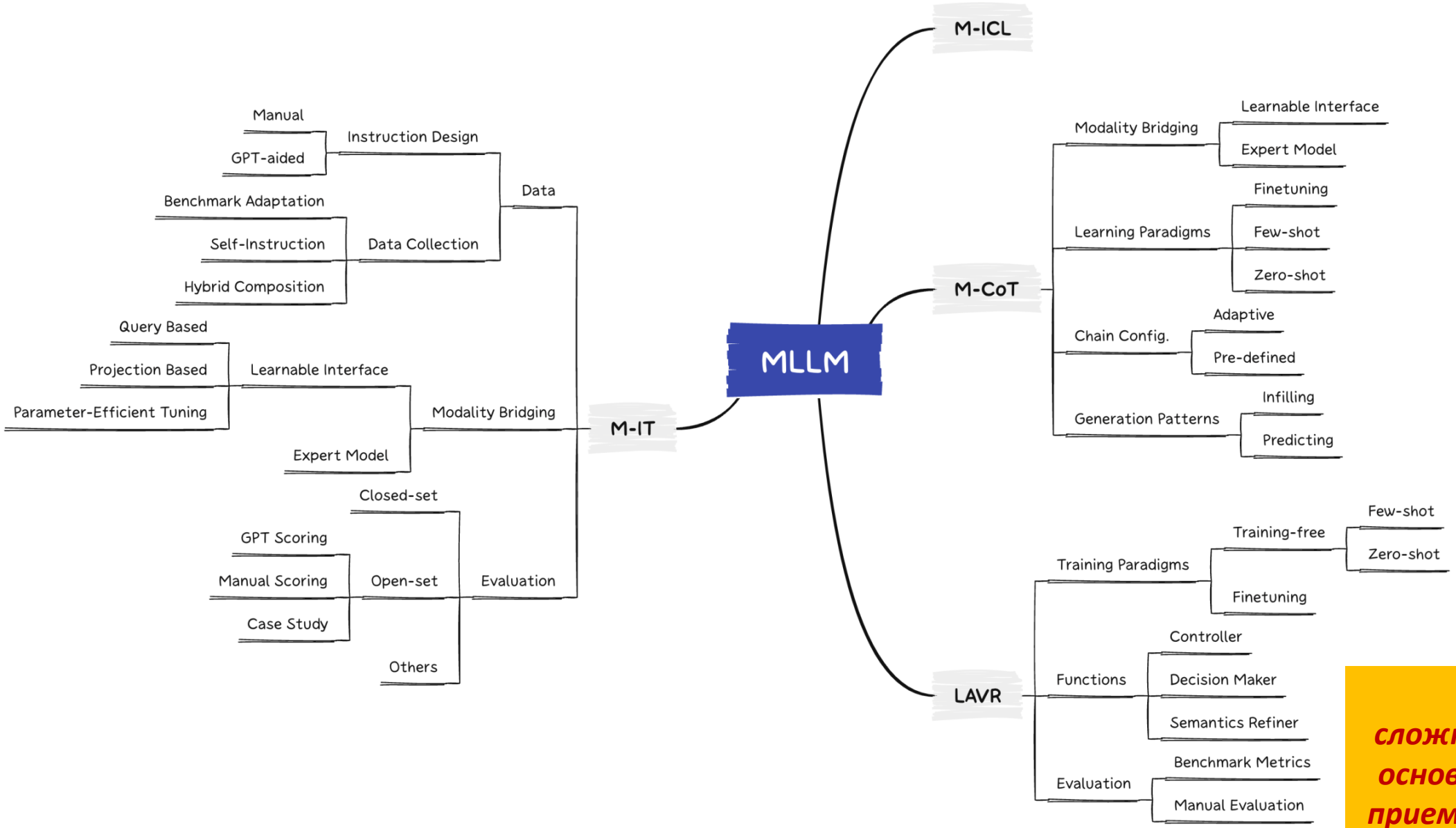
Описание характера жильца по изображению комнаты

OpenAI. Gpt-4v(ision) system card. Technical report, OpenAI, 2023.

Google Gemini Team. Gemini: A family of highly capable multimodal models. 2023.

Gemini vs GPT-4V : A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases, 2023

MLLM: Awesome-Multimodal-Large-Language-Models

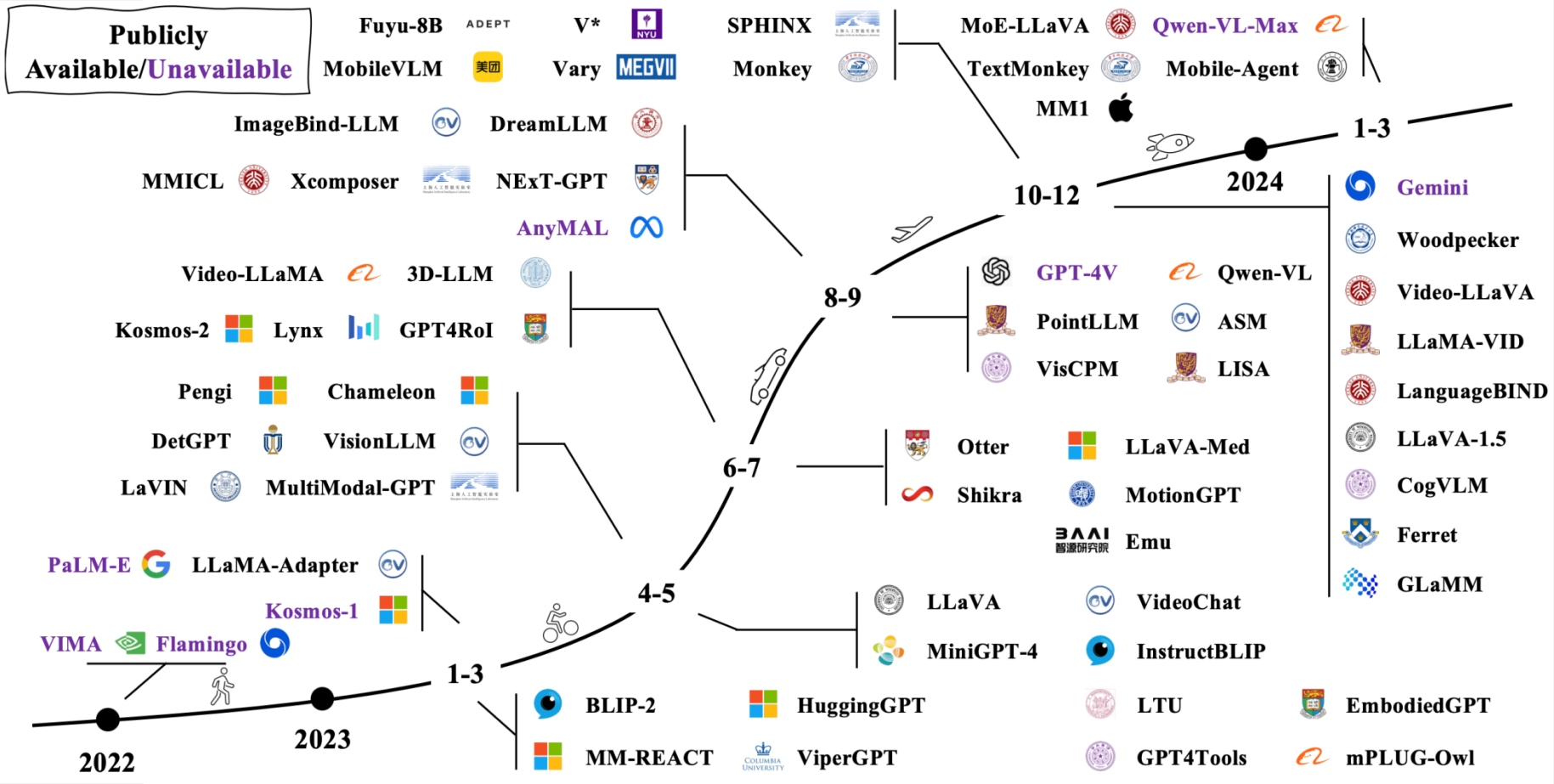


A Survey on Multimodal Large Language Models, 2024

MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, 2024

MLLM: сложная технология, основанная на общих приемах работы LLM, о которых см. в отдельном треке

MLLM: Awesome-Multimodal-Large-Language-Models



- Awesome Papers
  - Multimodal Instruction Tuning
  - Multimodal In-Context Learning
  - Multimodal Chain-of-Thought
  - LLM-Aided Visual Reasoning
  - Foundation Models
  - Evaluation
  - Multimodal Hallucination
  - Multimodal RLHF
  - Others
- Awesome Datasets
  - Datasets of Pre-Training for Alignment
  - Datasets of Multimodal Instruction Tuning
  - Datasets of In-Context Learning
  - Datasets of Multimodal Chain-of-Thought
  - Datasets of Multimodal RLHF
  - Benchmarks for Evaluation
  - Others

Latest Papers and Datasets on Multimodal Large Language Models, and Their Evaluation

https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

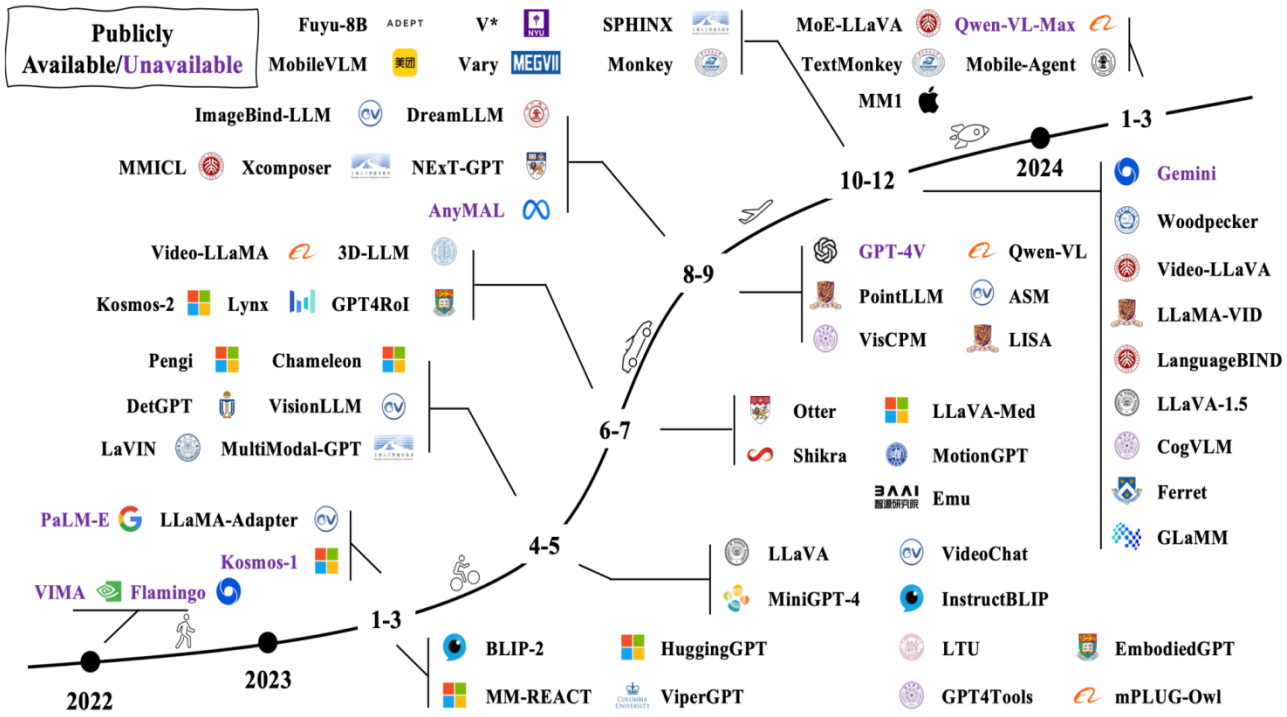
A Survey on Multimodal Large Language Models, 2024

MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, 2024

Теперь есть одна страница, где стоит читать про MLLM!

Ландшафт MLLM (2024)

Современный ландшафт LLM…



…постоянно меняется!

May 13, 2024

Hello GPT-4o

We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

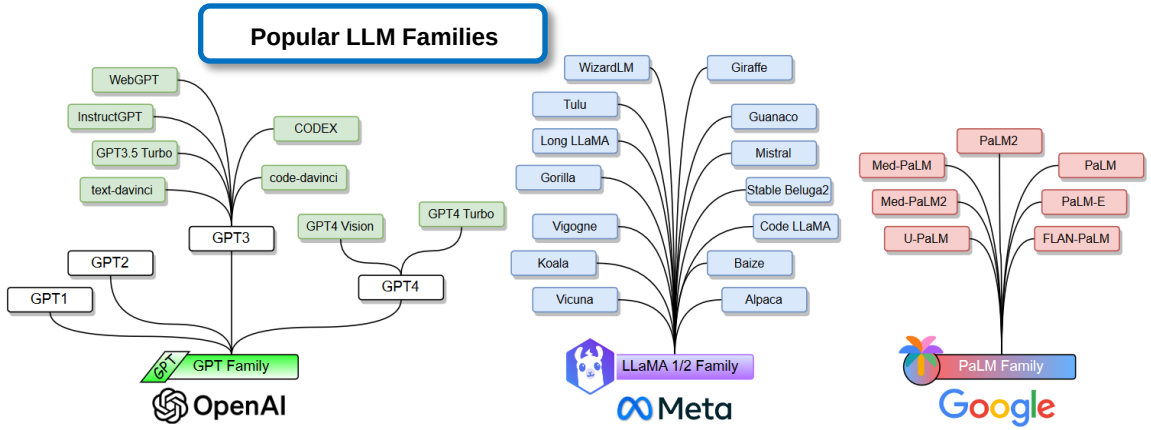
As measured on traditional benchmarks, **GPT-4o achieves GPT-4 Turbo**-level performance on text, reasoning, and coding intelligence, while **setting new high watermarks on multilingual, audio, and vision capabilities**.

Eval Sets	GPT-4o	GPT-4T 2024-04-09	Gemini 1.0 Ultra	Gemini 1.5 Pro	Claude Opus
MMMU (%) (val)	69.1	63.1	59.4	58.5	59.4
MathVista (%) (testmini)	63.8	58.1	53.0	52.1	50.5
AI2D (%) (test)	94.2	89.4	79.5	80.3	88.1
ChartQA (%) (test)	85.7	78.1	80.8	81.3	80.8
DocVQA (%) (test)	92.8	87.2	90.9	86.5	89.3
ActivityNet (%) (test)	61.9	59.5	52.2	56.7	
EgoSchema (%) (test)	72.2	63.9	61.5	63.2	

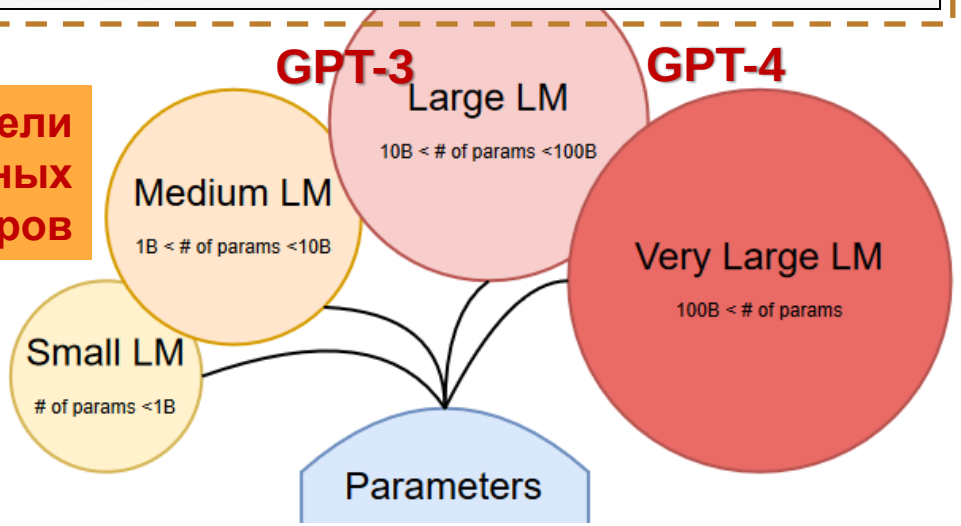
Vision understanding evals - GPT-4o achieves SOTA on visual perception benchmarks.

https://openai.com/index/hello-gpt-4o/

GPT-4o | GPT-4T 2024-04-09 | Gemini 1.0 Ultra | Gemini 1.5 Pro | Claude Opus



Модели разных размеров



Large Language Models: A Survey, 2024

CNN vs. Transformers:

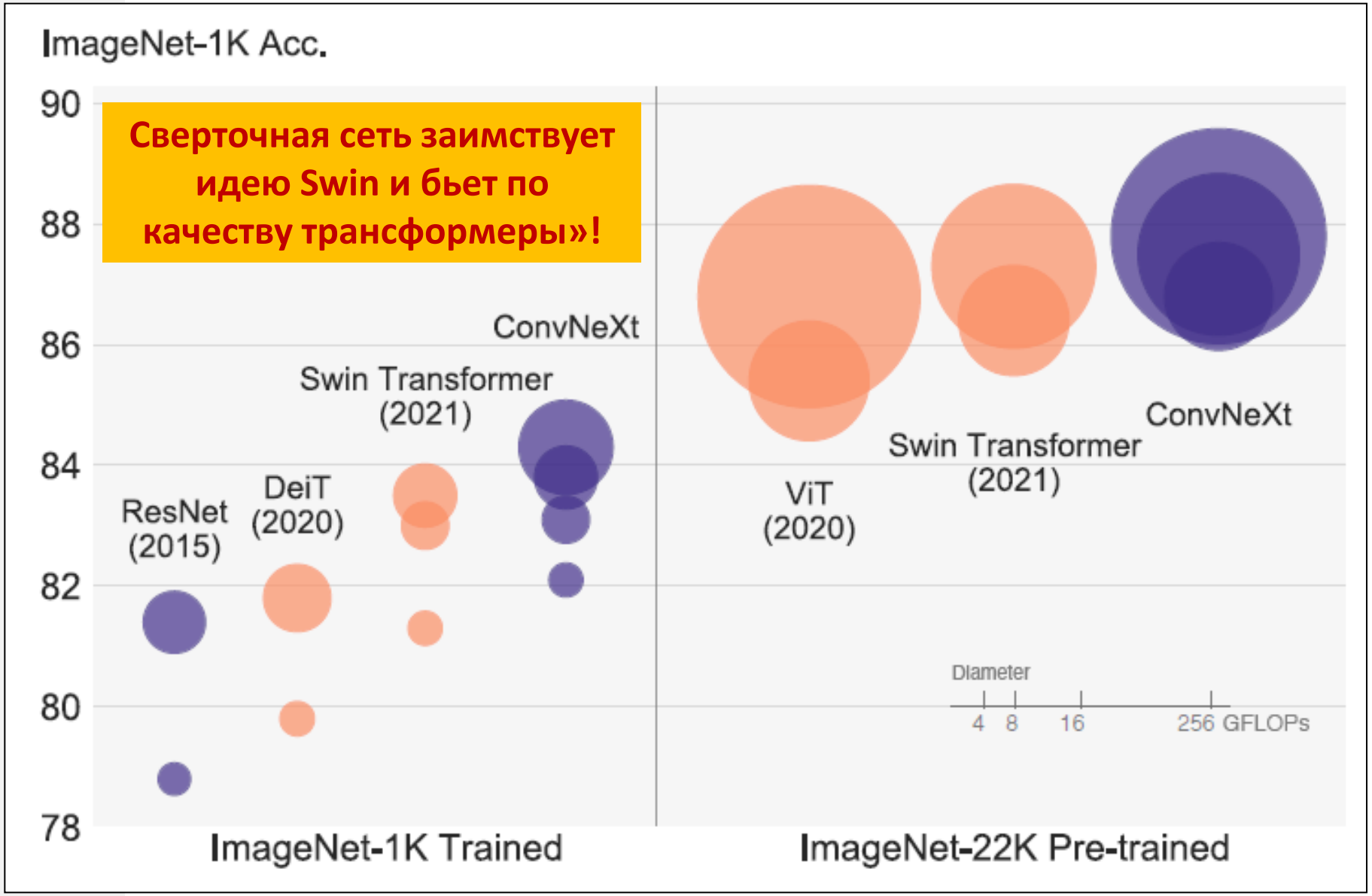
ничего еще не ясно!

(сверточные и гибридные архитектуры не хуже трансформеров)

ConvNeXt | **ConvNeXt V2** | **Large Kernels (RepLKNet)**

Sparse Large Kernel (SLaK) | **Auto-ML** | **MathNAS** | **DeepMAD**

Transformers vs. CNN: ничего еще не ясно! Например, ConvNeXt…

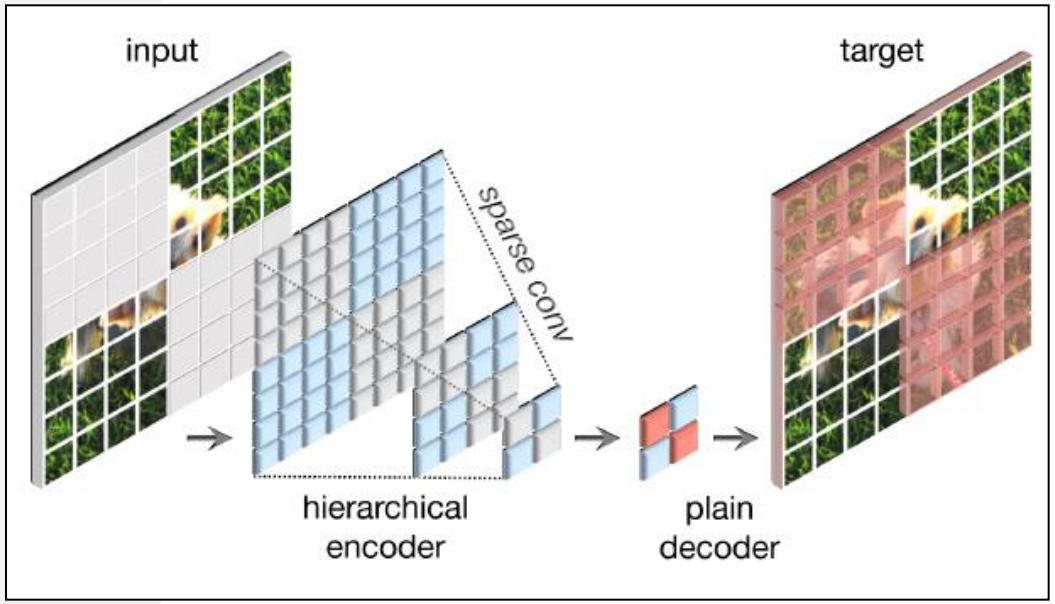


ImageNet-1K classification results for ConvNets and vision Transformers.
Each bubble's area is proportional to FLOPs of a variant in a model family.
ImageNet-1K/22K models here
take 224^2/384^2 images respectively.
ResNet and ViT results were obtained with improved training procedures over the original papers.
We demonstrate that a standard ConvNet model can achieve the same level of scalability as hierarchical vision Transformers while being much simpler in design.

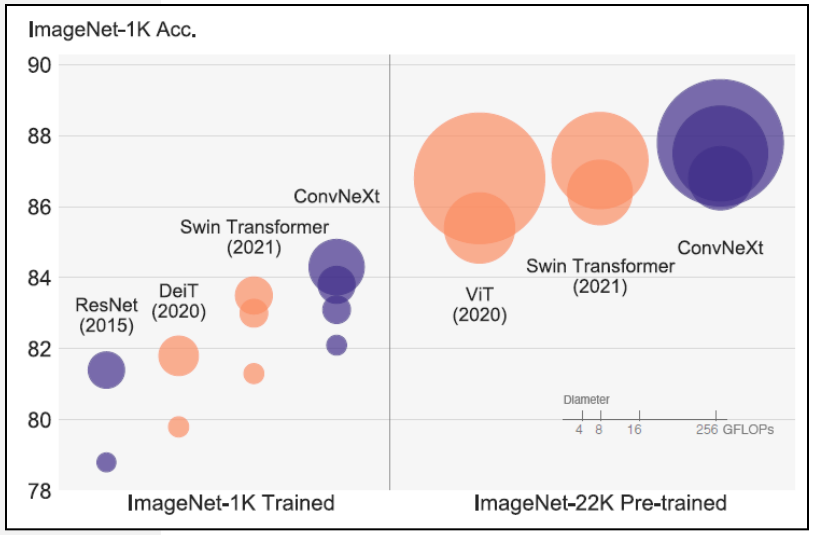
A ConvNet for the 2020s, Facebook AI Research (FAIR), 2022

https://github.com/facebookresearch/ConvNeXt

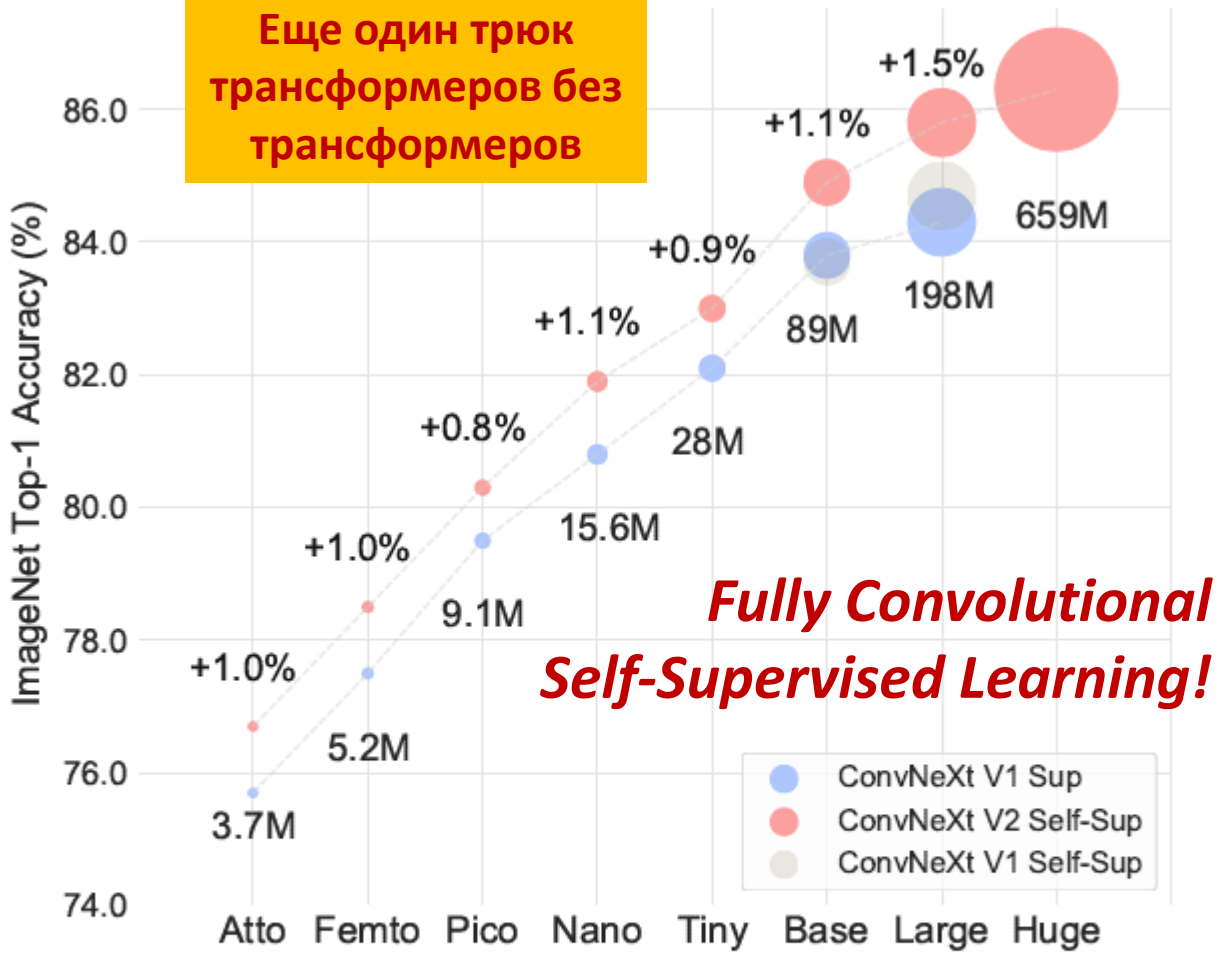
Transformers vs. CNN: ничего еще не ясно! ConvNeXt V2…(2023)



Fully Convolutional Masked Autoencoder



Еще один трюк трансформеров без трансформеров

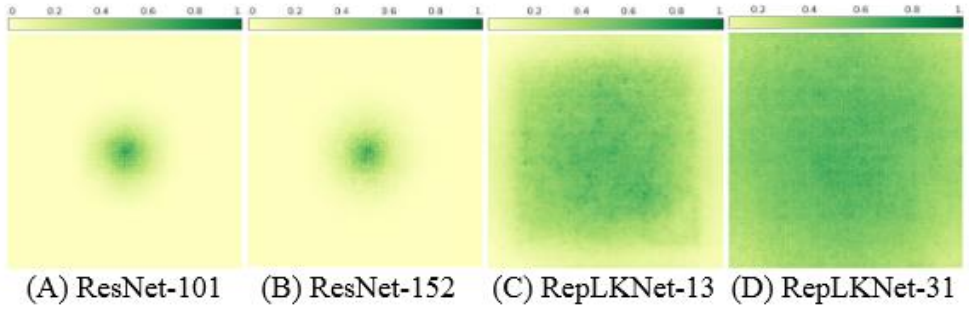


Fully Convolutional Self-Supervised Learning!

Figure 1. **ConvNeXt V2 model scaling**. The ConvNeXt V2 model, which has been pre-trained using our fully convolutional masked autoencoder framework, performs significantly better than the previous version across a wide range of model sizes.

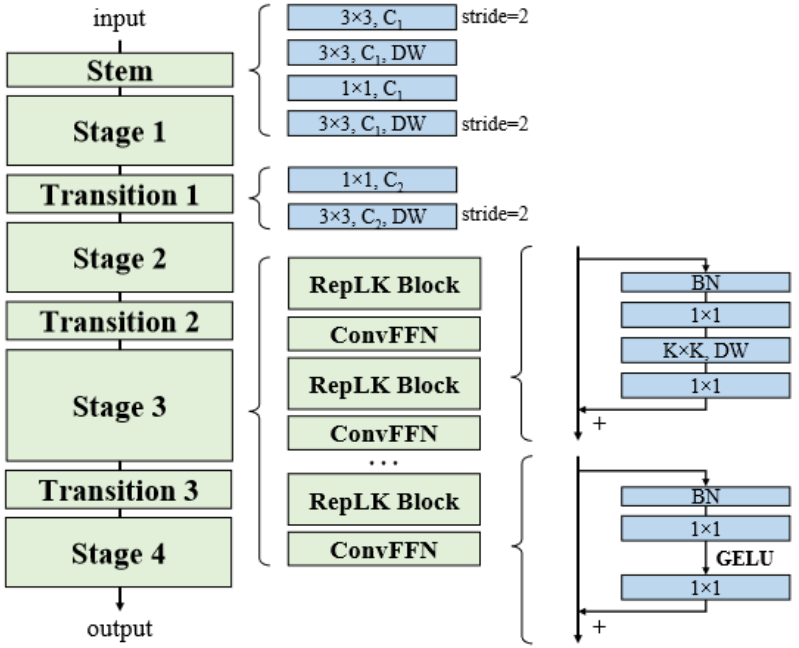
ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, KAIST, Meta AI, FAIR, 2023

Transformers vs. CNN: ничего еще не ясно! Revisiting Large Kernels (RepLKNet)



*The **Effective Receptive Field** (ERF) of ResNet-101/152 and RepLKNet-13/31 respectively. A more widely distributed dark area indicates a larger ERF. More layers (e.g., from ResNet-101 to ResNet-152) help little in enlarging ERFs. Instead, our large kernel model **RepLKNet** effectively obtains large ERFs.
Briefly, we produce an aggregated contribution score matrix A (1024×1024), where each entry a (0 ≤ a ≤ 1) measures the contribution of the corresponding pixel on the input image to the central point of the feature map produced by the last layer.

*Scaling Up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs, Ding et al., 2022



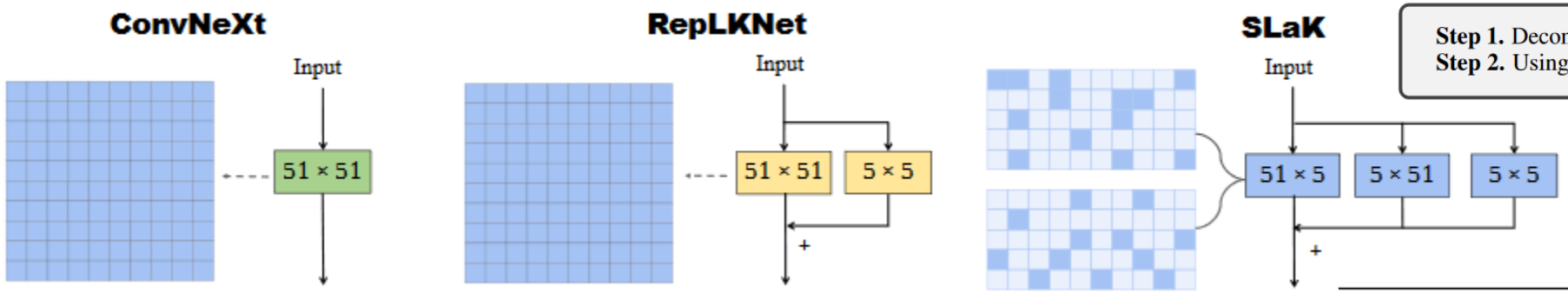
RepLKNet comprises Stem, Stages and Transitions. Except for depth-wise (DW) large kernel, the other components include DW 3×3, dense 1×1 conv, and batch normalization (BN). Note that every conv layer has a following BN, which are not depicted. Such conv-BN sequences use ReLU as the activation function, except those before the shortcut-addition and those preceding GELU.

Model	Input resolution	Top-1 acc	Params (M)	FLOPs (G)	Throughput examples/s
RepLKNet-31B	224×224	83.5	79	15.3	295.5
Swin-B	224×224	83.5	88	15.4	226.2
RepLKNet-31B	384×384	84.8	79	45.1	97.0
Swin-B	384×384	84.5	88	47.0	67.9
RepLKNet-31B ‡	224×224	85.2	-	-	-
Swin-B ‡	224×224	85.2	-	-	-
RepLKNet-31B ‡	384×384	86.0	-	-	-
Swin-B ‡	384×384	86.4	-	-	-
RepLKNet-31L ‡	384×384	86.6	172	96.0	50.2
Swin-L ‡	384×384	87.3	197	103.9	36.2
RepLKNet-XL ◇	320×320	87.8	335	128.7	39.1

ImageNet results.
The throughput is tested with FP32 and a batch size of 64 on 2080Ti. ‡ indicates ImageNet-22K pretrain-ing.
◊ indicates pretrained with extra data.

Свертка с большими ядрами может эффективно моделировать свойства трансформеров!

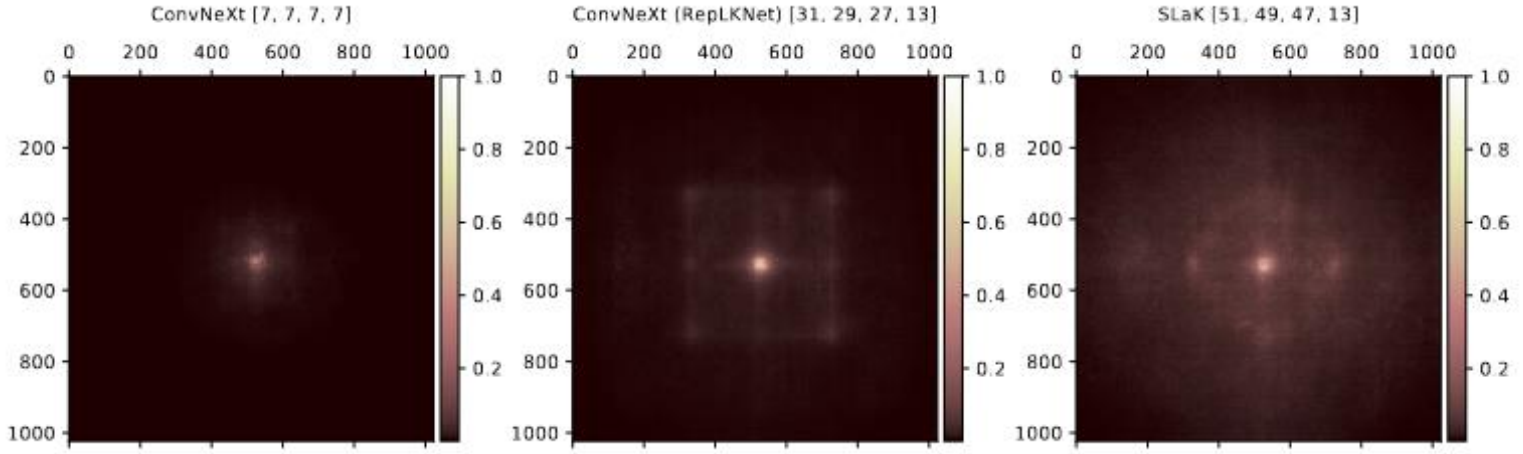
Transformers vs. CNN: ничего еще не ясно! Sparse Large Kernel Network (SLaK)



Step 1. Decomposing a large kernel into two rectangular, parallel kernels.
Step 2. Using *sparse* groups, expanding more width.

Разреженные ядра свертки могут быть еще больше и еще эффективнее

Large depth-wise kernel (e.g., 51×51) paradigms of ConvNeXt, RepLKNet, and SLaK.
Dark blue squares refer to the dense weights in convolutional kernels.
Light blue squares refer to the sparse weights in convolutional kernels.



Effective receptive field (ERF) of models with various kernel sizes. SLaK is not only able to capture long-range dependence but also the local context features

Model	Image Size	#Param.	FLOPs	Top-1 Acc
ResNet-50 (He et al., 2016)	224×224	26M	4.1G	76.5
ResNeXt-50-32×4d (Xie et al., 2017)	224×224	25M	4.3G	77.6
ResMLP-24 (Touvron et al., 2021a)	224×224	30M	6.0G	79.4
DeiT-S (Touvron et al., 2021b)	224×224	22M	4.6G	79.8
Swin-T (Liu et al., 2021e)	224×224	28M	4.5G	81.3
TNT-S (Han et al., 2021a)	224×224	24M	5.2G	81.3
T2T-ViT$_t$-14 (Yuan et al., 2021a)	224×224	22M	6.1G	81.7
ConvNeXt-T (Liu et al., 2022b)	224×224	29M	4.5G	82.1
SLaK-T	224×224	30M/50M	5.0G/8.7G	**82.5**
Mixer-B/16 (Tolstikhin et al., 2021)	224×224	59M	11.6G	76.4
ResNet-101 (He et al., 2016)	224×224	45M	7.9G	77.4
ResNeXt101-32x4d (Xie et al., 2017)	224×224	44M	8.0G	78.8
PVT-Large (Wang et al., 2021b)	224×224	61M	9.8G	81.7
T2T-ViT$_t$-19 (Yuan et al., 2021a)	224×224	39M	9.8G	82.4
Swin-S (Liu et al., 2021e)	224×224	50M	8.7G	83.0
ConvNeXt-S (Liu et al., 2022b)	224×224	50M	8.7G	83.1
SLaK-S	224×224	55M/91M	9.8G/16.7G	**83.8**
DeiT-Base/16 (Touvron et al., 2021b)	224×224	87M	17.6G	81.8
RepLKNet-31B (Ding et al., 2022)	224×224	79M	15.3G	83.5
Swin-B (Liu et al., 2021e)	224×224	88M	15.4G	83.5
ConvNeXt-B (Liu et al., 2022b)	224×224	89M	15.4G	83.8
SLaK-B	224×224	95M/158M	17.1G/28.5G	**84.0**
ViT-Base/16 (Dosovitskiy et al., 2021)	384×384	87M	55.4G	77.9
DeiT-B/16 (Touvron et al., 2021b)	384×384	86M	55.4G	83.1
Swin-B (Liu et al., 2021e)	384×384	88M	47.1G	84.5
RepLKNet-31B (Ding et al., 2022)	384×384	79M	45.1G	84.8
ConvNeXt-B (Liu et al., 2022b)	384×384	89M	45.0G	85.1
SLaK-B	384×384	95M/158M	50.3G/83.8G	**85.5**

Classification accuracy on ImageNet-1K. For SLaK models both theoretical, sparsity-aware numbers parameter & FLOPs (in black), and measured = no sparsity-aware acceleration (in blue).

***MORE CONVNETS IN THE 2020S: SCALING UP KERNELS BEYOND 51 × 51 USING SPARSITY, Liu et al., 2023**

Transformers vs. CNN: ничего еще не ясно! Math CNN Design (DeepMAD)

Theorem 1. *The normalized Gaussian entropy upper bound of the MLP $f(\cdot)$ is*

$$H_f = w_{L+1} \sum_{i=1}^{L} \log(w_i). \tag{1}$$

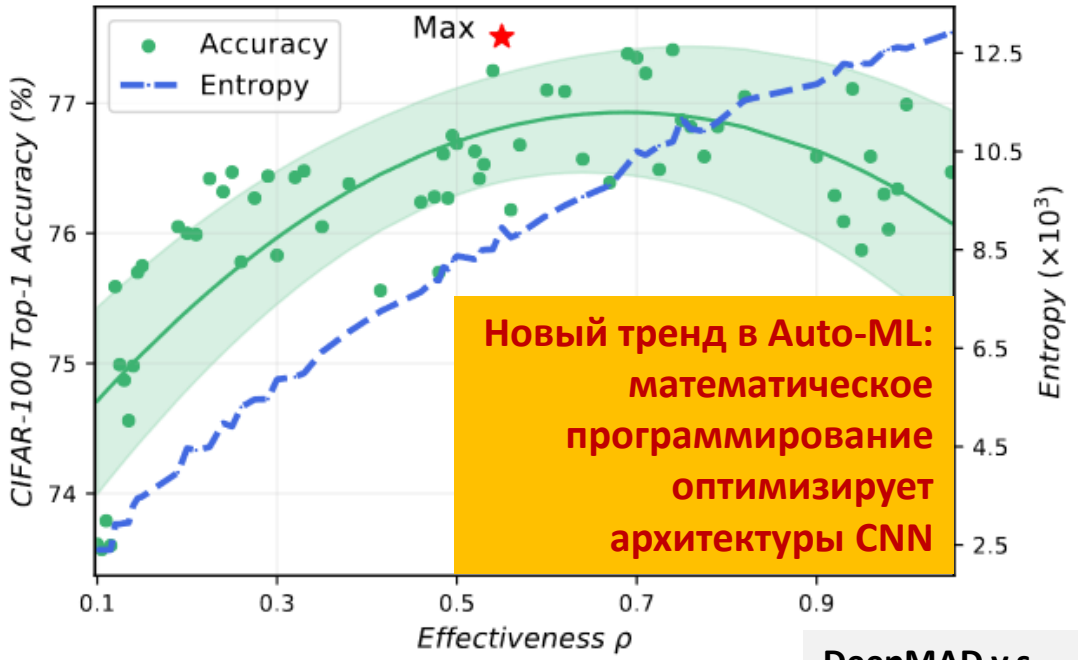
4.3. Final DeepMAD Formula

We gather everything together and present the final mathematical programming problem for DeepMAD. Suppose that we aim to design an L-layer CNN model $f(\cdot)$ with M stages. The entropy of the i-th stage is denoted as H_i defined in Eq. (4). Within each stage, all blocks use the same structural parameters (width, kernel size, etc.). The width of each CNN layer is defined by $w_i = c_i k_i^2 / g_i$. The depth of each stage is denoted as L_i for $i = 1, 2, \cdots, M$. We propose to optimize $\{w_i, L_i\}$ via the following mathematical programming (MP) problem:

$$\begin{aligned} \max_{w_i, L_i} \quad & \sum_{i=1}^{M} \alpha_i H_i - \beta Q, \\ \text{s.t.} \quad & L \cdot (\prod_{i=1}^{L} w_i)^{-1/L} \le \rho_0, \\ & \text{FLOPs}[f(\cdot)] \le \text{budget}, \\ & \text{Params}[f(\cdot)] \le \text{budget}, \\ & Q \triangleq \exp[\text{Var}(L_1, L_2, \cdots, L_M)], \\ & w_1 \le w_2 \le \cdots \le w_L. \end{aligned} \tag{5}$$

Zero-Shot Auto-ML (DeepMAD): Нужно оптимизировать соотношение глубины и ширины архитектурных блоков CNN, максимизируя *энтропию* (сложность) при ограничении *эффективности* (отношения глубины к средней ширине блока)

Даже с учетом *дополнительных ограничений на число вычислений и параметров* задача решается линейным программированием!

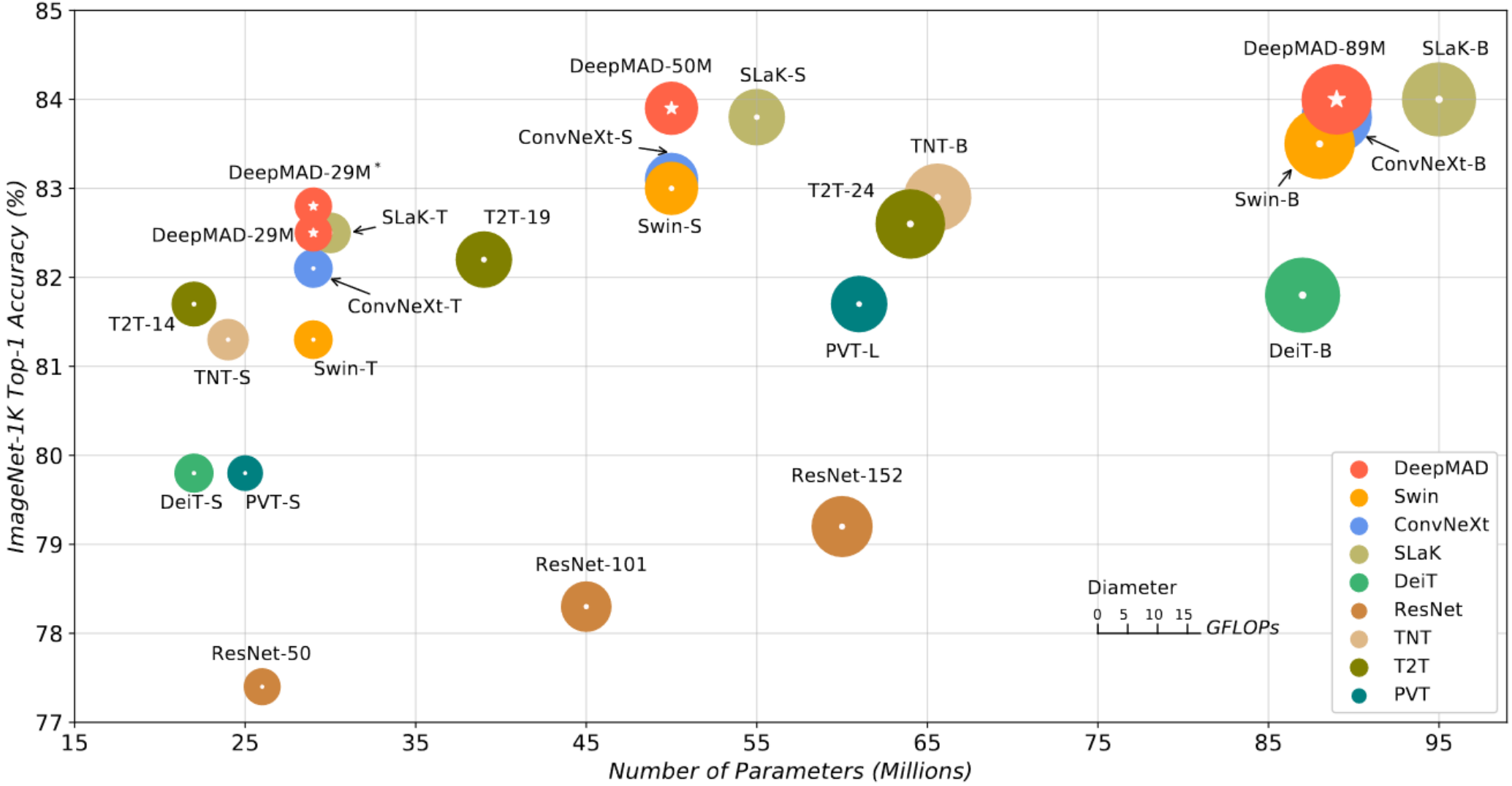


Model	# Param.	FLOPs	ρ	Acc. (%)
ResNet-18 [21]	11.7 M	1.8 G	0.01	70.9
ResNet-18†	11.7 M	1.8 G	0.01	72.2
DeepMAD-R18	11.7 M	1.8 G	0.1	76.9
DeepMAD-R18	11.7 M	1.8 G	**0.3**	**77.7**
DeepMAD-R18	11.7 M	1.8 G	0.5	77.5
DeepMAD-R18	11.7 M	1.8 G	0.7	75.7
ResNet-34 [21]	21.8 M	3.6 G	0.02	74.4
ResNet-34†	21.8 M	3.6 G	0.02	75.6
DeepMAD-R34	21.8 M	3.6 G	0.3	**79.7**
ResNet-50 [21]	25.6 M	4.1 G	0.09	77.4
ResNet-50†	25.6 M	4.1 G	0.09	79.3
DeepMAD-R50	25.6 M	4.1 G	0.3	**80.6**

DeepMAD v.s. ResNet on ImageNet-1K, using ResNet building block. †: model trained by our pipeline. ρ is tuned for DeepMAD-R18. DeepMAD achieves consistent improvements to ResNet18/34/50 with the same Params and FLOPs.

*DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network, Shen et al., 2023

Transformers vs. CNN: ничего еще не ясно! Math CNN Design (DeepMAD)



Auto-ML (DeepMAD) бьет SOTA ViT and CNN (Swin, ConvNext, SLaK)

DeepMAD v.s. SOTA ViT and CNN models on ImageNet-1K. ρ = 0.5 for all DeepMAD models. All DeepMAD models except DeepMAD-29M∗ trained with 224 resolution. x-axis is the Params, the smaller the better. y-axis is the accuracy, the larger the better.

***DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network, Shen et al., 2023**

Тенденции и результаты 2020-2024 в области генерации данных

Diffusion models

Text-2-Image

Neural Radiance Fields (NeRF)

3D Gaussian Splatting (3DGS)

Text-2-Video

Text-2-3D

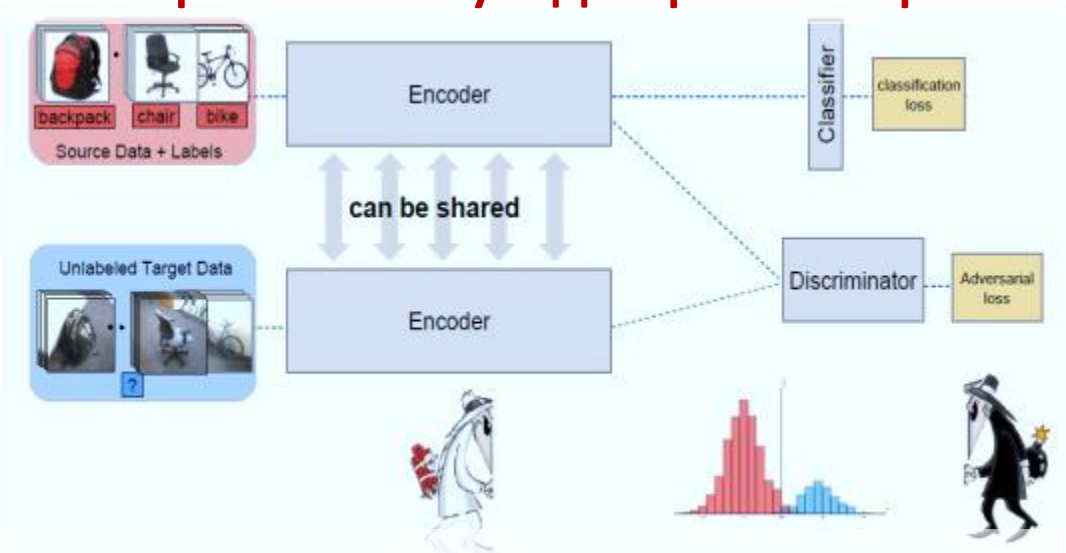
3D Avatars

Генерация данных

GM до 2020

- Generative Adversarial Networks (GAN)
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

Генератор создает визуальные образы, стараясь обмануть Дискриминатор…



….Дискриминатор старается отличить фантазии Генератора от реальности

Dmitry Vetrov, OpenTalks.AI/2022

Adaptive Deep Models across Visual Domains, Kate Saenko, ICCV, 2017

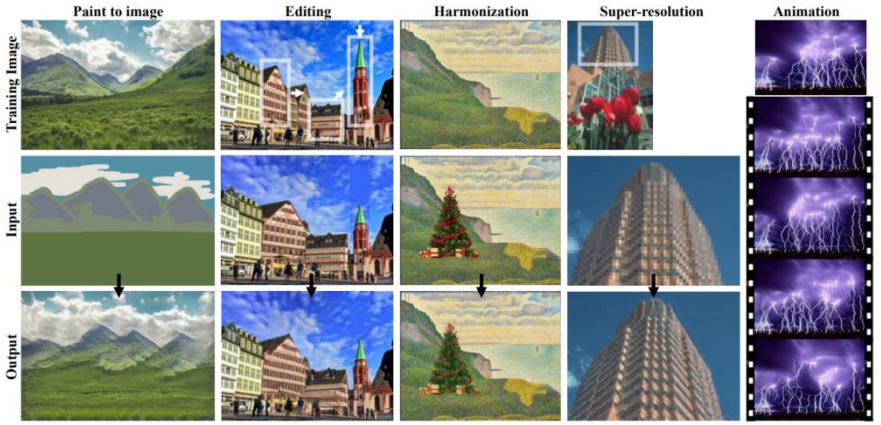
Generative models (GAN)

GM до 2020

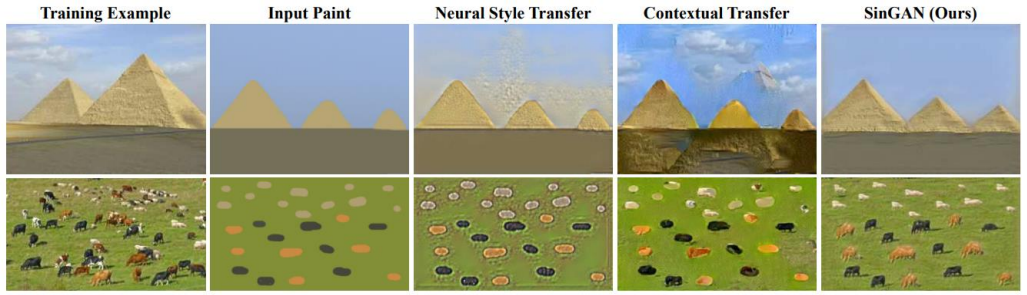
- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

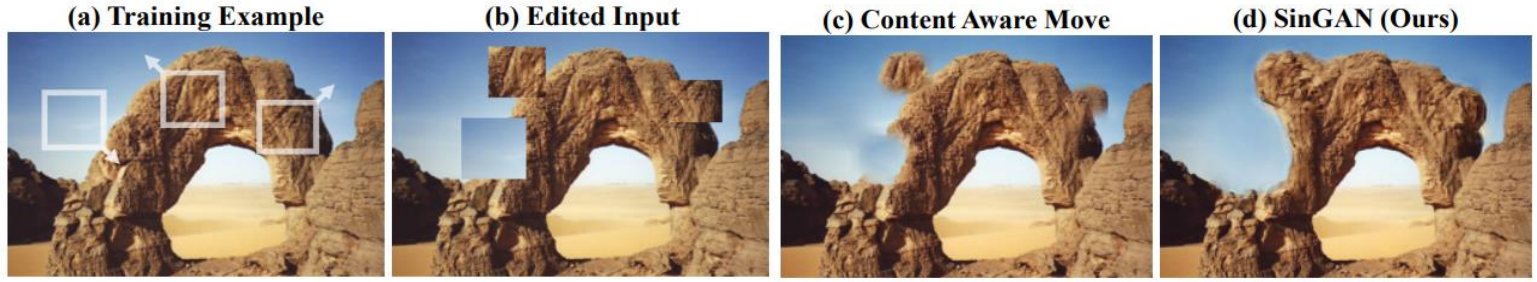
Dmitry Vetrov, OpenTalks.AI/2022

Практически неограниченные возможности реалистичного манипулирования данными (DeepFake)



Неоспоримое лидерство GAN до 2020…





SinGAN: Learning a Generative Model From a Single Natural Image, Shaham et. al, 2019

Проблемы и вызовы в GM (2020)

Слабая генерализация за пределами выборки

Нет генерации образов по словесному описанию

Diffusion models

Новое поколение генеративных моделей

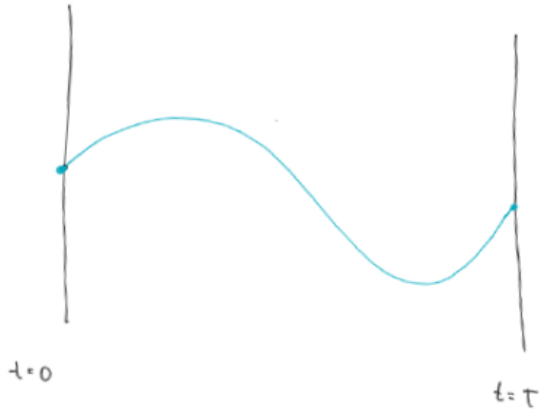
Diffusion models: новый лидер 2020+ (сменил GAN)

References

- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239, June 2020.
- Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

- Ordinary differential equation defines the evolution of a point

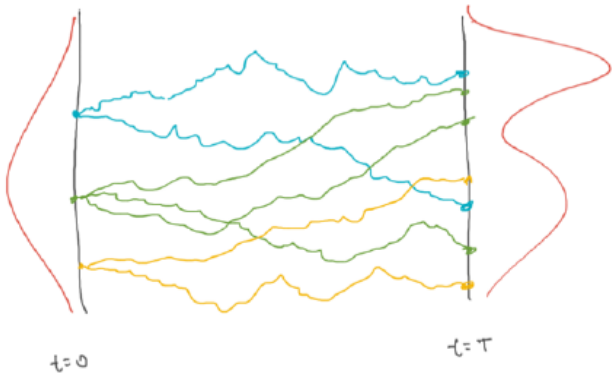
$$dx = f(x,t)dt$$



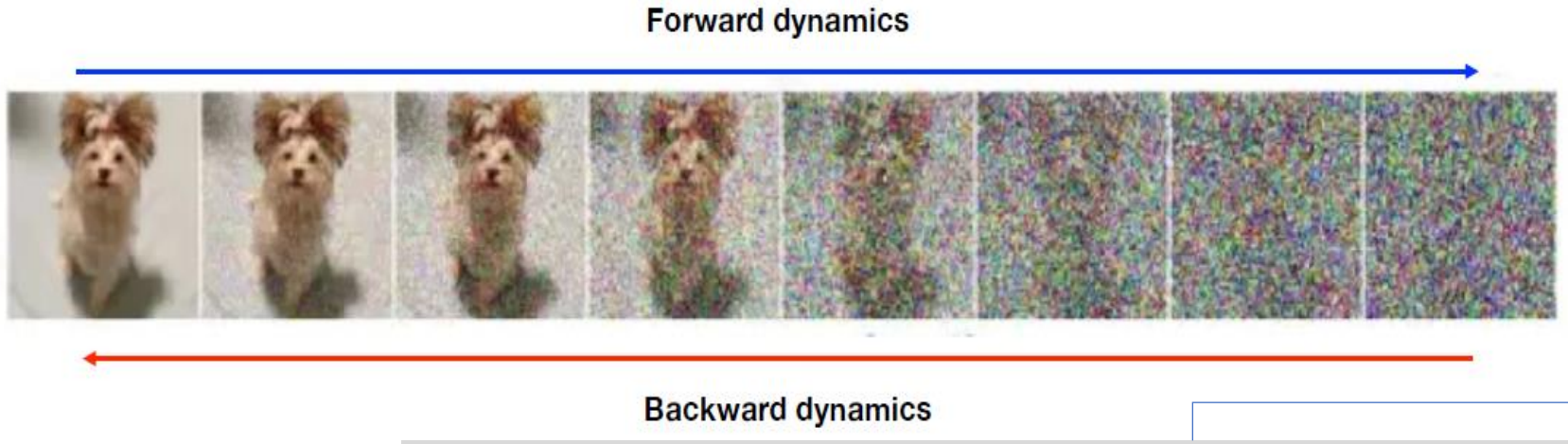
Модель 1: решение обратного стохасти-ческого ДУ

- Stochastic differential equation defines the evolution of a distribution

$$dx = f(x,t)dt + g(x,t)dW$$



Задача: выучить распределение образов для класса на примерах, а потом сэмплировать из этого распределения



Модель 2: поэтапное восстановление зашумленного образа

Кстати, это и есть «усилитель смысла» Эшби, извлекающий смысл из шума!

Denoising diffusion probabilistic models

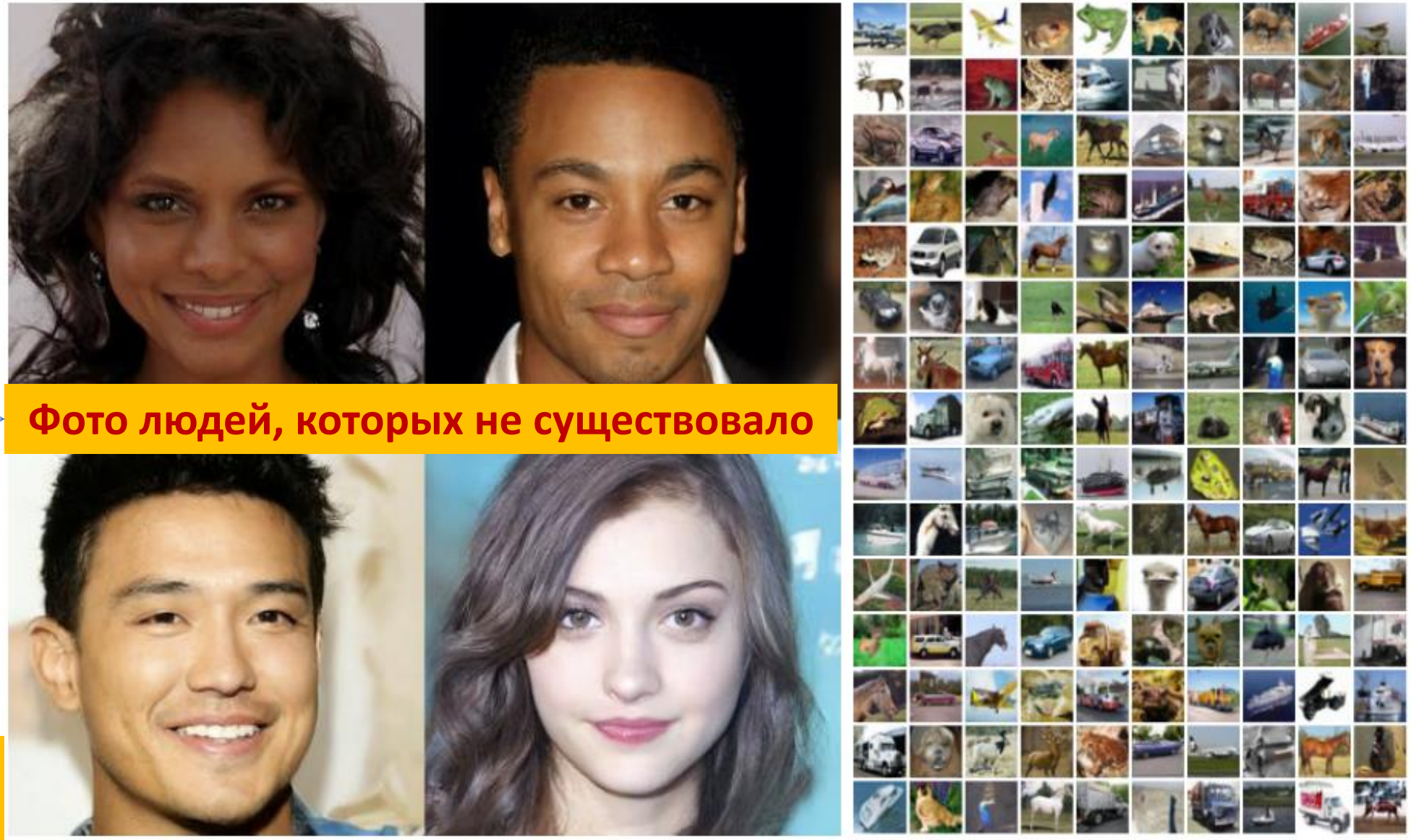


Фото людей, которых не существовало

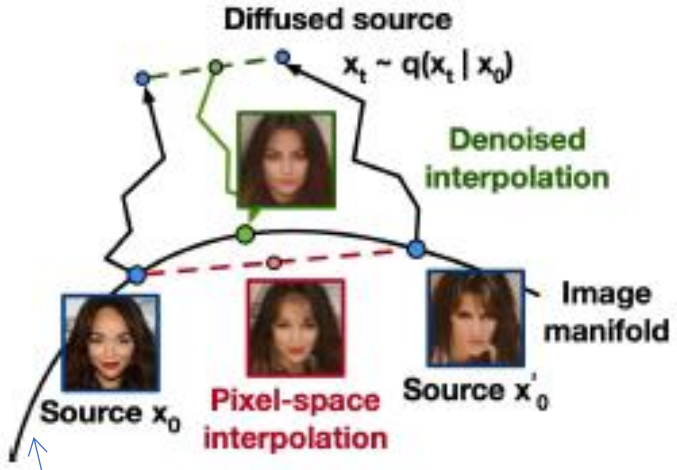
Выучили распределение для лиц

Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Denoising diffusion probabilistic models

Интерполяция между двумя лицами:
в каждой точке только лица!



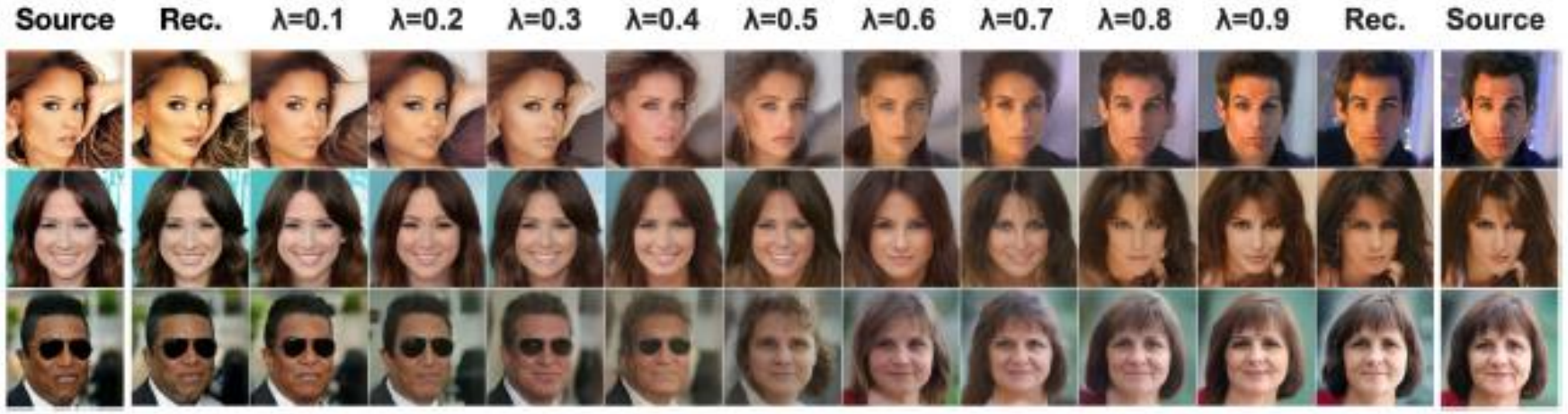


Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

Непрерывное пространство (распределение) изображений

Выучили распределение для лиц

Coarse-to-fine interpolation Figure 9 shows interpolations between a pair of source CelebA 256×256 images as we vary the number of diffusion steps prior to latent space interpolation. Increasing the number of diffusion steps destroys more structure in the source images, which the model completes during the reverse process. This allows us to interpolate at both fine granularities and coarse granularities. In the limiting case of 0 diffusion steps, the interpolation mixes source images in pixel space. On the other hand, after 1000 diffusion steps, source information is lost and interpolations are novel samples.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

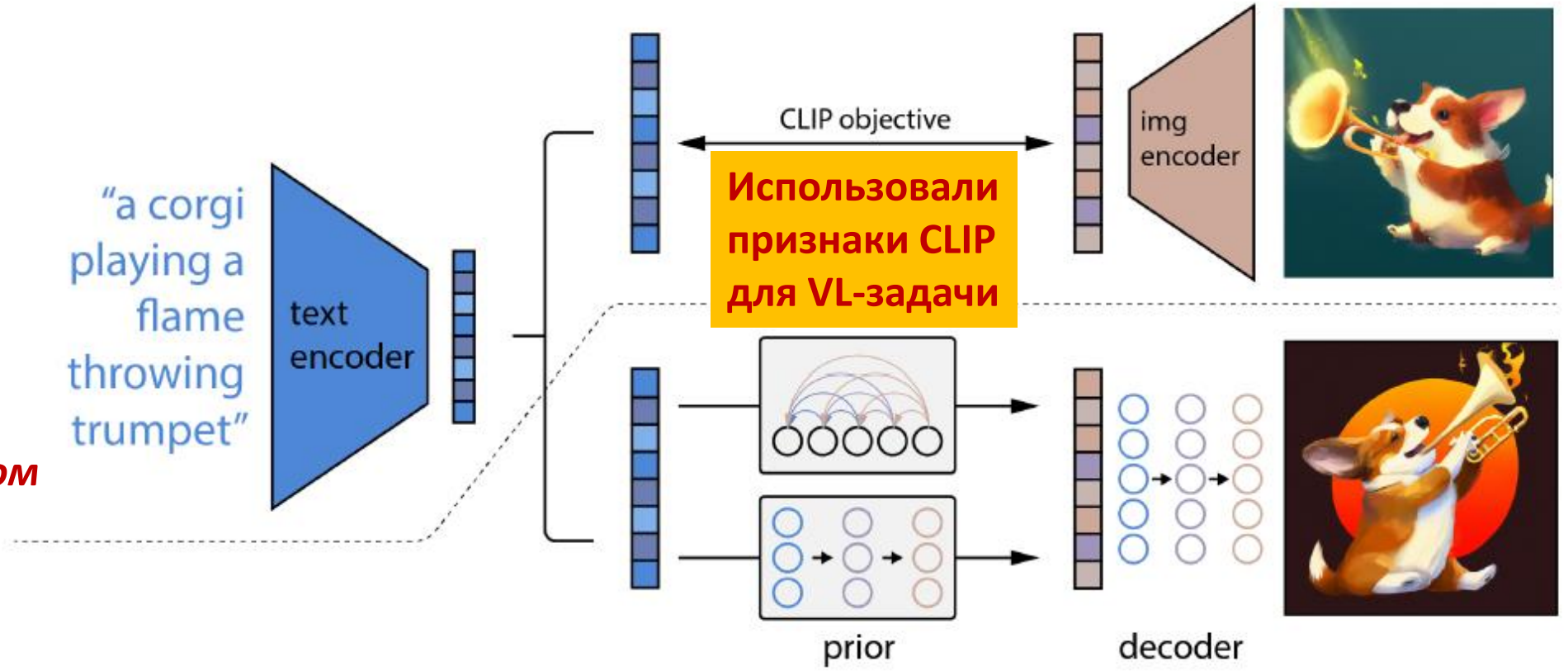
Text-2-Image

(генерация и редактирование изображений по словесному описанию)

DALL·E 2: images and art from natural language description

Задача: text-to-image generation

Генерация изображений и рисунков по текстовому описанию на естественном языке



Использовали признаки CLIP для VL-задачи

A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or **diffusion prior** to produce an image embedding, and then this embedding is used to condition a **diffusion decoder** which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

We develop methods for training **diffusion priors in latent space**, and show that they achieve comparable performance to autoregressive priors, while being more compute-efficient.

DALL·E 2: images and art from natural language description

Непрерывное пространство изображений (распределение) стало семантическим!

a photo of a cat → an anime drawing of a super saiyan cat, artstation

Пространство признаков CLIP

a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

diffusion priors in latent space

DALL·E 2: images and art from natural language description

Главные компоненты образа в CLIP теперь можно визуализировать!

Our decoder model provides a unique opportunity to explore CLIP latent space by allowing us to directly visualize what the CLIP image encoder is seeing. Visualization of *reconstructions of CLIP latents from progressively more PCA dimensions (20, 30, 40, 80, 120, 160, 200, 320 dimensions), with the original source image on the far right.* **The lower dimensions preserve coarse-grained semantic information**, whereas the **higher dimensions encode finer-grained details** of the objects in the scene.

Непрерывное пространство изображений (распределение) стало семантическим

diffusion priors in latent space

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

DALL·E 2: images and art from natural language description

Качество генерации изображений остается высоким даже в нереальных сценах

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

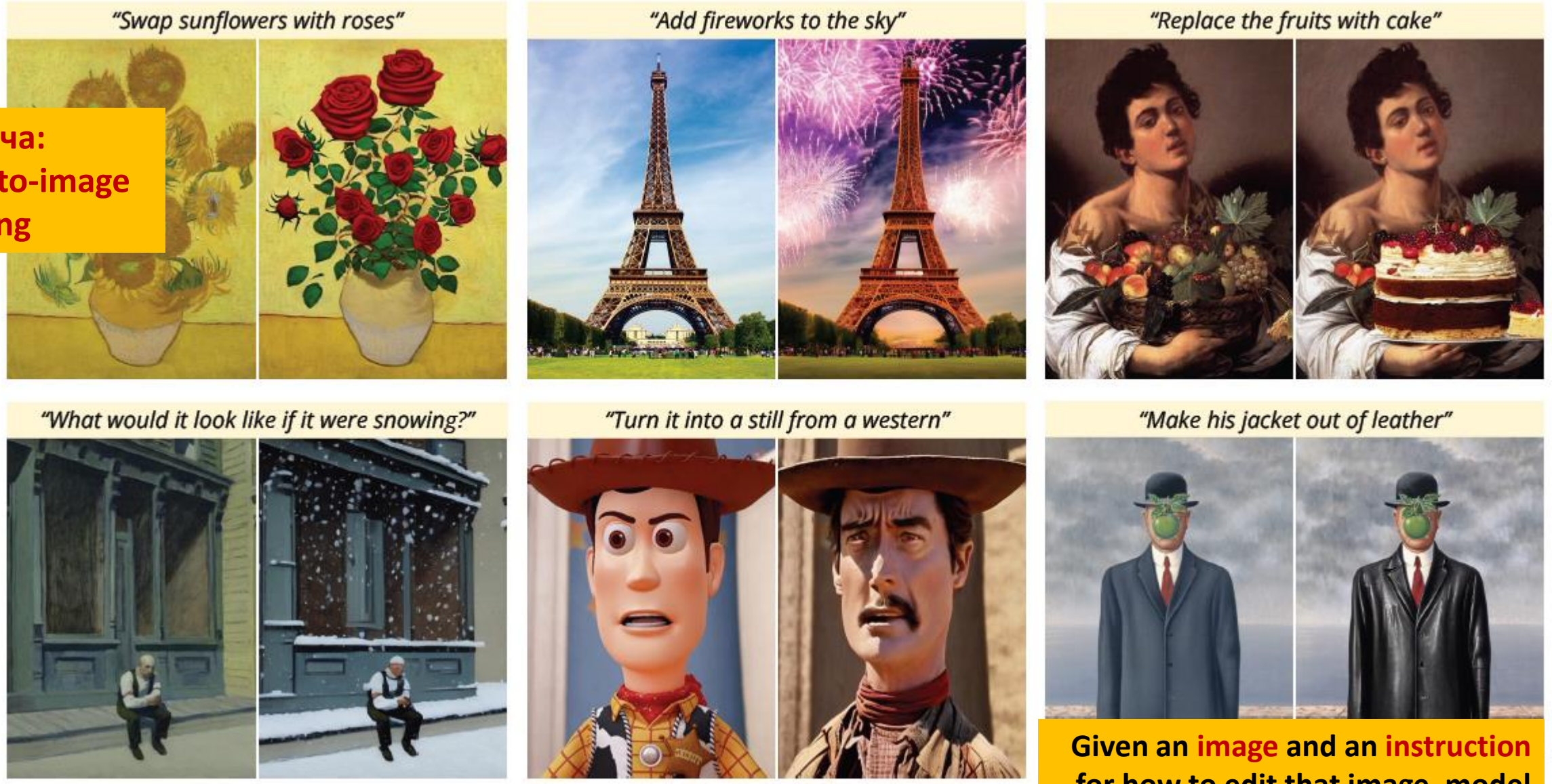
https://openai.com/dall-e-2/#demos

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

diffusion priors in latent space

InstructPix2Pix: как объяснить художнику, чего мы хотим

Задача: text-to-image editing

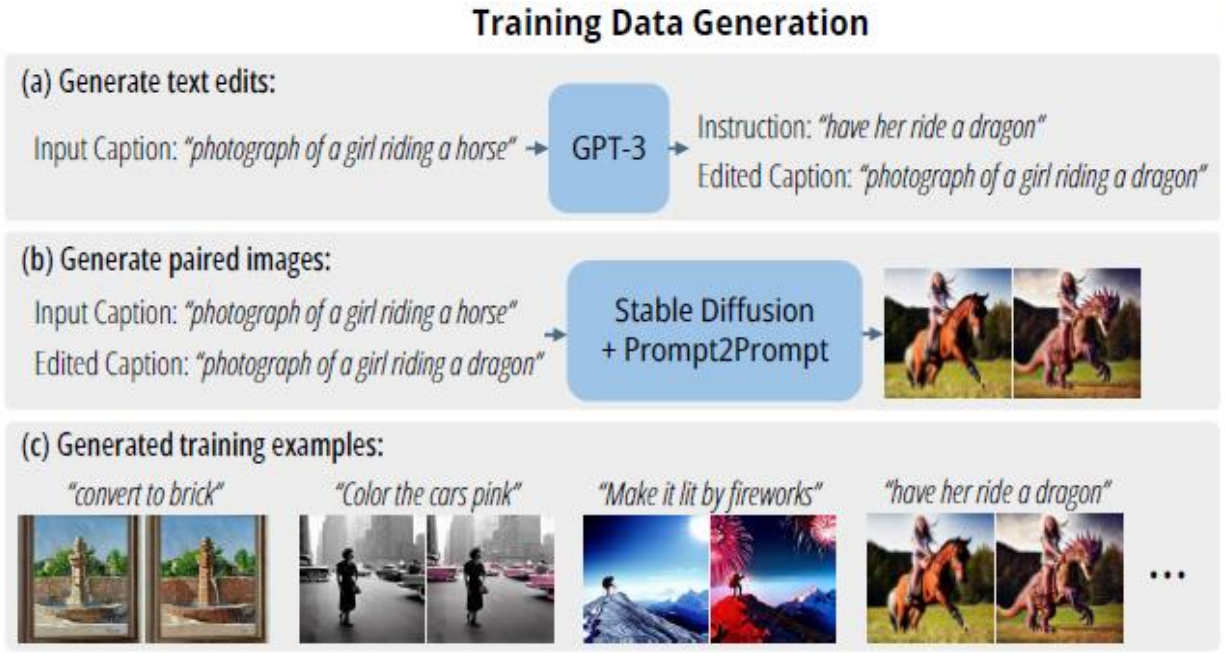


Given an image and an instruction for how to edit that image, model performs the appropriate edit.

InstructPix2Pix: Learning to Follow Image Editing Instructions,
Tim Brooks, Aleksander Holynski, Alexei A. Efros, 2022

InstructPix2Pix: как объяснить художнику, чего мы хотим

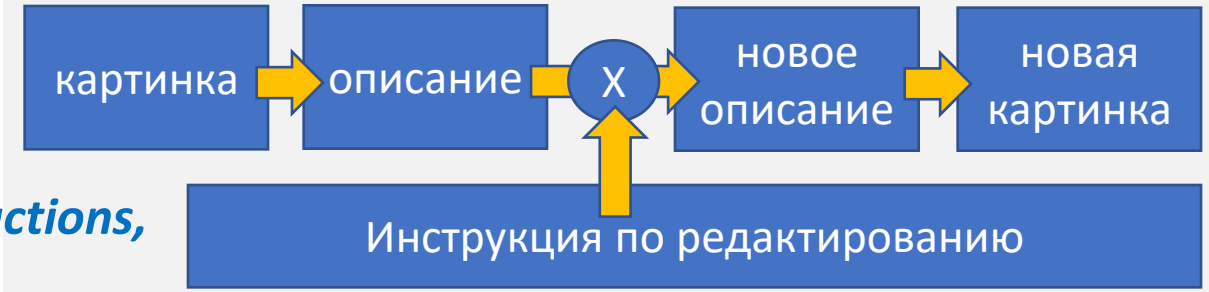
Задача сводится к преды-дущей!



Описываем картину-редактируем описание-генерируем новую картину

Learning consists of two parts: **generating an image editing dataset**, and **training a diffusion model on that dataset**. (a) We first use a finetuned GPT-3 to generate instructions and edited captions. (b) We then use **StableDiffusion** in combination with **Prompt-to-Prompt** to generate pairs of images from pairs of captions. We use this procedure to create a dataset (c) of over 450,000 training examples. (d) Finally, our **InstructPix2Pix diffusion model** is trained on our generated data to edit images from instructions.
At inference time, our model generalizes to edit real images from human-written instructions.

Генерация датасета для обучения

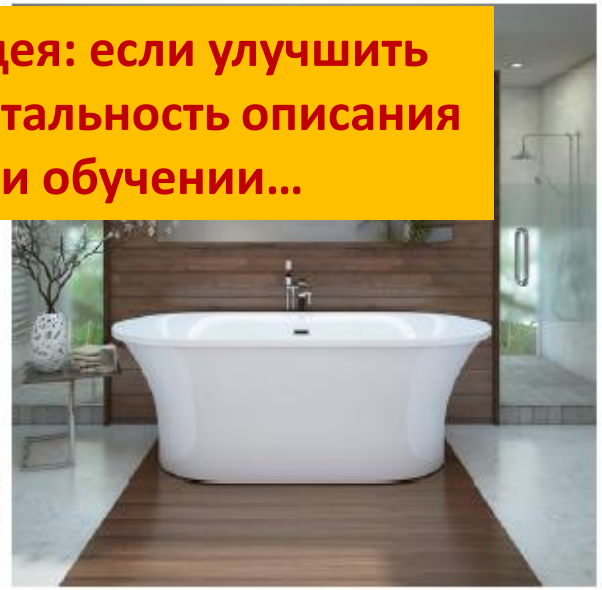


InstructPix2Pix: Learning to Follow Image Editing Instructions, Tim Brooks, Aleksander Holynski, Alexei A. Efros, 2022

DALL-E 3: Improving Image Generation with Better Captions

2023

Идея: если улучшить детальность описания при обучении…

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Important details that are commonly omitted in all-text:

1. The **presence of objects** like stop signs along a sidewalk and **descriptions of them**.
2. The **position of objects** in a scene and the **number of those objects**.
3. **Common sense details like the colors and sizes of objects** in a scene.
4. **The text that is displayed in an image**.

To improve the captions in our image generation dataset, we want our captioner to produce image descriptions useful for learning a text-to-image model.

Examples of **alt-text** accompanying selected images scraped from the internet, **short synthetic captions (SSC)**, **and descriptive synthetic captions (DSC)**

Improving Image Generation with Better Captions. 2023.

DALL-E 3: Improving Image Generation with Better Captions

Effect of using "upsampled" drawbench captions to create samples with DALL-E 3. Original drawbench captions on top, upsampled captions on bottom. Images are best of 4 for each caption.

A bird scaring a scarecrow.

Paying for a quarter-sized pizza with a pizza-sized quarter.

A smafml vessef epropoeilled on watvewr by ors, sauls, or han engie.

A large, vibrant bird with an impressive wingspan swoops down from the sky, letting out a piercing call as it approaches a weathered scarecrow in a sunlit field. The scarecrow, dressed in tattered clothing and a straw hat, appears to tremble, almost as if it's coming to life in fear of the approaching bird.

A person is standing at a pizza counter, holding a gigantic quarter the size of a pizza. The cashier, wide-eyed with astonishment, hands over a tiny, quarter-sized pizza in return. The background features various pizza toppings and other customers, all of them equally amazed by the unusual transaction.

A small vessel, propelled on water by oars, sails, or an engine, floats gracefully on a serene lake. the sun casts a warm glow on the water, reflecting the vibrant colors of the sky as birds fly overhead.

…то качество генерации изображений также резко улучшится!

Important details that are commonly omitted in all-text:

1. The **presence of objects** like stop signs along a sidewalk and **descriptions of them**.
2. The **position of objects** in a scene and the **number of those objects**.
3. **Common sense details like the colors and sizes of objects** in a scene.
4. **The text that is displayed in an image**.

To improve the captions in our image generation dataset, we want our captioner to produce image descriptions useful for learning a text-to-image model.

Improving Image Generation with Better Captions. 2023.

ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models

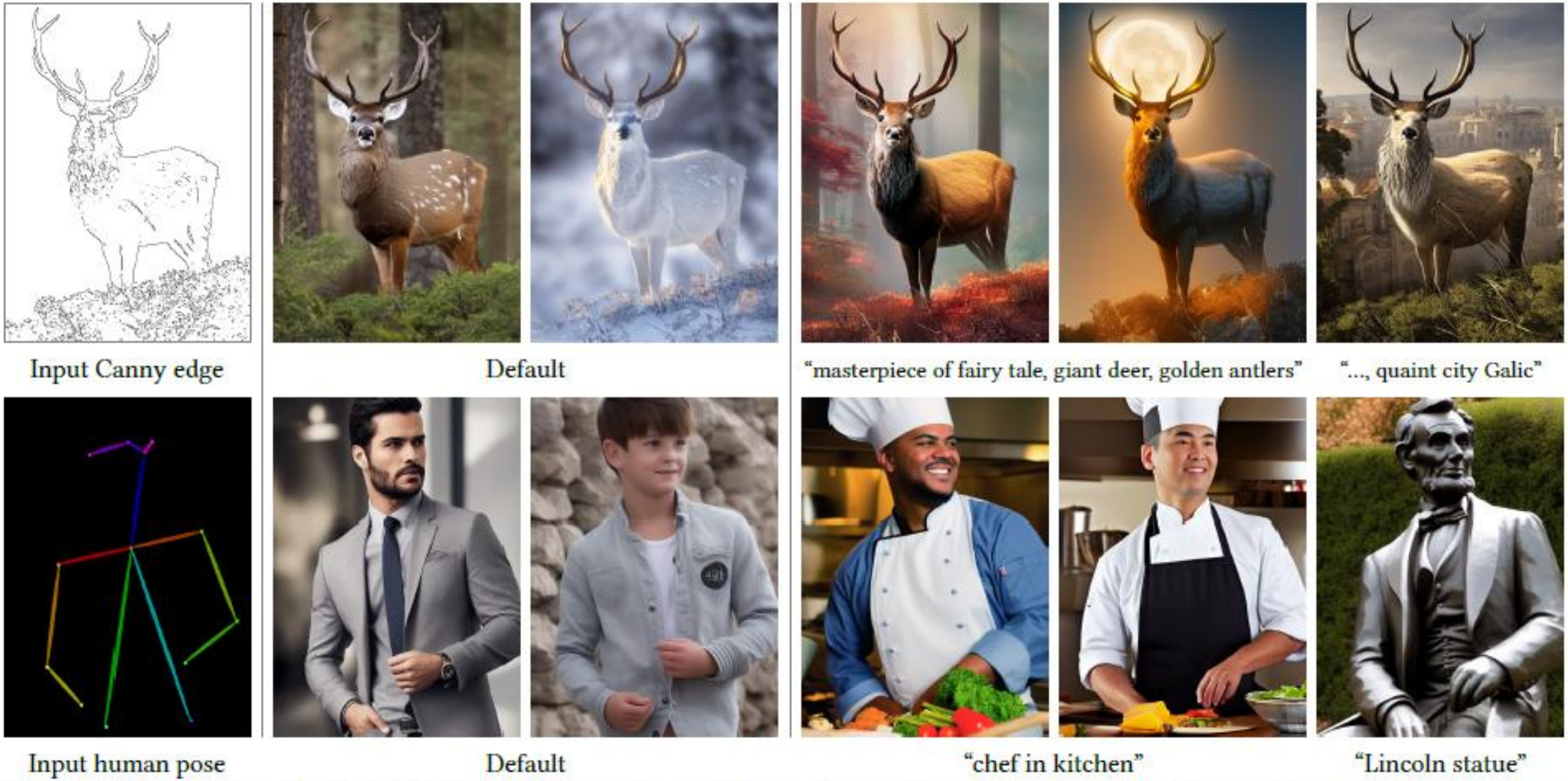


Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

Text-2-Image: ***Добавим к текстовому запросу визуальный скетч!***

Задача: управлять генерацией изображения не только словесно

Adding Conditional Control to Text-to-Image Diffusion Models. 2023.

Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models

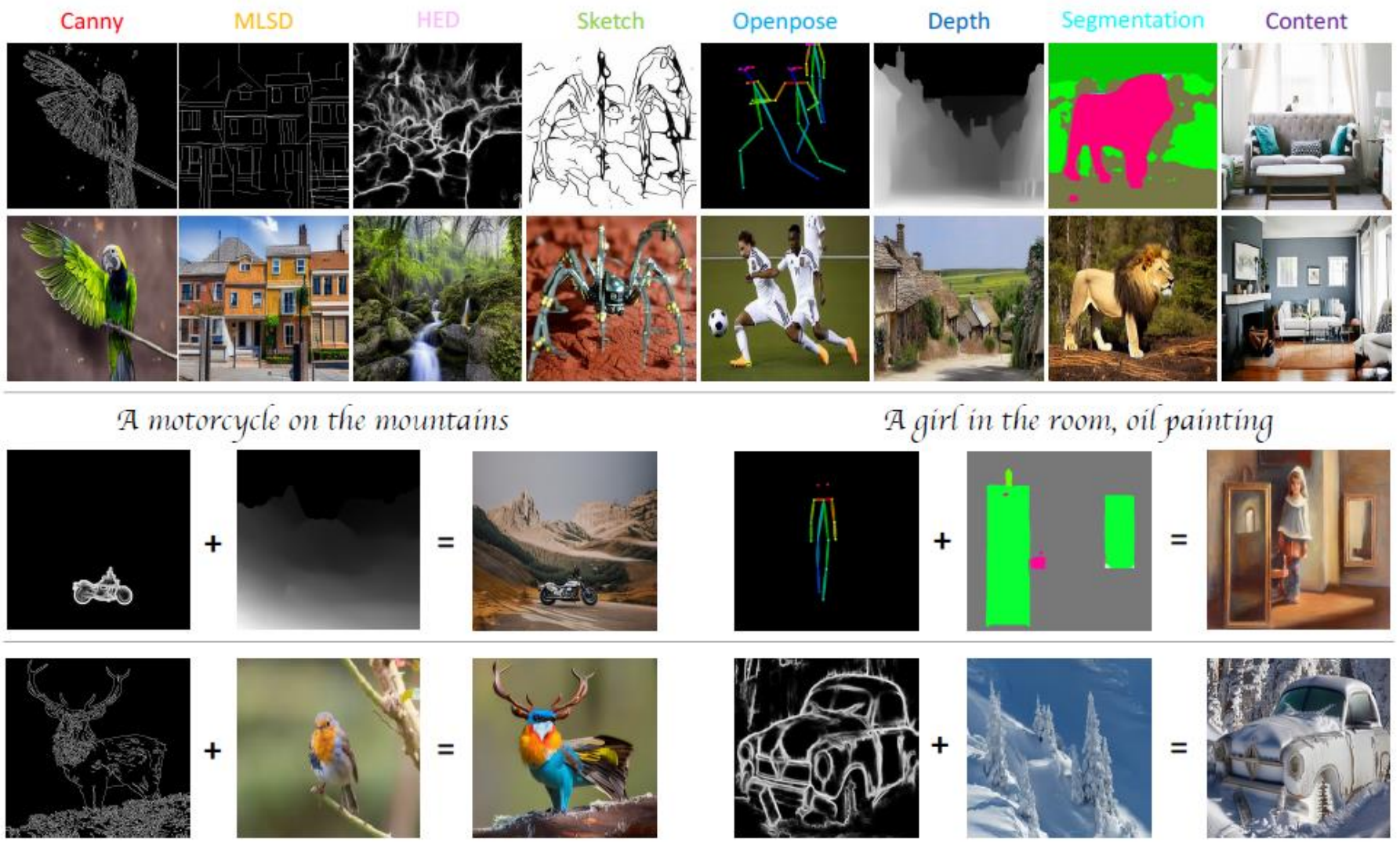


Figure 1: Visual results of our proposed Uni-ControlNet. The top and bottom two rows are results for single condition and multi-conditions respectively.

Text-2-Image: ***Добавим к локальным визуальным свойствам глобальные***

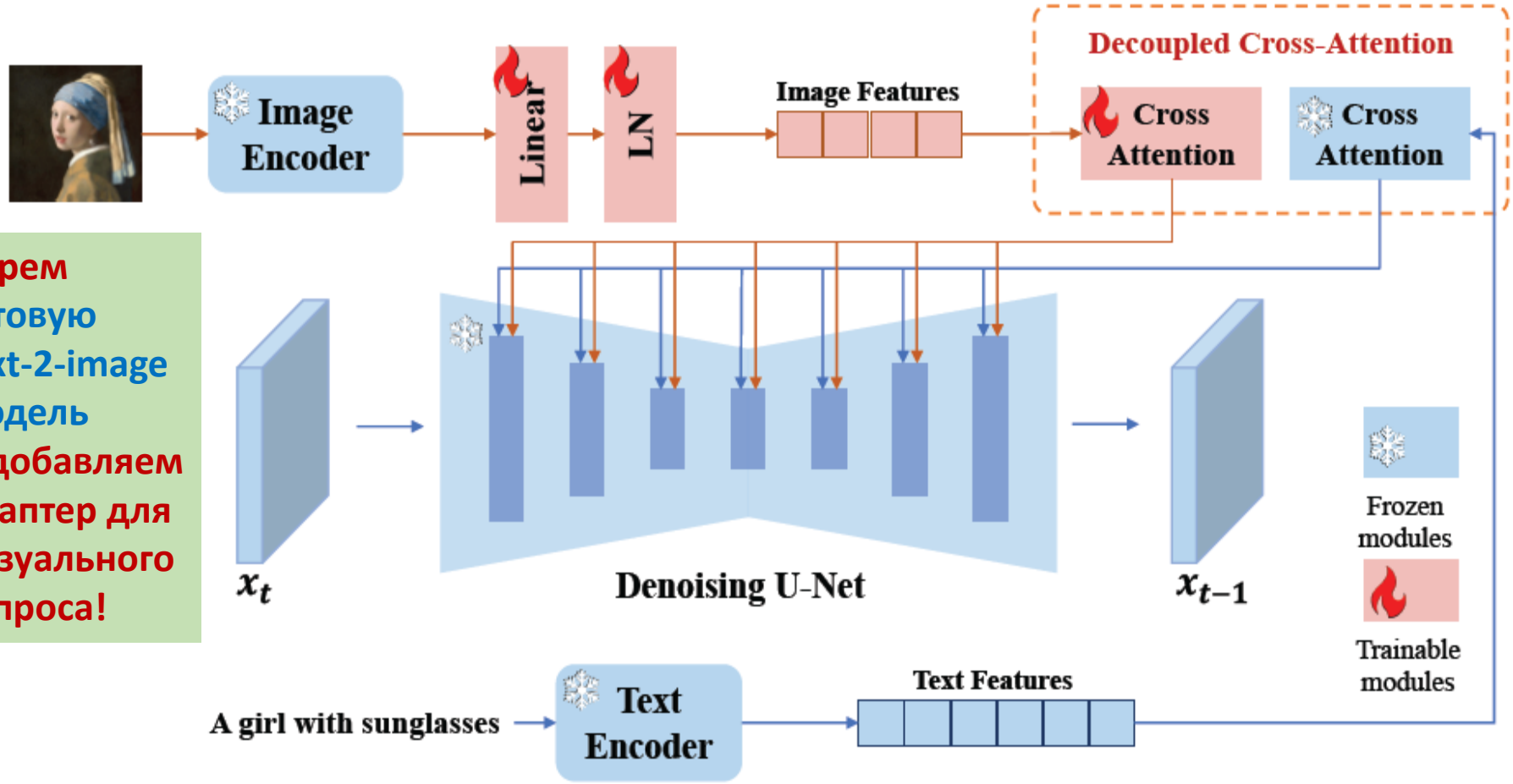
Uni-ControlNet, a unified framework that allows for the **simultaneous utilization of different local controls** (e.g., edge maps, depth map, segmentation masks) **and global controls** (e.g., CLIP image embeddings) in a flexible and composable manner within one single model

Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. 2023.

Text-2-Image:
Не добавим что-то к тексту, а подадим полноценный визуальный запрос!

IP-Adapter: Text Compatible Image Prompt Adapter

The overall architecture of our proposed IP-Adapter with decoupled cross-attention strategy. Only the newly added modules (in red color) are trained while the pretrained text-to-image model is frozen.



Берем готовую text-2-image модель и добавляем адаптер для визуального запроса!

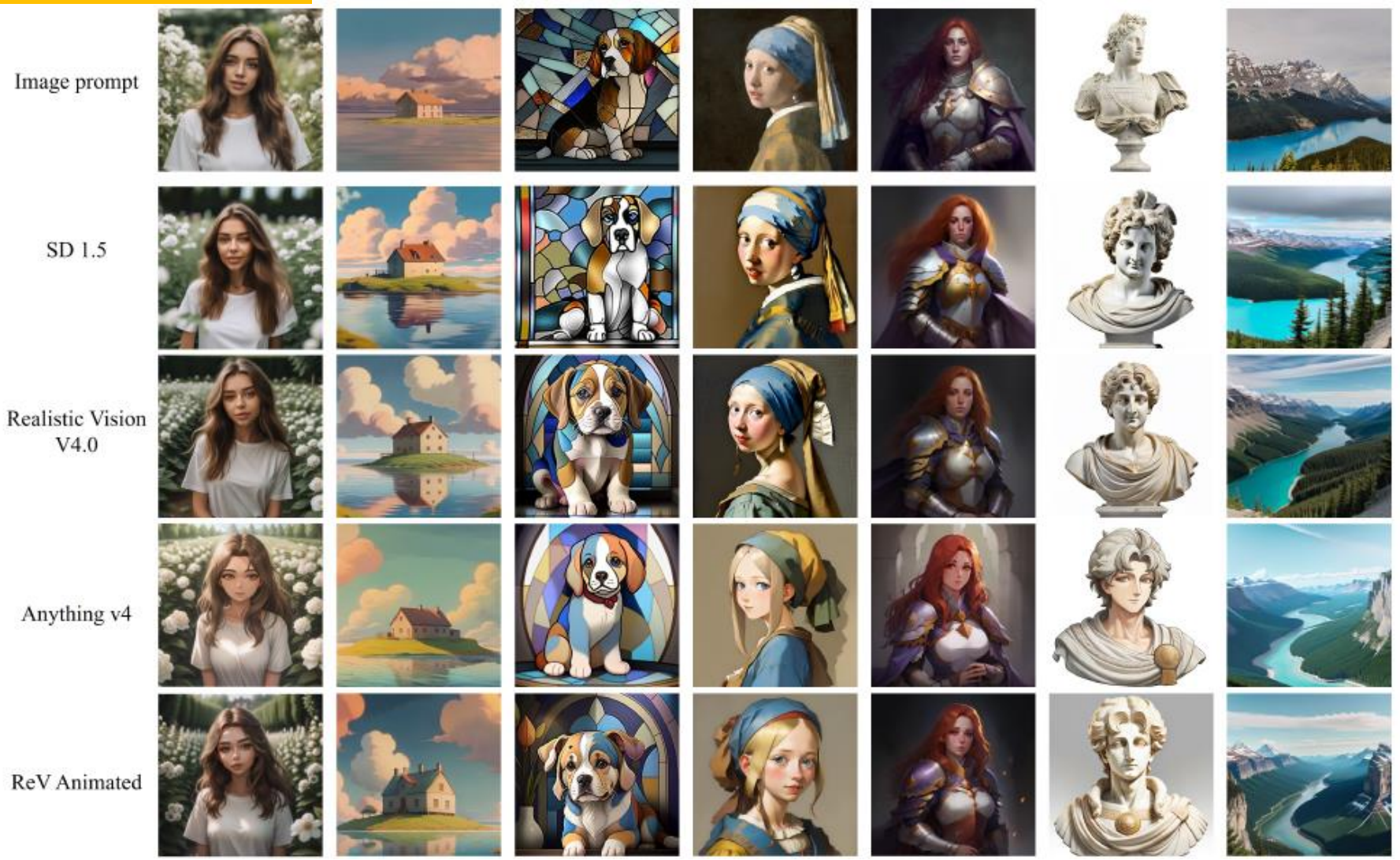
The main problem: cross-attention layers in the pretrained diffusion model are trained to adapt the text features. Consequently, merging image features and text features only accomplishes the alignment of image features to text features, but **potentially misses some image-specific information from the reference image.** Our IP-Adapter adopts a decoupled cross-attention mechanism**. For every cross-attention layer in the UNet of diffusion model, we add an additional cross-attention layer only for image features.** In the training stage, **parameters of the new cross-attention layers are trained**, while the **original UNet model remains frozen**.

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. 2023.

Text-2-Image: ***Текстовый и визуальный запрос!***

IP-Adapter: Text Compatible Image Prompt Adapter

Images of different diffusion models with our proposed IP-Adapter. The IP-Adapter is only trained once.



We present **IP-Adapter**, an effective and lightweight adapter to achieve image prompt capability for the pretrained text-to-image diffusion models. **The key design of our IP-Adapter is decoupled cross-attention mechanism that separates cross-attention layers for text features and image features**. IP-Adapter with only **22M parameters can achieve comparable or even better performance to a fully fine-tuned image prompt model**. As we freeze the pretrained diffusion model, the proposed IP-Adapter can be generalized not only to other custom models fine-tuned from the same base model, but also to controllable generation using existing controllable tools

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. 2023.

Stable Diffusion 3: текущий лидер (март 2024)

a space elevator, cinematic scifi art

A cheeseburger with juicy beef patties and melted cheese sits on top of a toilet that looks like a throne and stands in the middle of the royal chamber.

a hole in the floor of my bathroom with small gremlins living in it

a small office made out of car parts

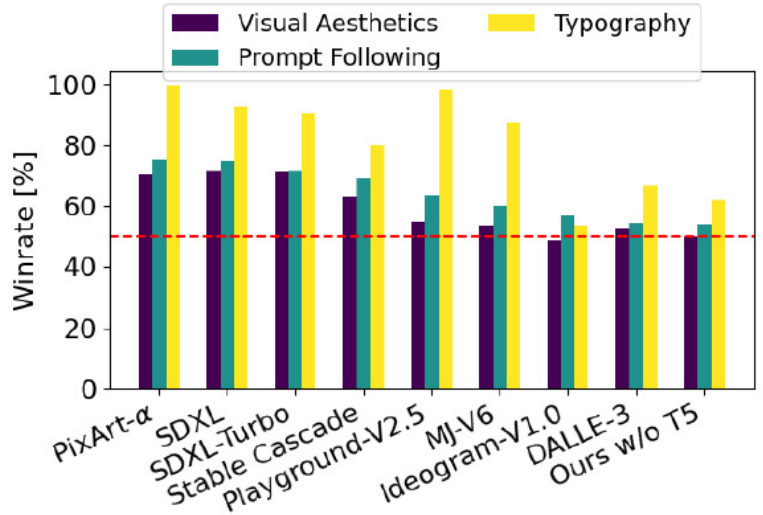
This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest.

human life depicted entirely out of fractals

an origami pig on fire in the middle of a dark room with a pentagram on the floor

an old rusted robot wearing pants and a jacket riding skis in a supermarket.

smiling cartoon dog sits at a table, coffee mug on hand, as a room goes up in flames. "This is fine," the dog assures himself.



Human Preference Evaluation against current closed and open SOTA generative image models. Our 8B model compares favorable against current state-of-the-art text-to-image models when evaluated on the parti-prompts (Yu et al., 2022) across the categories visual quality, prompt following and typography generation.

Outperforms DALL·E 3, Midjourney v6, and Ideogram v1

Our new **Multimodal Diffusion Transformer (MMDiT)** architecture **uses separate sets of weights for image and language representations**, which improves text understanding and spelling capabilities compared to previous Stable Diffusion.

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. 2024.

One-step Diffusion (2023): дистиллируем DM в один этап!

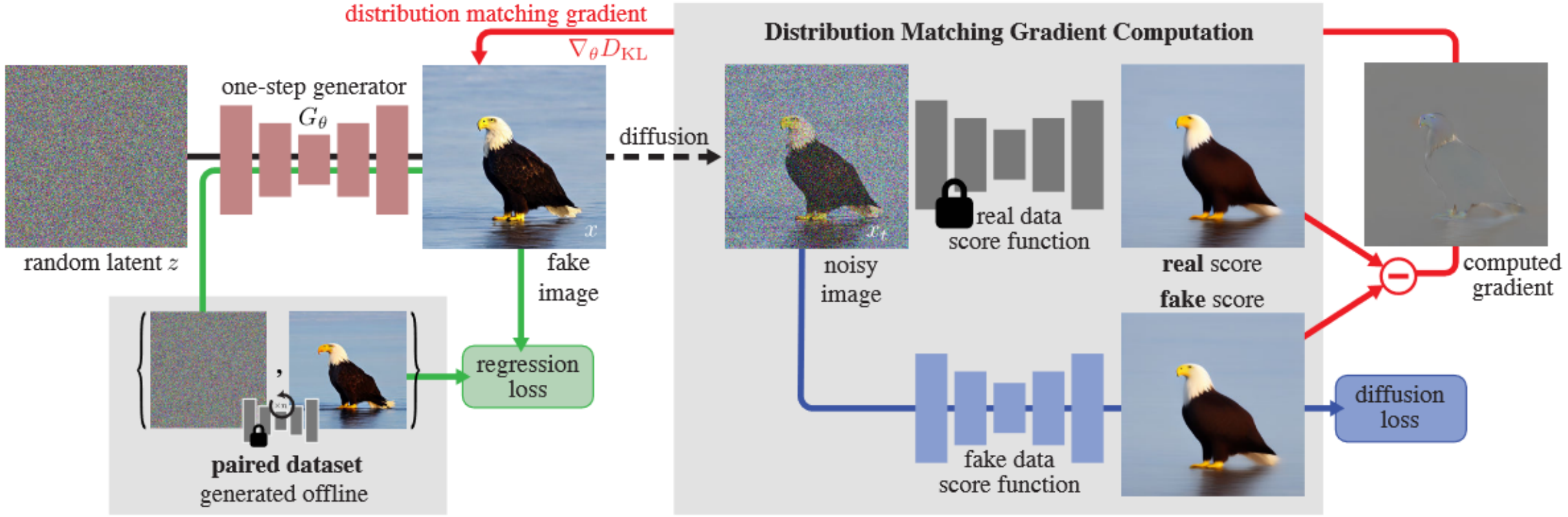
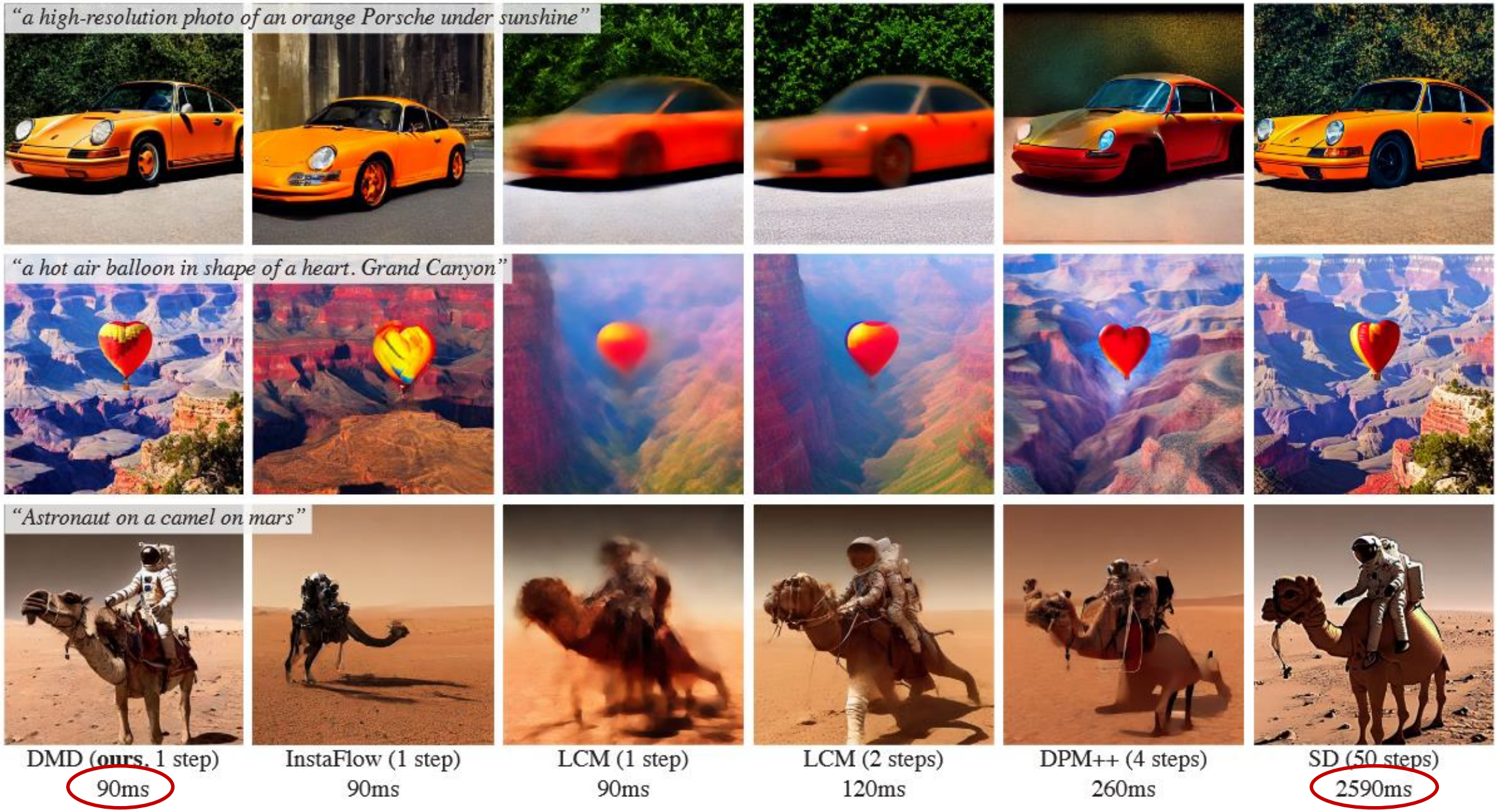


Figure 2. **Method overview.** We train one-step generator G_θ to map random noise z into a realistic image. To match the multi-step sampling outputs of the diffusion model, we pre-compute a collection of noise–image pairs, and occasionally load the noise from the collection and enforce LPIPS [85] regression loss between our one-step generator and the diffusion output. Furthermore, we provide **distribution matching gradient** $\nabla_\theta D_{KL}$ to the fake image to enhance realism. We inject a random amount of noise to the fake image and pass it to two diffusion models, one pretrained on the real data and the other continually trained on the fake images with a diffusion loss, to obtain its denoised versions. The denoising scores (visualized as mean prediction in the plot) indicate directions to make the images more realistic or fake. The difference between the two represents the direction toward more realism and less fakeness and is backpropagated to the one-step generator.

One-step Diffusion with Distribution Matching Distillation, 2023

Задача: ускорить генерацию изображений

One-step Diffusion (2023): генерация образов в 30 раз быстрее!



One-step Diffusion with Distribution Matching Distillation, 2023

Starting from Stable Diffusion (right), our distribution matching distillation algorithm yields a model that can generate images with the close quality (left), but faster!

Text-2-Video

IMAGEN VIDEO: video generation with diffusion models

Melting pistachio ice cream dripping down the cone.

Ролики короткие

A british shorthair jumping over a couch.

Проблема с лишними конечностями

Coffee pouring into a cup.

Камера на движется, 3D образ нарушается

IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS, Google Research, Brain Team, 2022

Генерация видео по текстовому описанию

SORA: Video generation models as world simulators (2024)

Ролики длиннее

https://cdn.openai.com/tmp/s/bike_1.mp4

3D consistency. SORA может создавать видео с динамическим движением камеры. Когда камера сдвигается и вращается, люди и элементы сцены последовательно перемещаются в трехмерном пространстве.

https://openai.com/sora#research

Video generation models as world simulators, OpenAI, 2024

Генерация видео по текстовому описанию

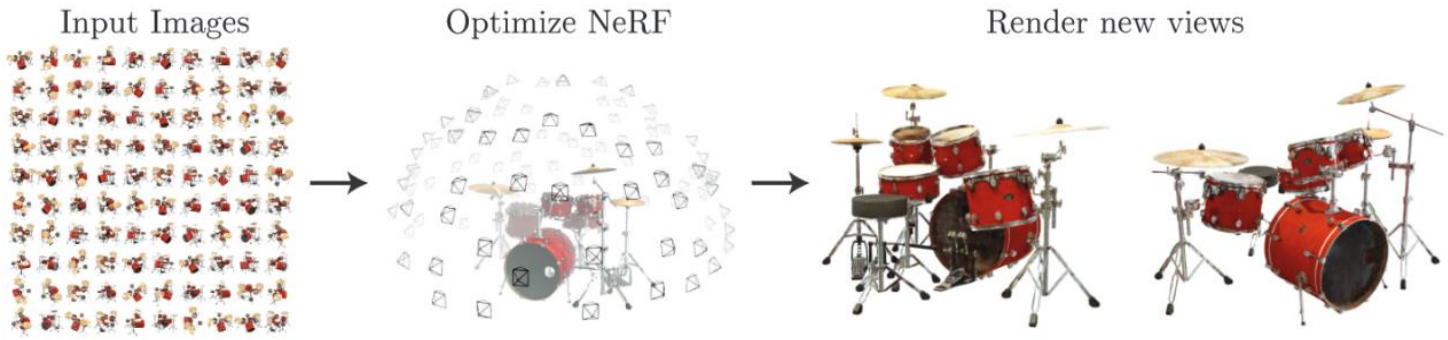
3D Scene Generation & Analysis

Предобученные по ракурсным изображениям описания 3D сцен

Neural Radiance Fields (NeRF)

Предобученные по ракурсным изображениям описания 3D сцен

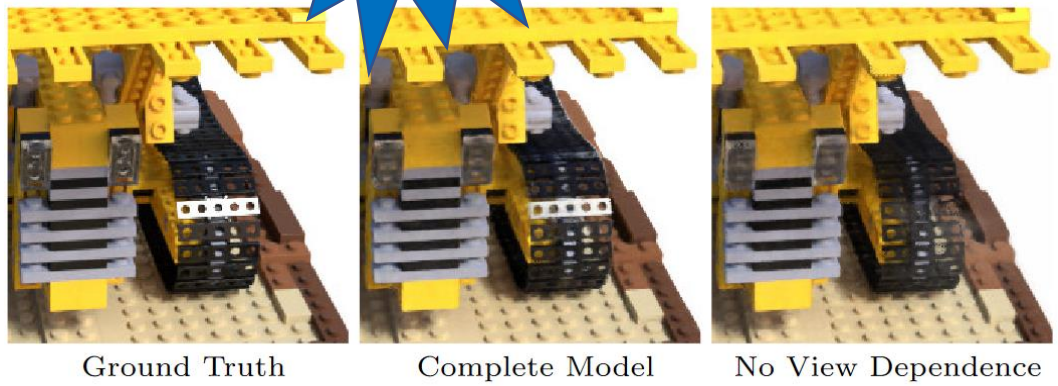
NeRF: Representing Scenes as Neural Radiance Fields

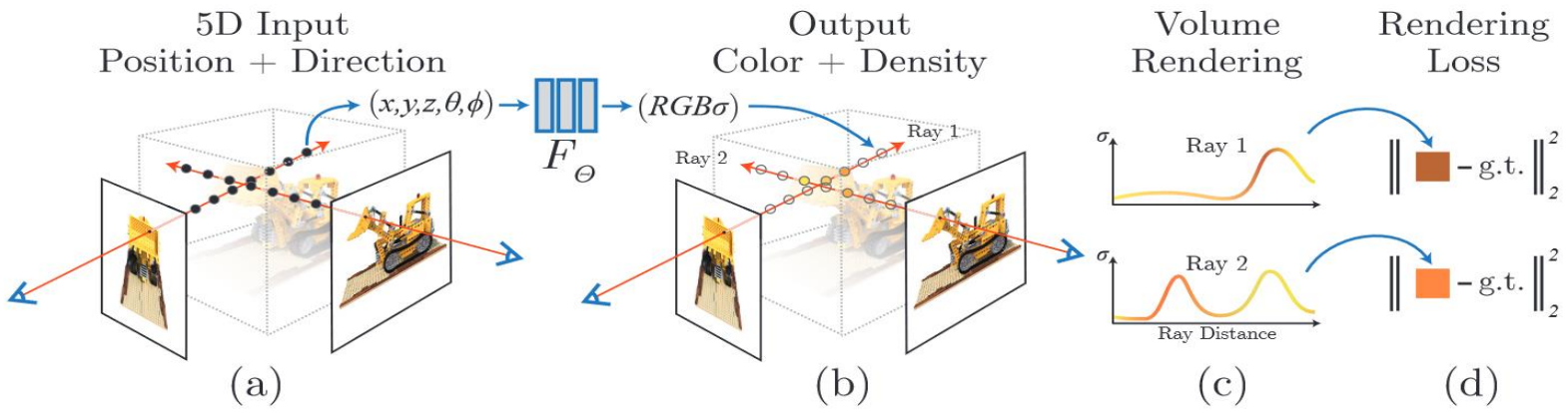


Как соединить традиционные технологии 3D рендеринга с нейронными сетями и машинным обучением

TOP AI 2021!

Continuous 5D neural radiance field representation (volume density and view-dependent color at any continuous location) of a scene from a set of input images. We use techniques from volume rendering to accumulate samples of this scene representation along rays to render the scene from any viewpoint.





- Пропускаем лучи через сцену для семплирования на них множества точек
- Для каждой точки из множества получаем плотность (не зависит от направления взгляда) и цвет (зависит от направления взгляда)
- Применяем рендеринг для создания изображений

Mildenhall et. al, Nerf: Representing scenes as neural radiance fields for view synthesis. UC Berkeley, Google Research, 2021

Развитие и приложения NeRF

Overview

Не только компьютерная графика, но и компьютерное зрение

В исходной задаче генерации изображений 3D сцен и объектов

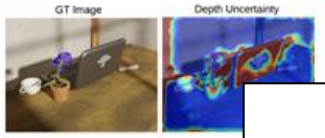
NeRF-RPN: object detection in NeRFs

Object Detection

LERF: Language Embedded Radiance Fields

Multi-modal Visual-Language Tasks

NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields



SLAM: Ориентация роботов и БЛА в пространстве

Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation

Neural Fields for Robotic Object Manipulation from a Single Image

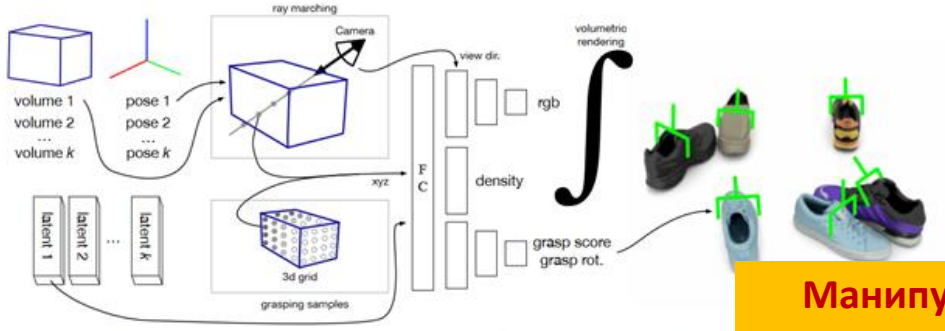
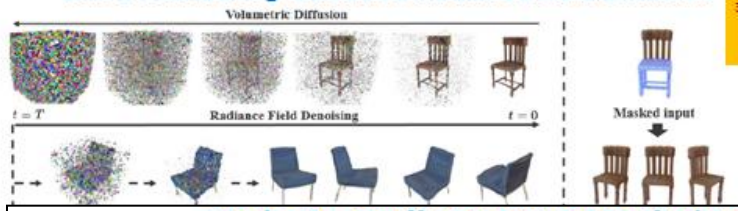


Fig. 1. We present a unified representation that can be used to re-render scenes and generate grasping poses. Given obj... left), and learned latent codes (bottom left), the system renders the scene in novel configurations and annotates each ob... image). The rendering ray-marches through a set of volumes (cuboids in which the objects reside). In order to generate grasps (shown in green on the right), we use a grid sampling approach that covers each object volume. Grasp orientation and score are predicted using a decoder which is partially shared with the volumetric rendering process. We select the grasp point that yields the highest confidence score. FC refers to a fully connected layer.

Манипулирование предметами

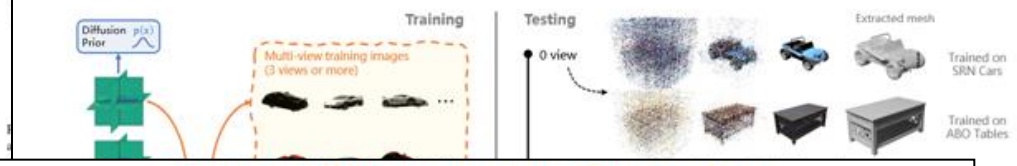
Манипулирование предметами

DiffRF: Rendering-Guided 3D Radiance Field Diffusion



Вернемся к задаче генерации изображений 3D сцен и объектов

Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction



DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models

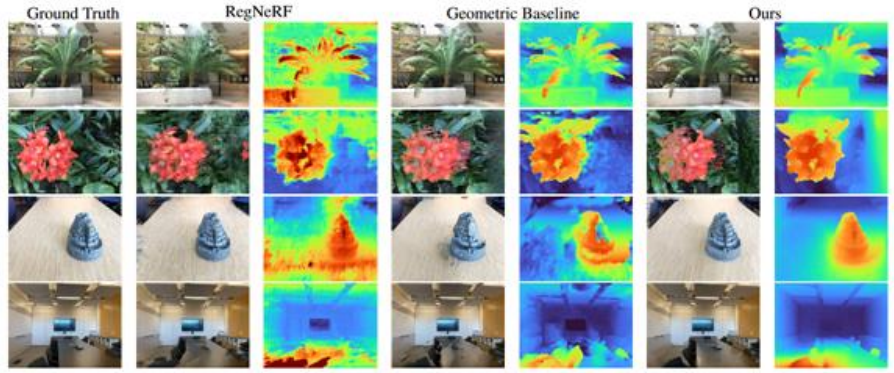
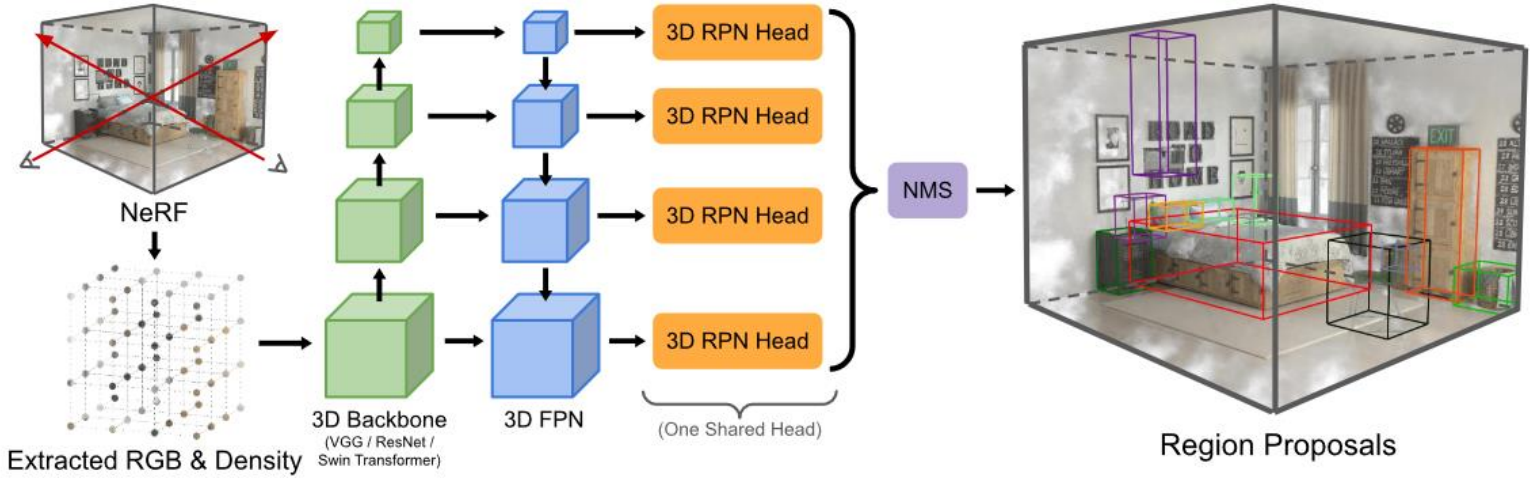


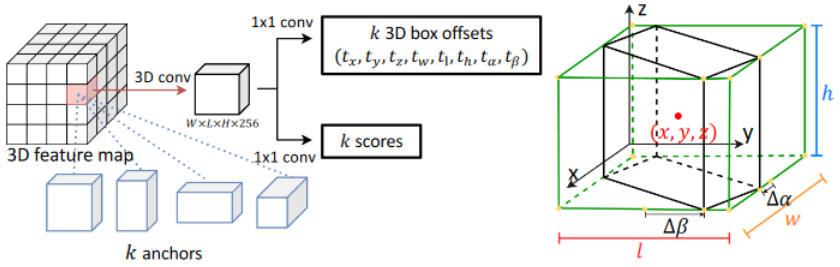
Figure 4. Qualitative results for the task of novel view synthesis on LLFF dataset. NeRF models are trained with 3 views and rendered from one of test views. Our DDM model encourages more realistic geometry as seen in the depth maps.

Соединяем NeRF с диффузными моделями

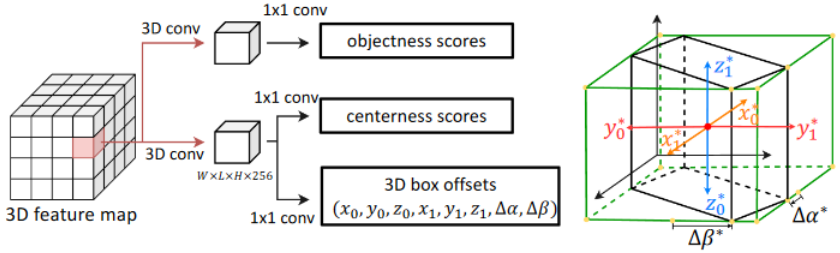
Соединяем NeRF с диффузными моделями

NeRF-RPN: object detection in NeRFs





(a) Anchor-based 3D RPN head and bounding box representation



(b) Anchor-free 3D RPN head and bounding box representation

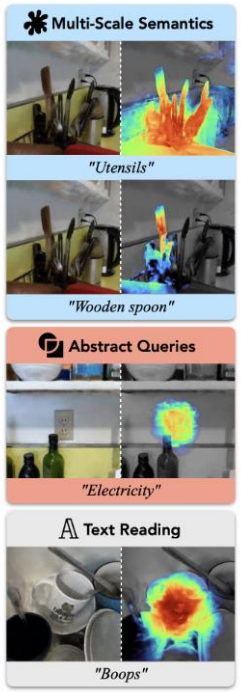
NeRF-RPNdirectly operates on NeRF. Given a pre-trained NeRF model, NeRF-RPN aims to detect all bounding boxes of objects in a scene. By exploiting a novel voxel representation that incorporates multi-scale 3D neural volumetric features, we demonstrate it is possible to regress the 3D bounding boxes of objects in NeRF directly without rendering the NeRF at any viewpoint.

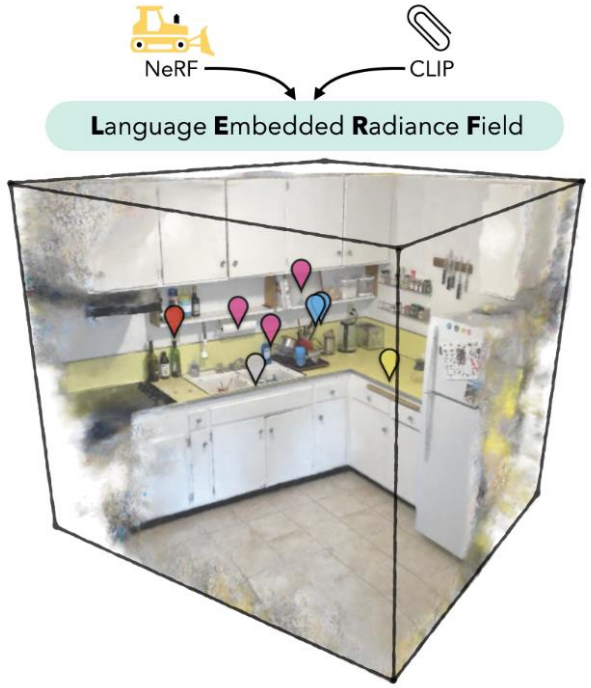
Figure 5. **3D-FRONT NeRF Dataset Samples.** Rows 1–2 show the NeRF reconstruction quality and ground-truth bounding box quality of our 3D-FRONT NeRF dataset. Rows 3–4 show groundtruth boxes with diverse object appearance in the dataset.

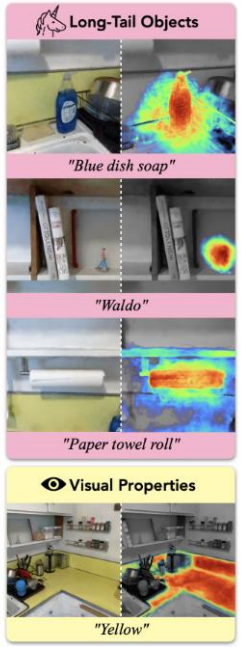
Не только компьютерная графика, но и компьютерное зрение

NeRF-RPN: A general framework for object detection in NeRFs, 2022

LERF: Language Embedded Radiance Fields

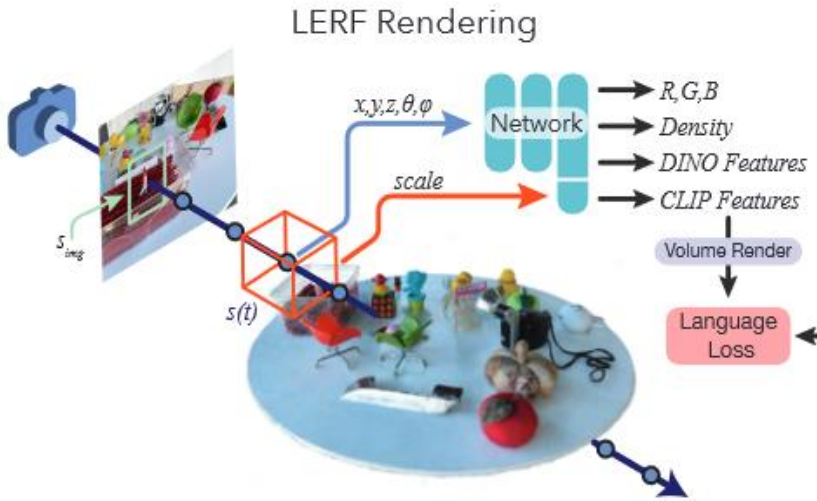


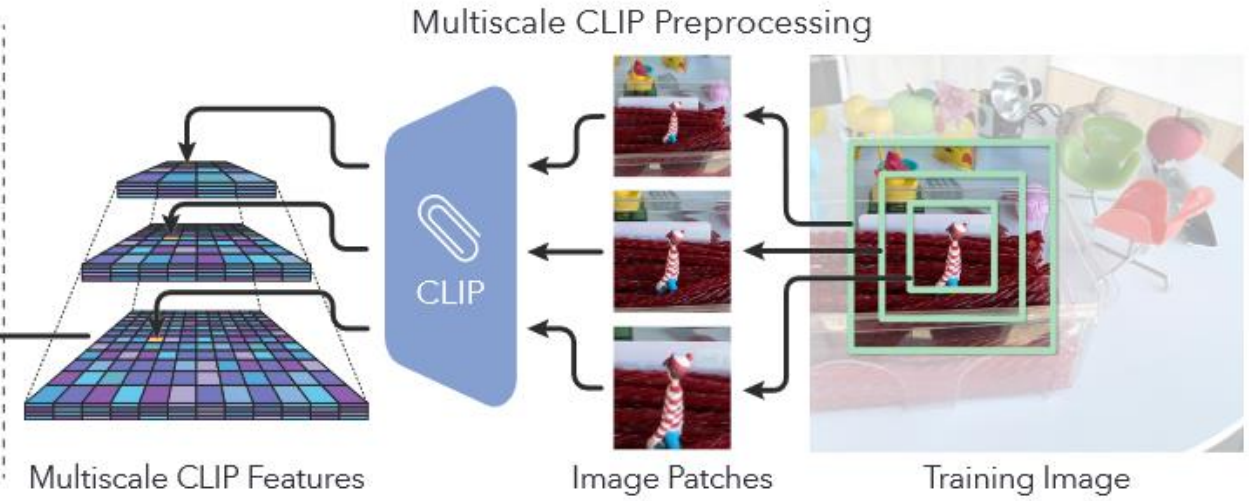




LERF grounds CLIP representations in a dense, multi-scale 3D field. A LERF can be reconstructed from a hand-held phone capture within 45 minutes, then can **render dense relevancy maps given textual queries interactively in real-time**. LERF enables a broad range of concepts to be queried via natural language, from abstract queries like “Electricity”, visual properties like “Yellow”, long-tail objects such as “Waldo”, and even reading text like “Boops” on the mug. For each prompt, an RGB image and relevancy map are rendered focusing on the location with maximum relevancy activation

We pre-compute a multi-scale feature pyramid of CLIP embeddings over training views, and during training interpolate this pyramid with simg and the ray’s pixel location to obtain CLIP supervision. The CLIP loss maximizes cosine similarity, and other outputs are supervised with mean squared-error using standard per-pixel rendering.





LERF: Language Embedded Radiance Fields, UC Berkeley, 2023

Multi-modal Visual-Language Tasks

NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields

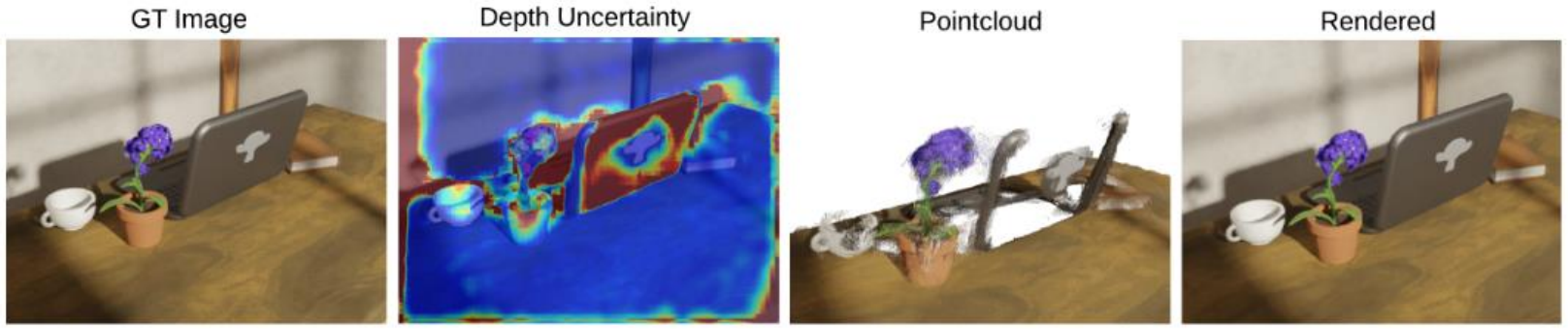


Figure 1. From left to right, input RGB image, estimated depth uncertainty, back-projected depth-maps into a pointcloud, after thresholding the depth by its uncertainty ($\sigma_d \leq 1.0$) for visualization, and our resulting neural radiance field, rendered from the same viewpoint as the input image. Our pipeline is capable of reconstructing neural radiance fields in real-time given only a stream of RGB images.

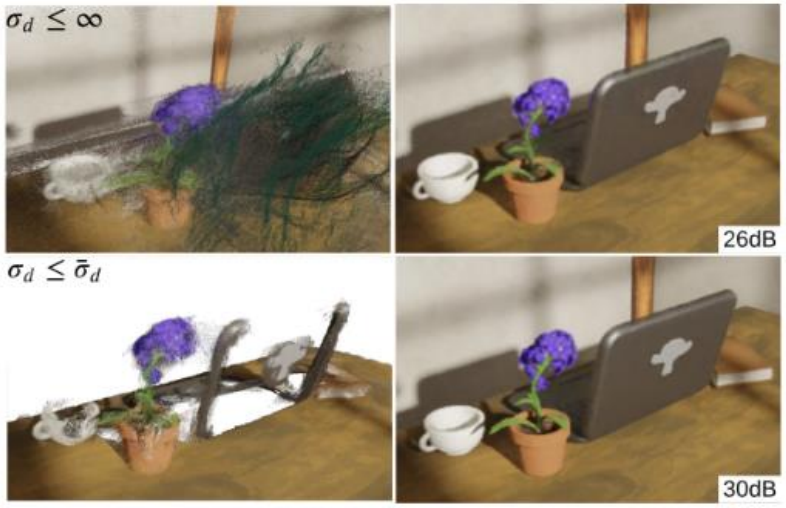


Figure 5. (Top-Left) Raw pointcloud estimated by the tracking module, (Bottom-Left) Pointcloud after thresholding the depth uncertainty ($\sigma_d \leq 1.0$) for visualization. (Right Column) Radiance field reconstructions after 120s of optimization, with and without depth weighting (top-right and bottom-right respectively). Room scene in Cube-Diorama dataset [1].

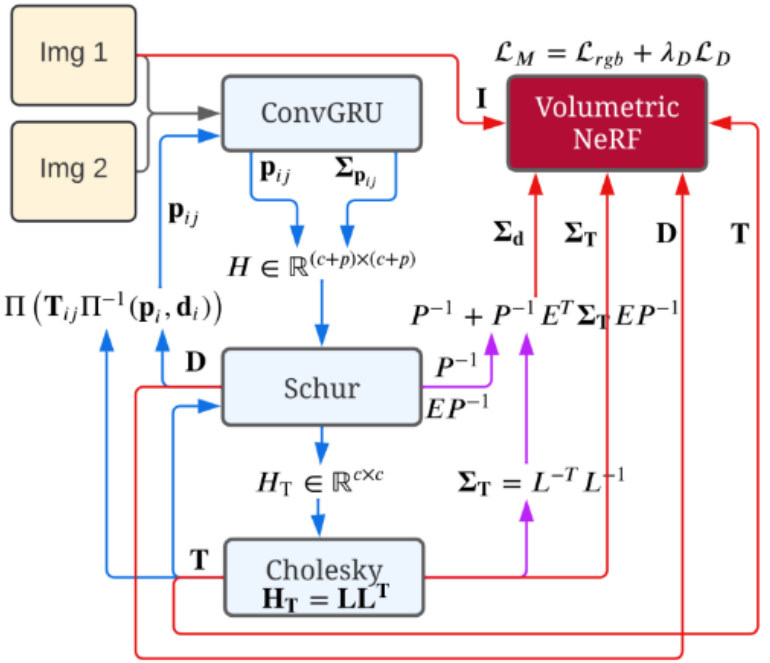


Figure 2. The input to our pipeline consists of sequential monocular images (here represented as Img 1 & Img 2). Starting from the top-right, our architecture fits a NeRF using Instant-NGP [17], which we supervise using RGB images $\mathbf{I}$, depths $\mathbf{D}$, where the depths are weighted by their marginal covariance, $\mathbf{\Sigma_D}$. Inspired by Rosinol et al. [23], we compute these covariances from dense monocular SLAM. In our case, we use Droid-SLAM [31]. We provide more details about the flow of information in Sec. 3.1. In blue, we show Droid-SLAM's [31] contributions and flow of information, similarly, in pink are Rosinol's contribution [23], and in red, our contribution.

SLAM: Ориентация роботов и БЛА в пространстве

Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation

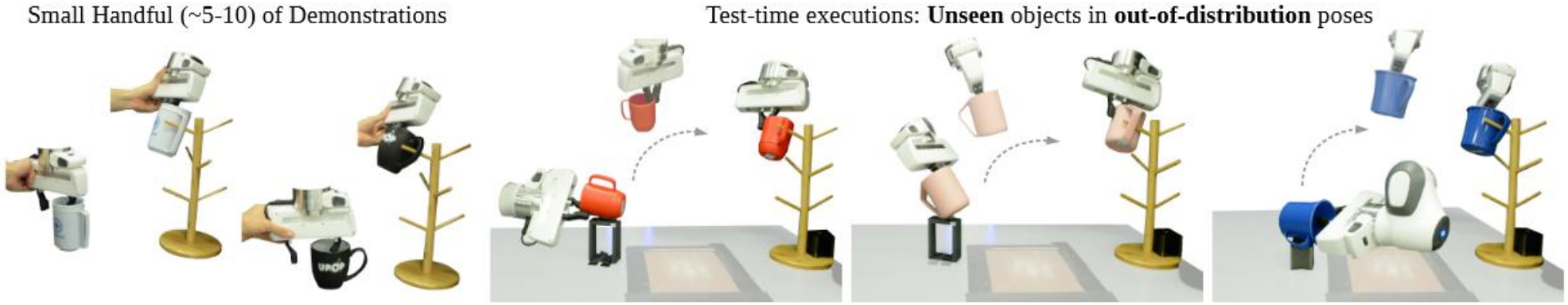


Fig. 1: Given a few (∼5-10) demonstrations of a manipulation task (left), Neural Descriptor Fields (NDFs) generalize the task to novel object instances in any 6-DoF configuration, *including those unobserved at training time*, such as mugs with arbitrary 3D translation and rotation (right). NDFs are continuous functions that map 3D spatial coordinates to spatial descriptors. We generalize this to functions which encode SE(3) poses, such as those used for grasping and placing. NDFs are trained self-supervised for the surrogate task of 3D reconstruction, do not require labeled keypoints, and are SE(3)-equivariant, guaranteeing generalization to unseen object configurations.

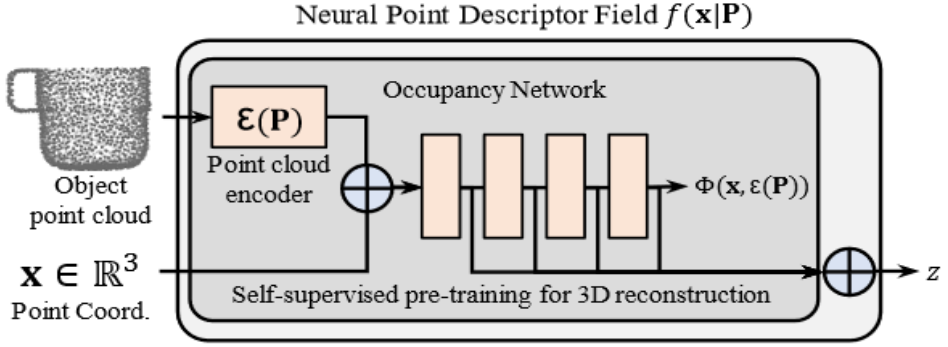


Fig. 2: **Point Descriptor Fields** – We propose to parameterize a Neural Point Descriptor Field f as the concatenation of the layer-wise activations of an occupancy network $\Phi(\mathbf{x}, \mathcal{E}(\mathbf{P}))$. Both the point cloud encoder and the point descriptor function can be pre-trained with a 3D reconstruction task.

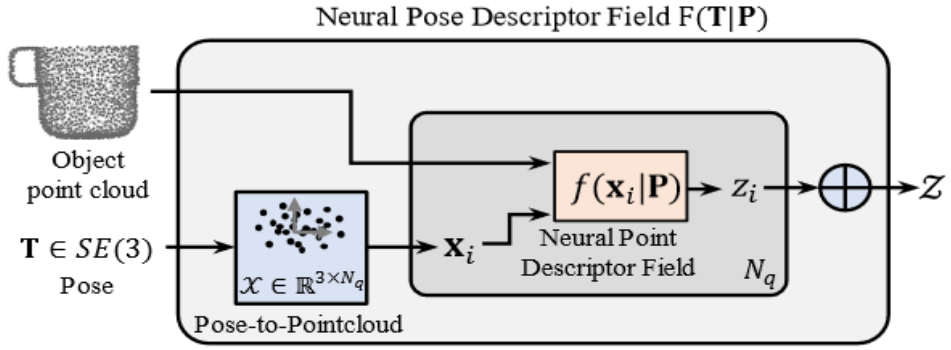


Fig. 3: **Pose Descriptor Fields** – NDFs can extract *pose* descriptors by representing a pose via its action on a *query pointcloud* $\mathcal{X}$, and then extracting point-level spatial descriptors z_i for each point $\mathbf{x}_i$ with a point-level NDF. Concatenation then yields the final pose descriptor $\mathcal{Z}$.

Манипулирование предметами

Neural Fields for Robotic Object Manipulation from a Single Image

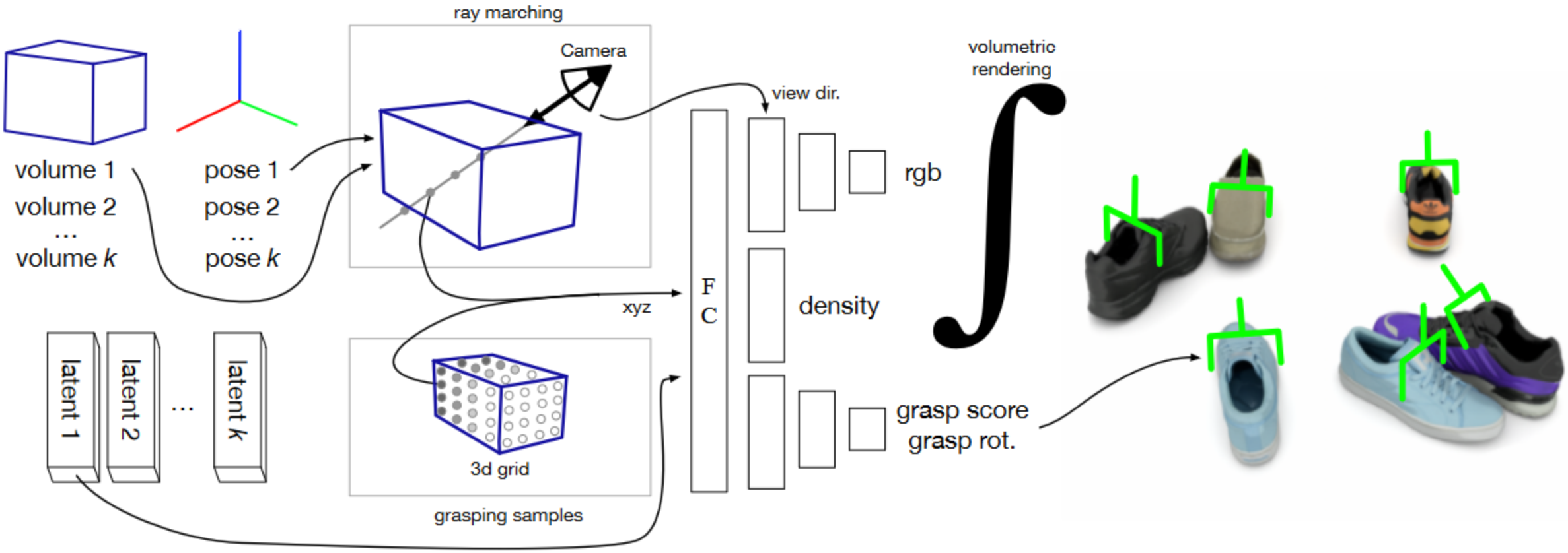


Fig. 1. We present a unified representation that can be used to re-render scenes and generate grasping poses. Given object poses (top left), volumes (top left), and learned latent codes (bottom left), the system renders the scene in novel configurations and annotates each object with a grasping pose (right image). The rendering ray-marches through a set of volumes (cuboids in which the objects reside). In order to generate grasps (shown in green on the right), we use a grid sampling approach that covers each object volume. Grasp orientation and score are predicted using a decoder which is partially shared with the volumetric rendering process. We select the grasp point that yields the highest confidence score. FC refers to a fully connected layer.

Манипулирование предметами

DiffRF: Rendering-Guided 3D Radiance Field Diffusion

Вернемся к задаче генерации изображений 3D сцен и объектов

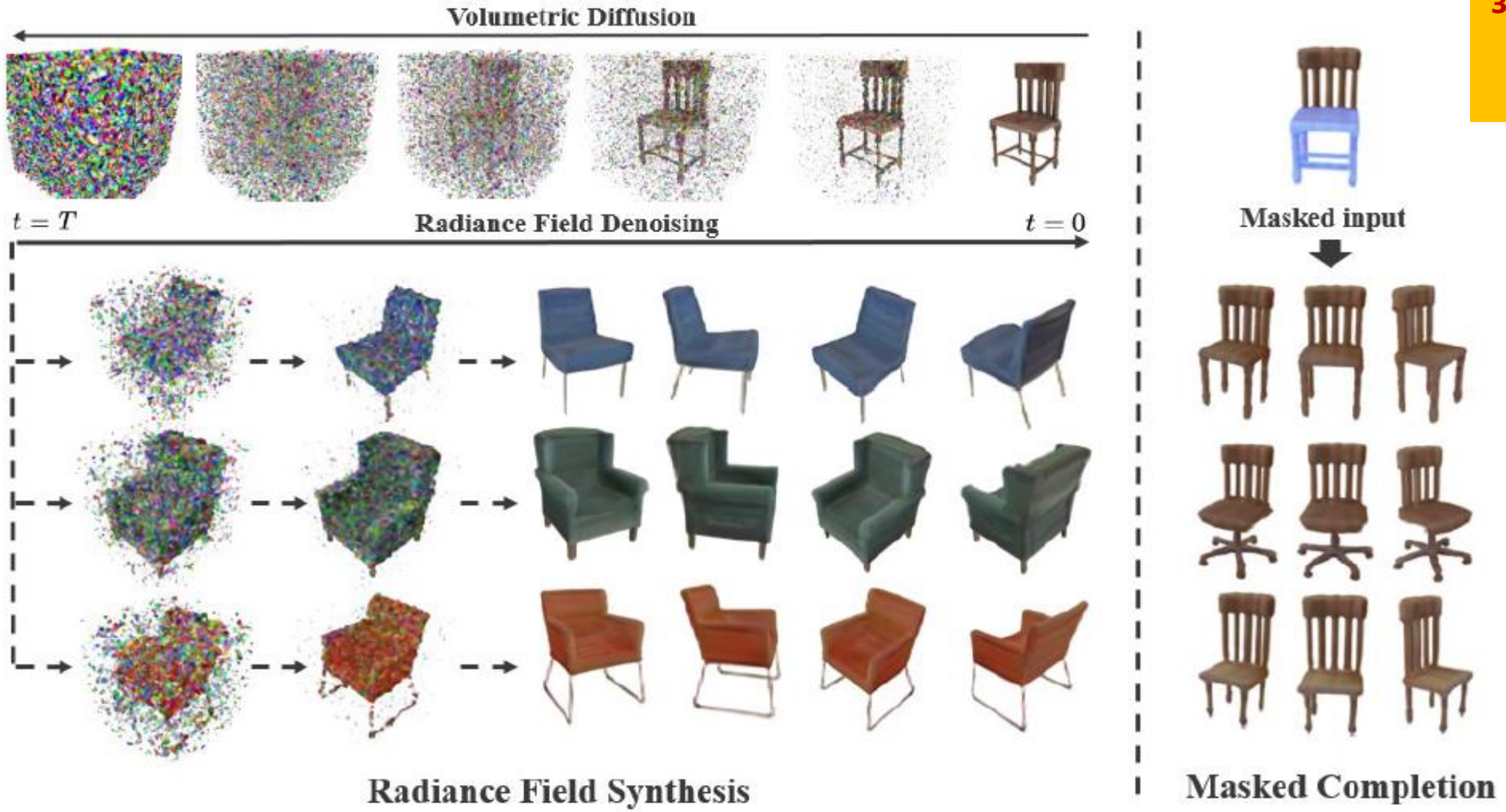


Figure 1. Our method performs denoising of a probabilistic diffusion process applied to 3D radiance fields. Guided by 3D supervision and volumetric rendering, our model enables the unconditional synthesis of high-fidelity 3D assets (left). We further introduce the novel application of *masked completion* (right), *i.e.*, the task of recovering shape and appearance from incomplete objects (highlighted in light-blue on the top right chair), solved by our model as conditional inference without task-specific training.

Соединяем NeRF с диффузными моделями

DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models

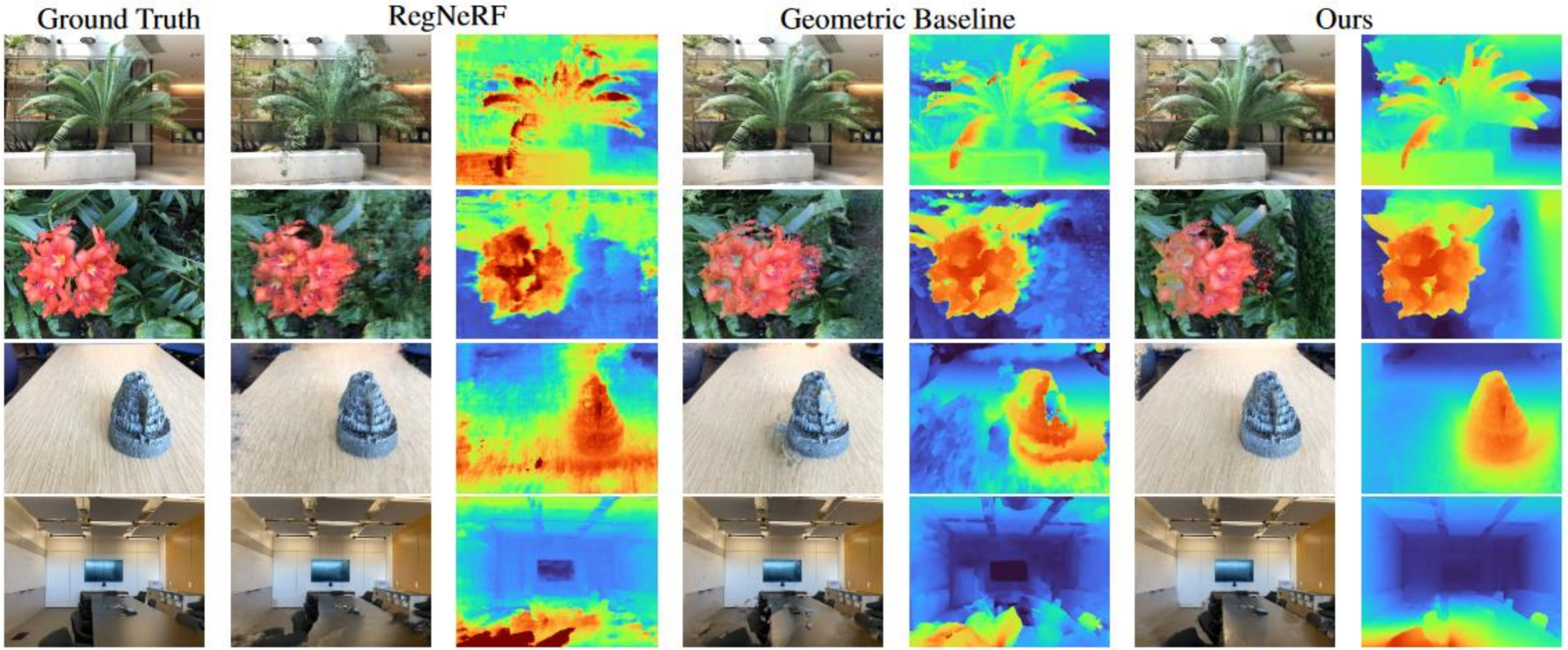


Figure 4. Qualitative results for the task of novel view synthesis on LLFF dataset. NeRF models are trained with 3 views and rendered from one of test views. Our DDM model encourages more realistic geometry as seen in the depth maps.

Соединяем NeRF с диффузными моделями

Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction

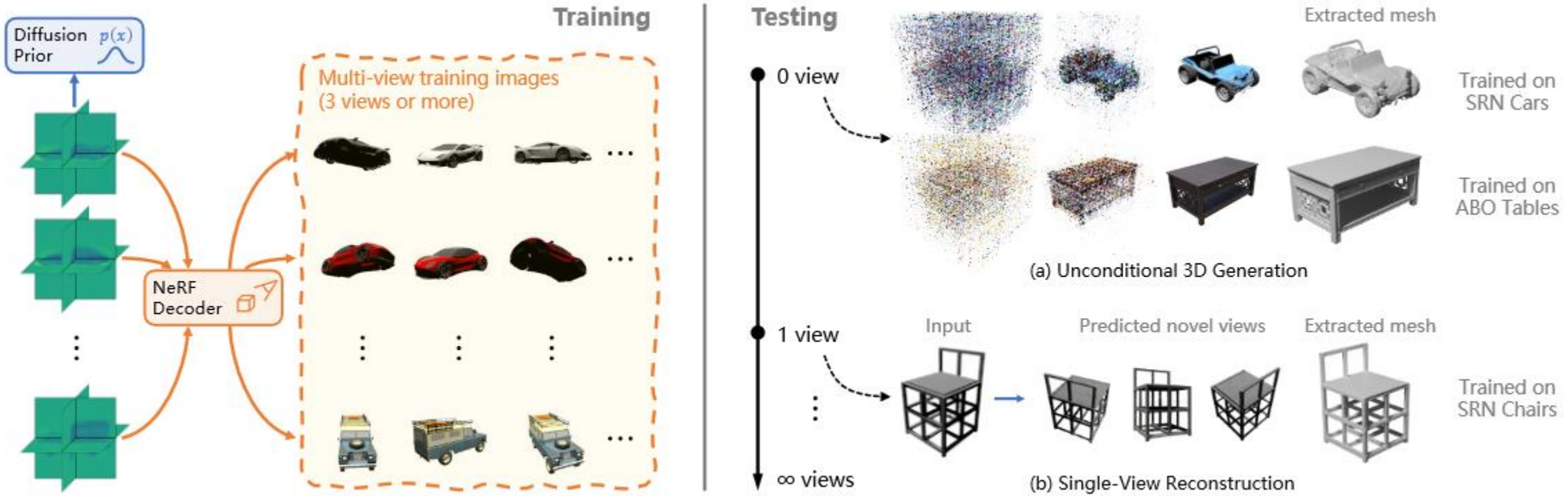
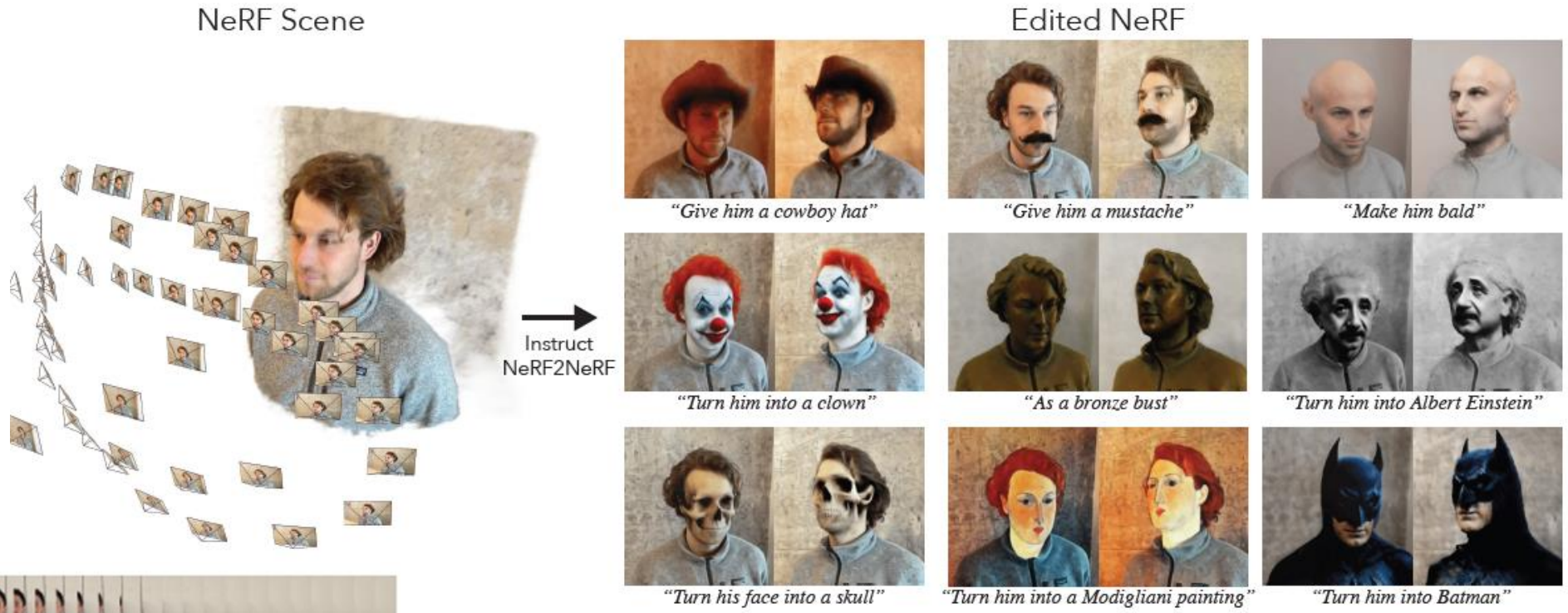


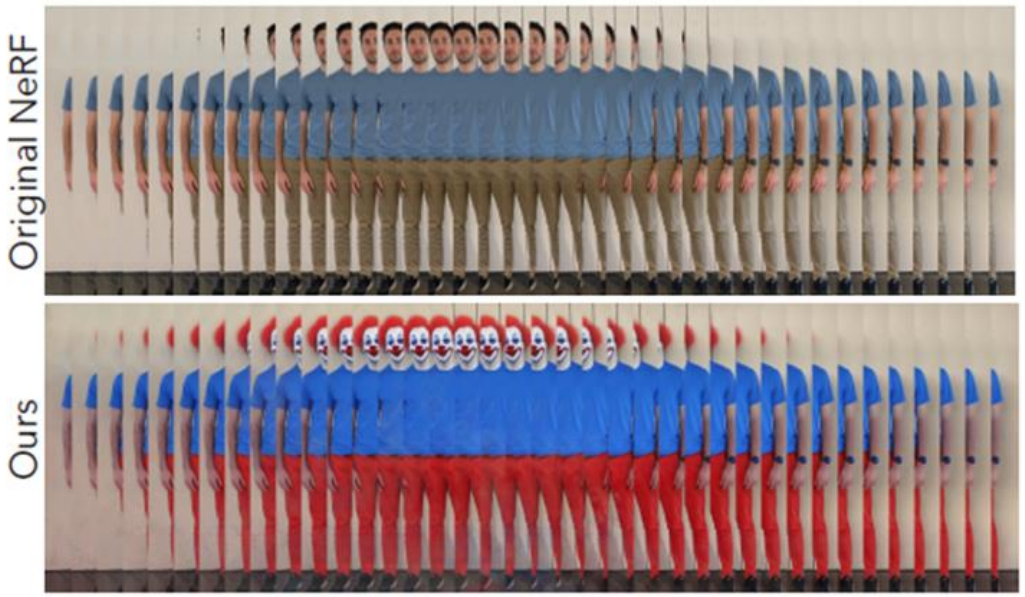
Figure 1. During training, SSDNeRF jointly learns triplane features of individual scenes, a shared NeRF decoder, and a triplane diffusion prior. During testing, it can perform (a) unconditional generation, (b) single-view reconstruction, as well as multi-view reconstruction.

Соединяем NeRF с диффузными моделями

Instruct-NeRF2NeRF: вербальное редактирование 3D сцен

Теперь к уже нейросетевой 3D модели NeRF можно добавить различные генеративные возможности нейросетей, в том числе – синтез и редактирование по текстовому описанию…





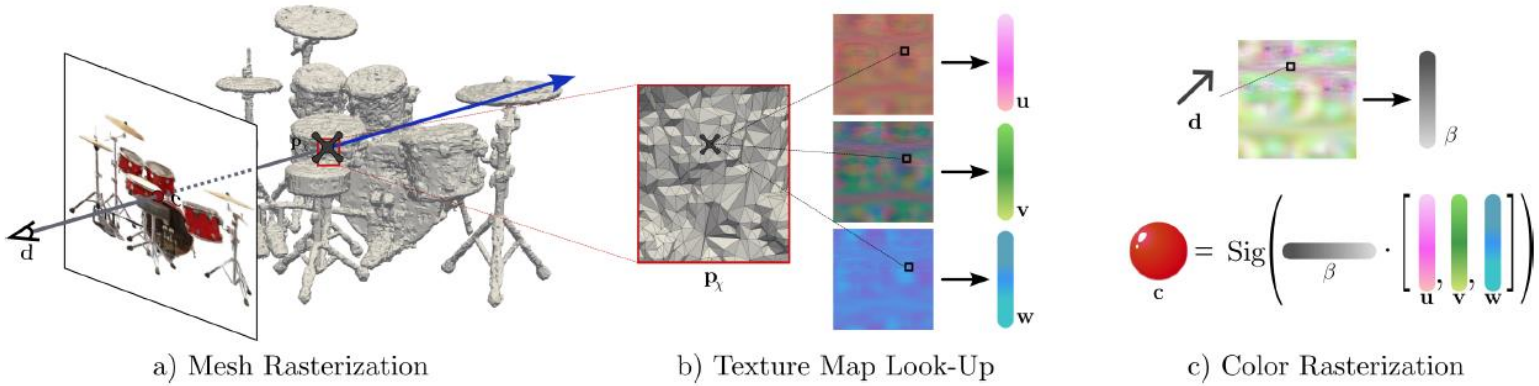
3D consistency. На основе NeRF также можно создавать видео с динамическим движением камеры *в искусственно измененном мире*. Когда камера сдвигается и вращается, измененные элементы сохраняют свой вид.

Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions, 2023

2023

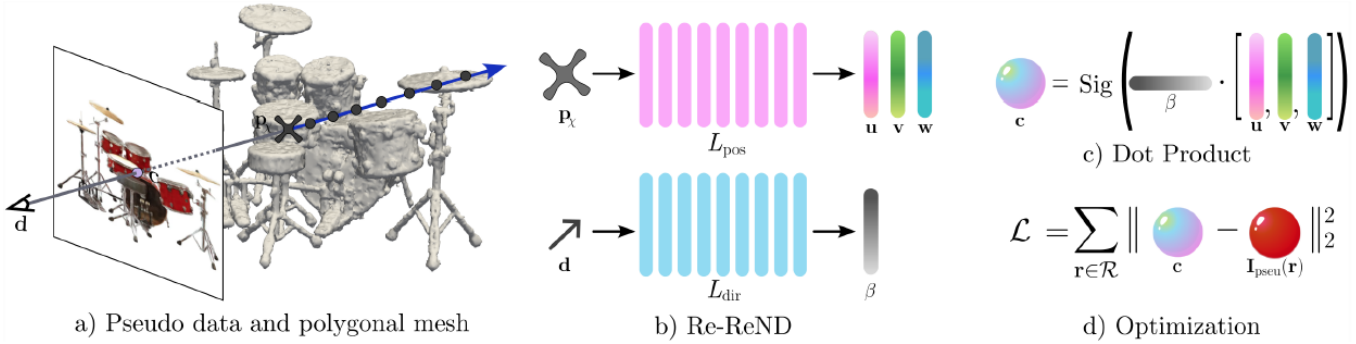
Re-ReND: distill NeRFs to shaders

Как ускорить NeRF



a) Mesh Rasterization b) Texture Map Look-Up c) Color Rasterization

Figure 2. **Rendering a NeRF using** Re-ReND**.** Given a pre-trained NeRF, Re-ReND renders it in real time by leveraging a mesh and light field embedding maps ($\mathbf{u}, \mathbf{v}, \mathbf{w}$ and β) which are compatible with a standard rasterization pipeline with programmable shaders. Rendering occurs in three steps: **a)** The mesh gets rasterized, and its vertices go through a regular vertex shader. **b)** The light field embedding maps are indexed (with the uv coordinates from the rasterization) to obtain values for the position-dependent ($\mathbf{u}, \mathbf{v}$ and $\mathbf{w}$) and direction-dependent (β) embeddings. **c)** Pixel colors are obtained by combining $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and β in a custom fragment shader implementing an inexpensive dot product. Since Re-ReND uses light fields, alpha compositing is unnecessary, and so we set alpha values to 1. Note that our method entirely disposes of MLPs at render time, and thus enjoys substantial boosts in rendering speed.



a) Pseudo data and polygonal mesh b) Re-ReND c) Dot Product d) Optimization

Figure 3. **Training** Re-ReND**. a)** Starting with a pre-trained NeRF, we extract a polygonal mesh and generate pseudo-images to train our factorized NeLF. These images are generated by rendering the NeRF from various points of view. **b)** Our factorized NeLF consists of two MLPs (L_{pos} and L_{dir}) that *independently* compute position and direction embeddings. L_{pos} outputs position-dependent deep radiance embeddings $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ for points *on* the mesh's surface, while L_{dir} outputs direction-dependent embeddings (β). Since this representation is amenable to "baking", *i.e.* pre-computing and storing outputs, it allows us to dispose of MLPs for deployment. **c)** Under this formulation, computing pixel color amounts to a dot product of $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ and β. **d)** We optimize our framework with an MSE reconstruction loss w.r.t. to the pseudo-images we extracted from the pre-trained NeRF.

We introduce Re-ReND, a method enabling Real-time Rendering of NeRFs across Devices. **Re-ReND is designed to achieve real-time performance by converting the NeRF into a representation that can be efficiently processed by standard graphics pipelines**. Re-ReND preserves remarkable photo-metric quality even when rendering at over **1,013 FPS on a desktop browser**, or at the capped **74 FPS of a VR headset**

Re-ReND: Real-time Rendering of NeRFs across Devices. 2023.

Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields

Как улучшить и ускорить NeRF

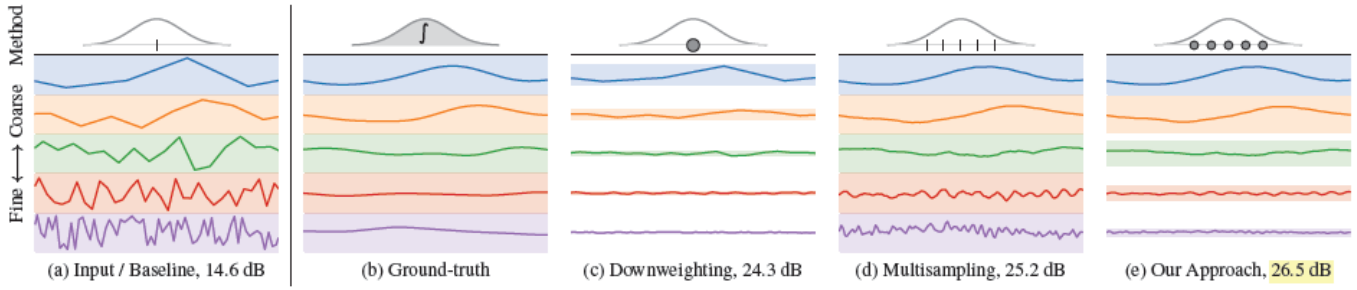


Figure 2: Here we show a toy 1-dimensional iNGP [21] with 1 feature per scale. Each subplot represents a different strategy

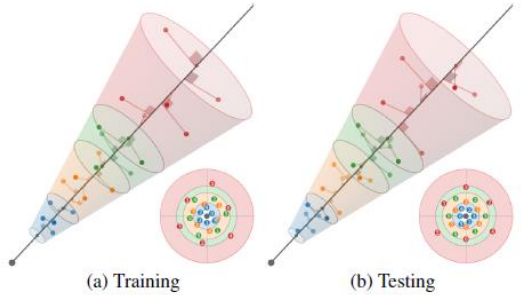


Figure 3: Here we show a toy 3D ray with an exaggerated pixel width (viewed along the ray as an inset) divided into 4 frustums denoted by color. We multisample each frustum with a hexagonal pattern that matches the frustum's first and second moments. Each pattern is rotated around the ray and flipped along the ray (a) randomly when training and (b) deterministically when rendering.

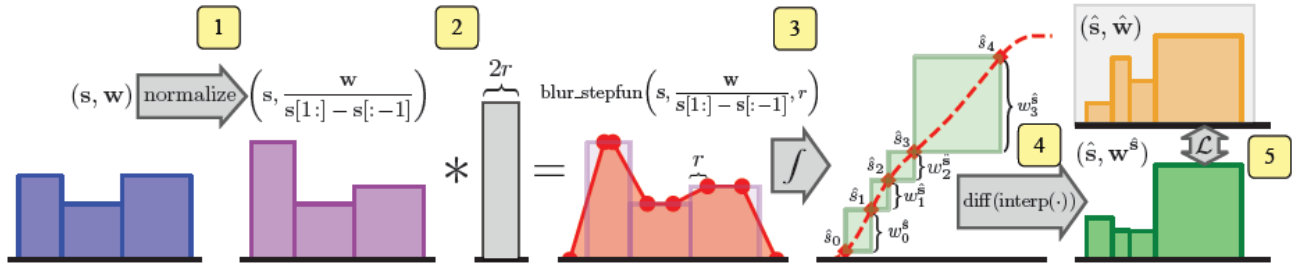


Figure 5: Computing our anti-aliased loss requires that we smooth and resample a NeRF histogram $(\mathbf{s}, \mathbf{w})$ into the same set of endpoints as a proposal histogram $(\hat{\mathbf{s}}, \hat{\mathbf{w}})$, which we outline here. (1) We divide $\mathbf{w}$ by the size of each interval in $\mathbf{s}$ to yield a piecewise constant PDF that integrates to ≤ 1. (2) We convolve that PDF with a rectangular pulse to obtain a piecewise linear PDF. (3) This PDF is integrated to produce a piecewise-quadratic CDF that is queried via piecewise quadratic interpolation at each location in $\hat{\mathbf{s}}$. (4) By taking the difference between adjacent interpolated values we obtain $\mathbf{w}^{\hat{\mathbf{s}}}$, which are the NeRF histogram weights $\mathbf{w}$ resampled into the endpoints of the proposal histogram $\hat{\mathbf{s}}$. (5) After resampling, we evaluate our loss $\mathcal{L}_{\text{prop}}$ as an element-wise function of $\mathbf{w}^{\hat{\mathbf{s}}}$ and $\hat{\mathbf{w}}$, as they share a common coordinate space.

1) MLPs are slow to train: The original **NeRF** model used a multilayer perceptron (MLP) to parameterize the mapping from spatial coordinates to colors and densities. Though compact and expressive, and recent work has accelerated training by replacing or augmenting MLPs with voxelgrid-like datastructures. One example **is Instant NGP (iNGP)**, which uses a pyramid of coarse and fine grids (the finest of which are stored using a hash map) to construct learned features that are processed by a tiny MLP, enabling greatly accelerated training.

2) In addition to being slow, the original NeRF model was also aliased: NeRF reasons about individual points along a ray, which results in "jaggies" in rendered images and limits NeRFs ability to reason about scale. **Mip-NeRF** resolved this issue by casting cones instead of rays, and by featurizing the entire volume within a conical frustum foruse as input to the MLP. Mip-NeRF and its successor **mipNeRF 360** showed that this approach enables highly accurate rendering on challenging real-world scenes.

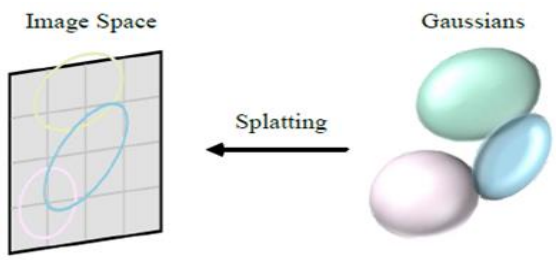
Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. Google Research 2023.

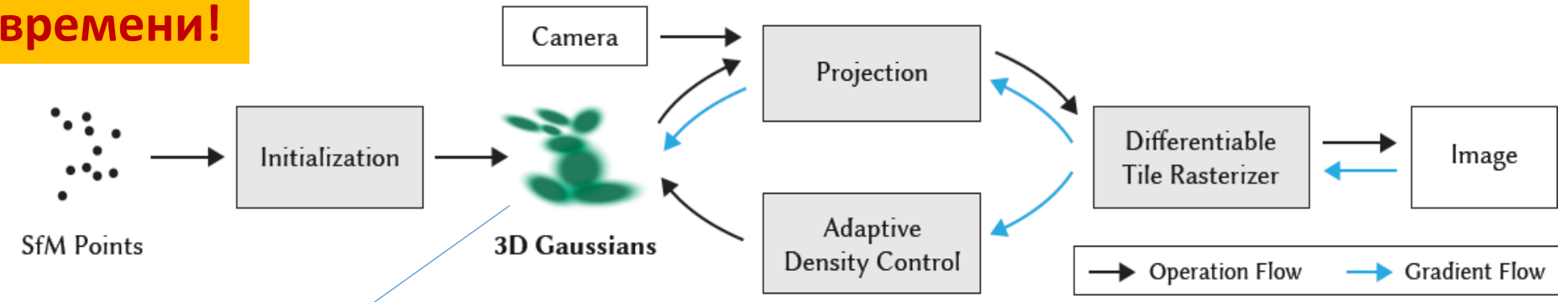
3D Gaussian Splatting (3DGS)

Предобученные по ракурсным изображениям описания 3D сцен

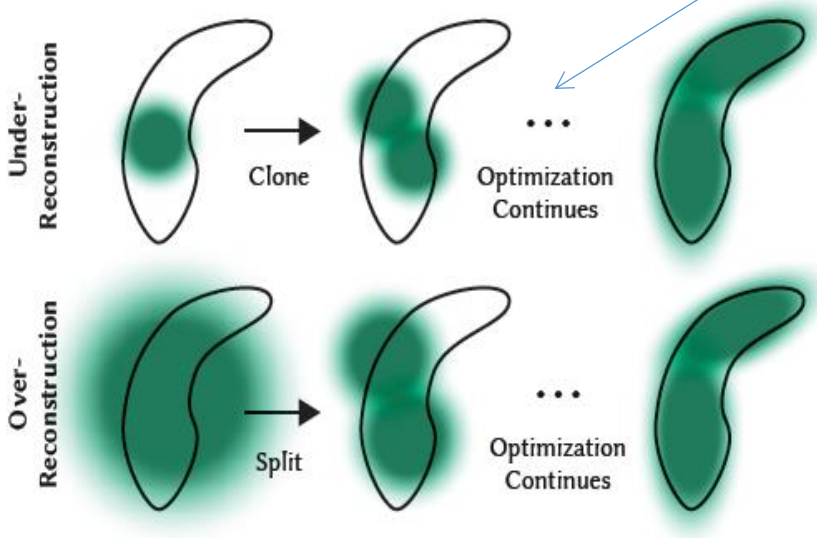
3D Gaussian Splatting for Real-Time Radiance Field Rendering 2023

Рендеринг в реальном времени!





Adaptive Gaussian densification scheme: split and merge centroids clustering. When small-scale geometry (black outline) is insufficiently covered, we clone the respective Gaussian. If small-scale geometry is represented by one large splat, we split it in two.



- The **anisotropic 3D Gaussians as a high-quality, unstructured representation of radiance fields**.
- An **optimization method of 3D Gaussian properties, interleaved with adaptive density control** for representations for captured scenes.
- A **fast, differentiable rendering approach for the GPU, which is visibility-aware, allows anisotropic splatting and fast backpropagation** to achieve high-quality novel view synthesis.

3D Gaussian Splatting for Real-Time Radiance Field Rendering. 2023.

3D Gaussian Splatting for Real-Time Radiance Field Rendering

Fig. 6. For some scenes (above) we can see that even at 7K iterations (~5min for this scene), our method has captured the train quite well. At 30K iterations (~35min) the background artifacts have been reduced significantly. For other scenes (below), the difference is barely visible; 7K iterations (~8min) is already very high quality.

Fig. 7. Initialization with SfM points helps. Above: initialization with a random point cloud. Below: initialization using SfM points.

Fig. 8. Ablation of densification strategy for the two cases "clone" and "split" (Sec. 5).

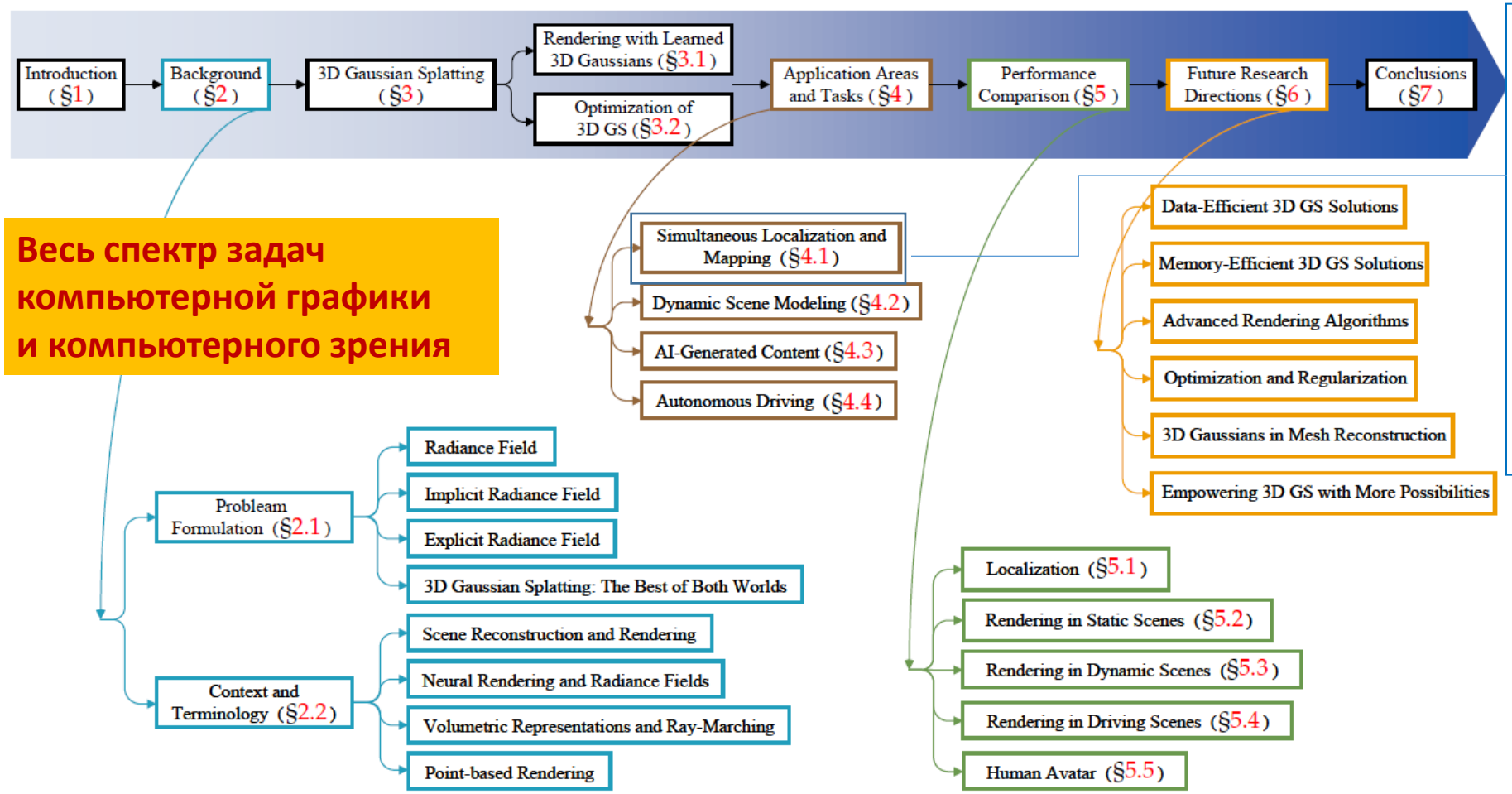
Не только рендеринг в реальном времени, но и быстрое обучение!

Our method achieves real-time rendering of radiance fields with quality that equals the previous method with the best quality [Barron et al. 2022], while only requiring optimization times competitive with the fastest previous methods. Note that for comparable training times to InstantNGP [Müller et al. 2022], we achieve similar quality to theirs; **by training for 51min we achieve state-of-the-art quality**, even slightly better than Mip-NeRF360 [Barron et al. 2022].

Новый стандарт в обучаемых Radiance Field Representations!

3D Gaussian Splatting for Real-Time Radiance Field Rendering. 2023.

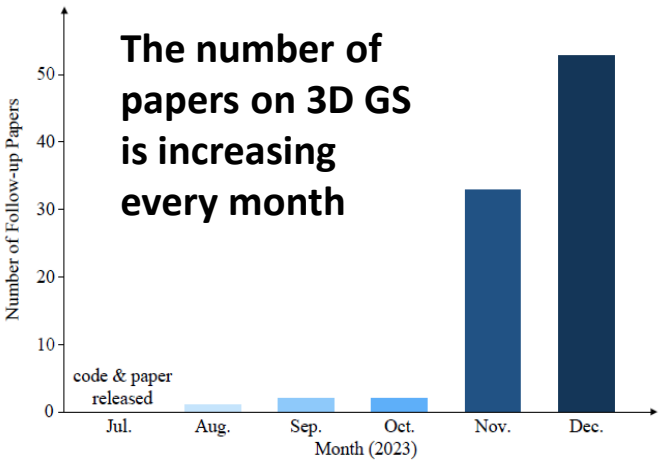
3D Gaussian Splatting: Game Changing



Весь спектр задач компьютерной графики и компьютерного зрения

Method		3D GS
F2F* [137]	[ICCV17]	
iMAP [138]	[ICCV21]	
Vox-Fusion [139]	[ISMAR22]	
NICE-SLAM [140]	[CVPR22]	
ESLAM [141]	[CVPR23]	
Point-SLAM [142]	[ICCV23]	
Co-SLAM [143]	[CVPR23]	
F2GM+F2F [64]	[arXiv]	✓
GSSLAM [63]	[arXiv]	✓
SplaTAM [66]	[arXiv]	✓
GS-SLAM [65]	[arXiv]	✓

Например: 3D GS для SLAM



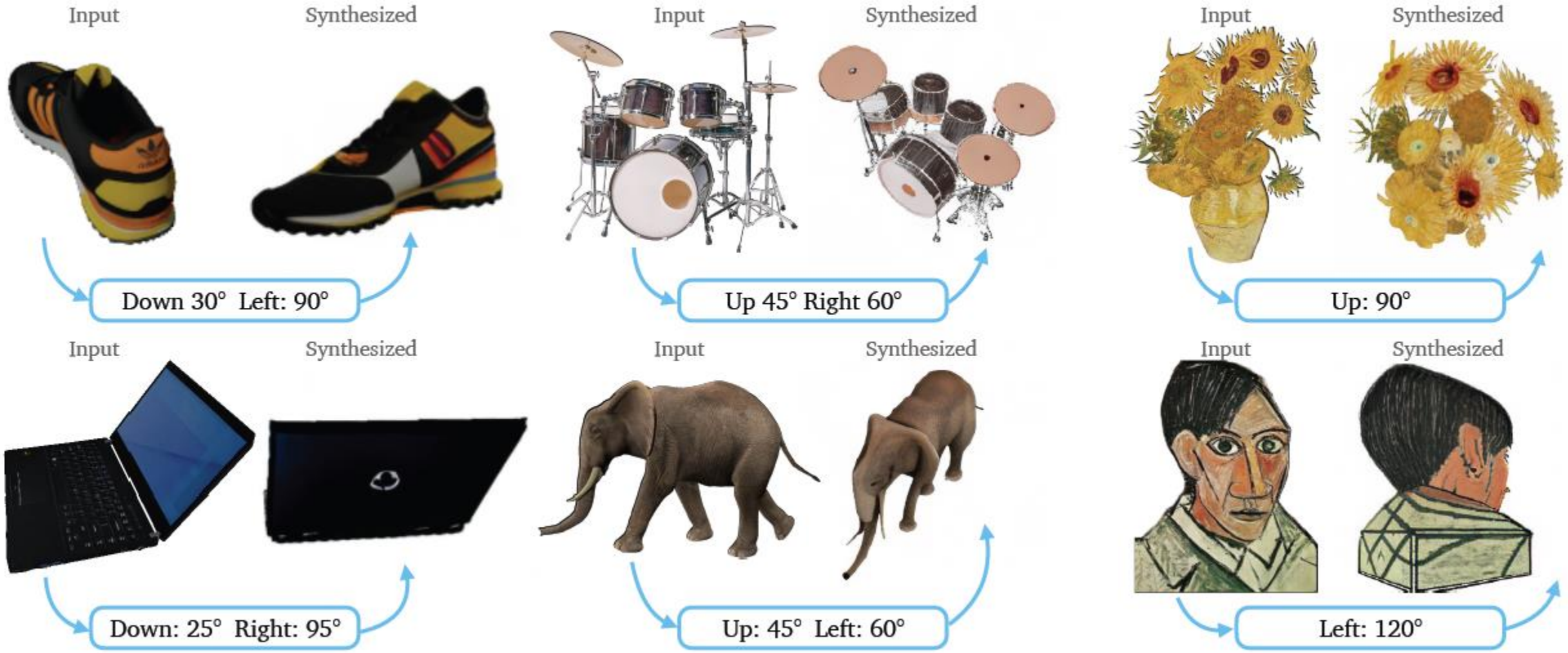
3D Gaussian splatting (3D GS) represents a significant departure from NeRF.
3D GS, with its explicit scene representations and differentiable rendering algorithms, not only promises real-time rendering capabilities but also introduces unprecedented levels of control and editability. **This positions 3D GS as a potential game-changer for the next generation of 3D reconstruction and representation**

A Survey on 3D Gaussian Splatting. 2024.

Новый стандарт в обучаемых Radiance Field Representations!

Stable Zero-123: Zero-shot One Image to 3D Object

Новый ракурс по одному снимку



Given a single RGB image of an object, we present Zero-1-to-3, a method to synthesize an image from a specified camera viewpoint. Our approach synthesizes views that contain rich details consistent with the input view for large relative transformations. It also achieves strong zero-shot performance on objects with complex geometry and artistic styles.

Zero-1-to-3: Zero-shot One Image to 3D Object. 2023.

Stable Zero-123: Zero-shot One Image to 3D Object

Новый ракурс по одному снимку

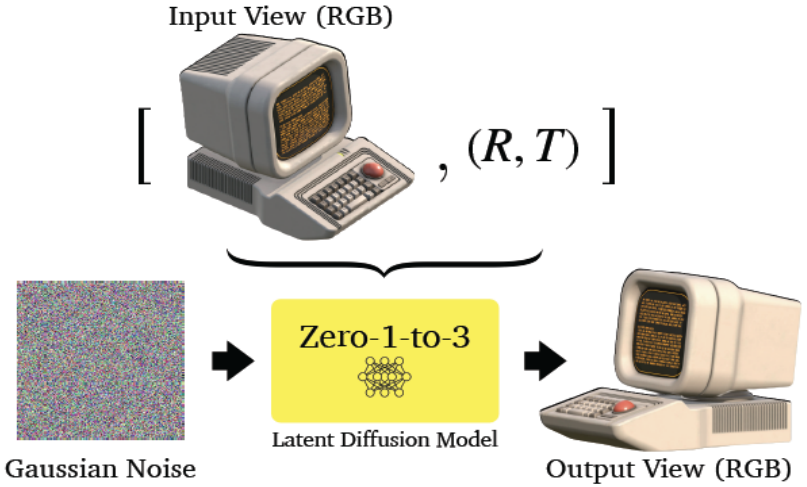


Figure 3: **Zero-1-to-3** is a viewpoint-conditioned image translation model using a conditional latent diffusion architecture. Both the input view and a relative viewpoint transformation are used as conditional information.

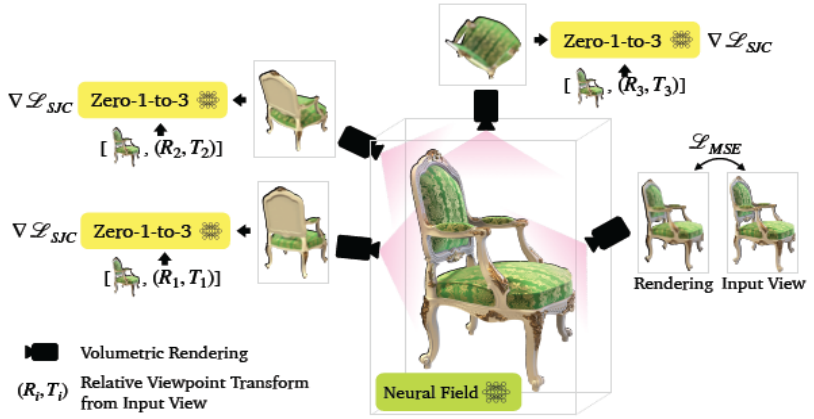
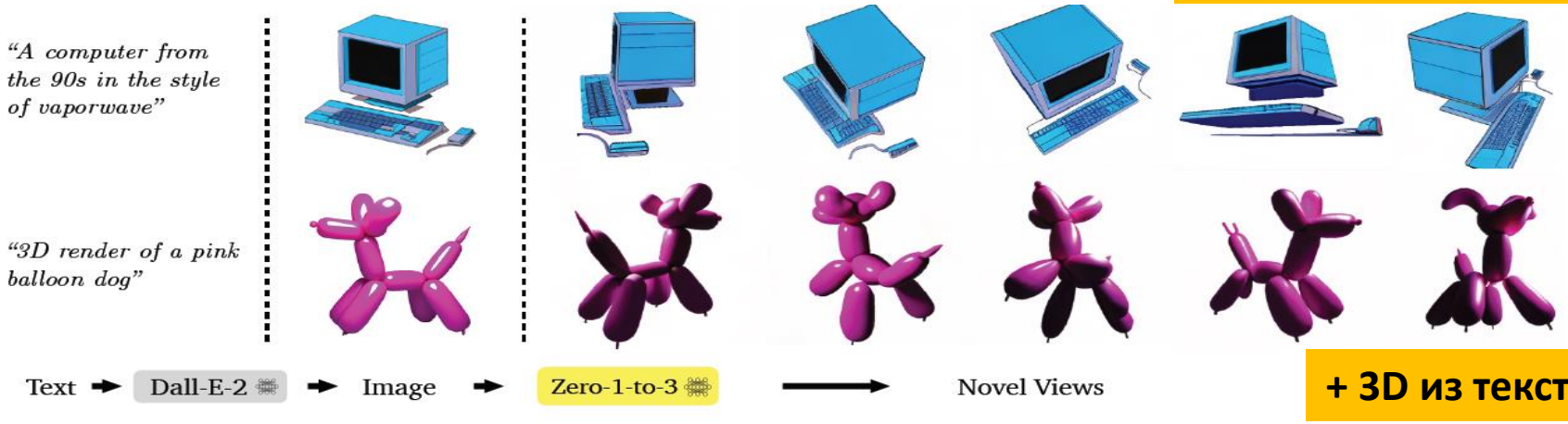


Figure 4: **3D reconstruction with Zero-1-to-3.** Zero-1-to-3 can be used to optimize a neural field for the task of 3D reconstruction from a single image. During training, we randomly sample viewpoints and use Zero-1-to-3 to supervise the 3D reconstruction.

+ 3D из изображения



+ 3D из текста!

Figure 10: **Novel View Synthesis from Dall-E-2 Generated Images.** The composition of multiple objects (1st row) and the lighting details (2nd row) are preserved in our synthesized novel views.

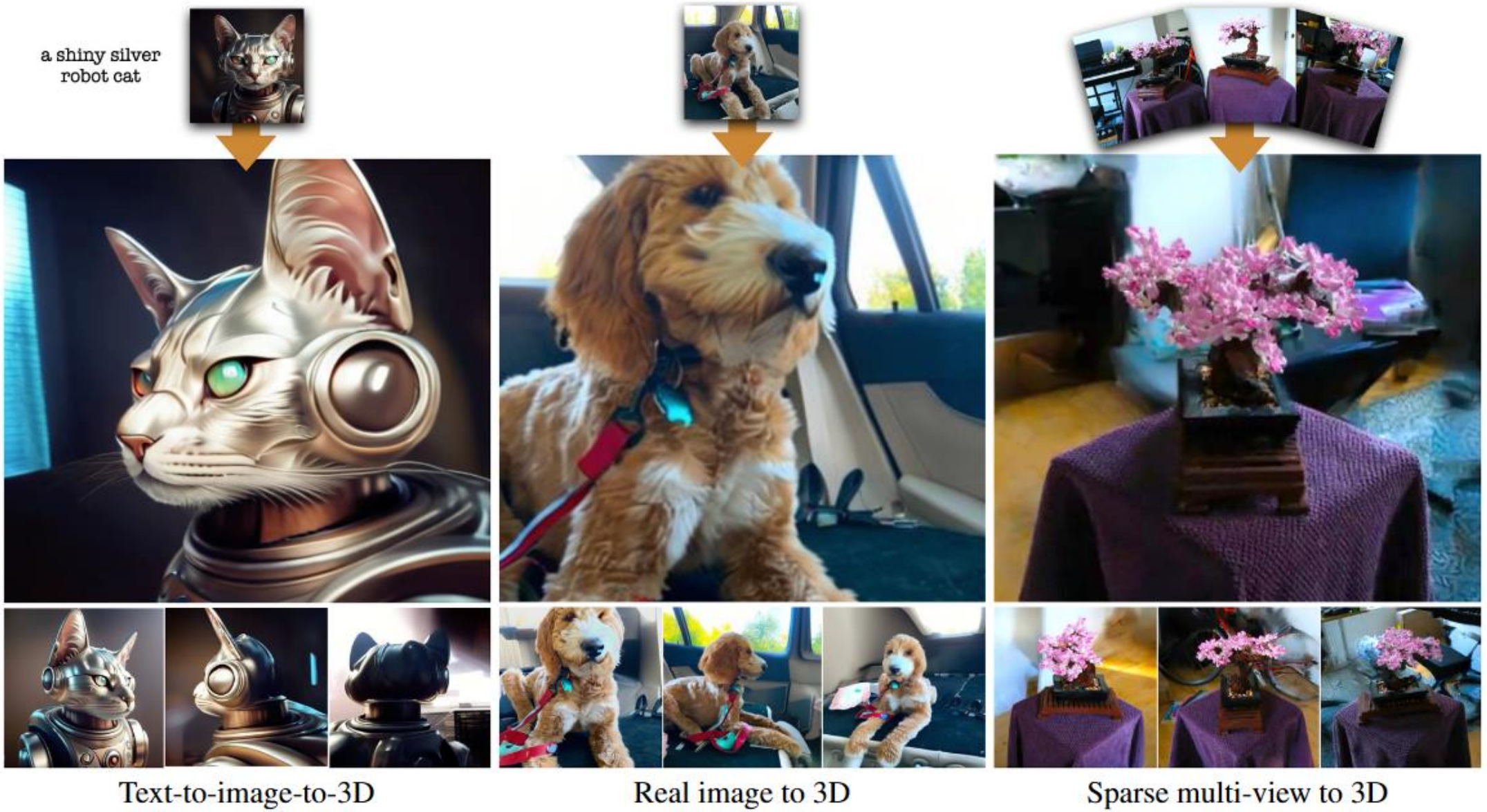
Our conditional diffusion model uses a synthetic dataset to learn controls of the relative camera viewpoint, which allow new images to be generated of the same object under a specified camera transformation. Even though it is trained on a synthetic dataset, our model retains a strong zero-shot generalization ability to out-of-distribution datasets as well as in-the-wild images, including impressionist paintings. **Our viewpoint-conditioned diffusion approach can further be used for the task of 3D reconstruction from a single image.** Qualitative and quantitative experiments show that our method significantly **outperforms stateof-the-art single-view 3D reconstruction and novel view synthesis models by leveraging Internet-scale pre-training.**

Zero-1-to-3: Zero-shot One Image to 3D Object. 2023.

Text-2-3D
3D Avatars

CAT3D: 3D with Multi-View Diffusion Models

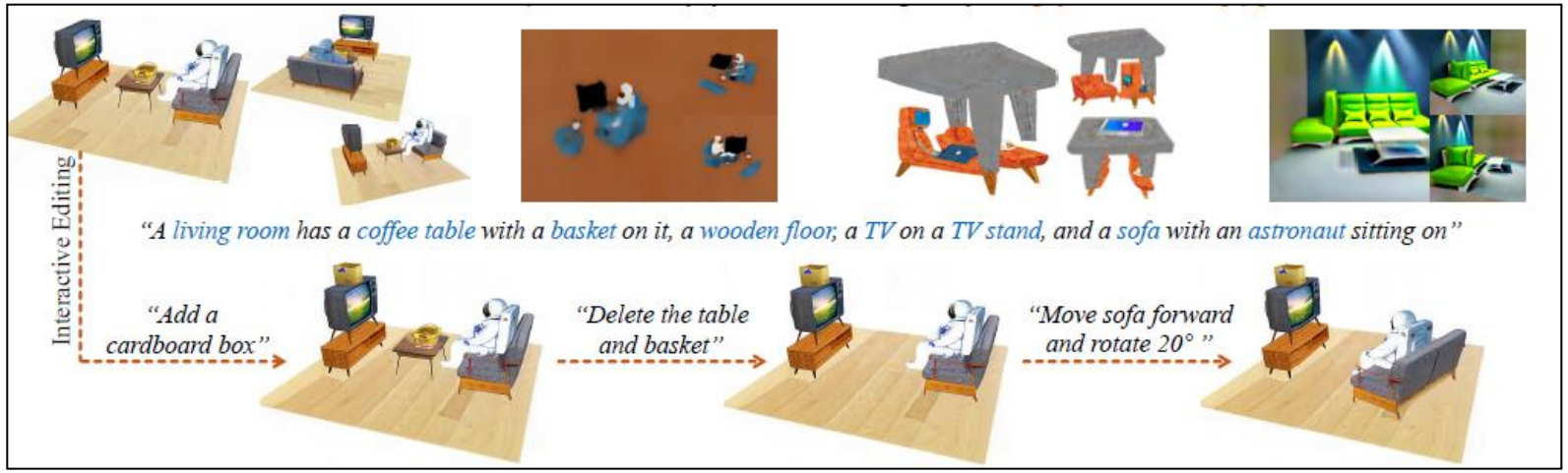
Новый ракурс по одному снимку -> 3D



We present CAT3D, **a method for creating anything in 3D by simulating this real-world capture process with a multi-view diffusion model**. Given any number of input images and a set of target novel viewpoints, our model generates highly consistent novel views of a scene. These generated views can be used as input to robust 3D reconstruction techniques to produce 3D representations that can be rendered from any viewpoint in real-time. CAT3D can create entire 3D scenes in as little as one minute, and outperforms existing methods for single image and few-view 3D scene creation.
See our project page for results and interactive demos: cat3d.github.io.

CAT3D: Create Anything in 3D with Multi-View Diffusion Models. 2024.

GALA3D: Text-to-3D Complex Scene Generation via Layout-guided GS

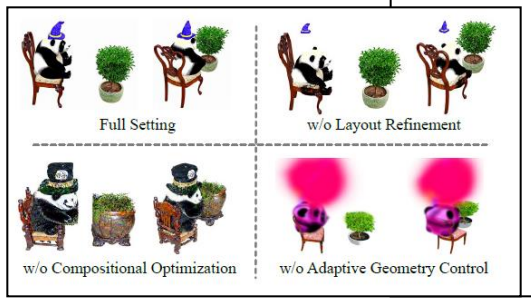


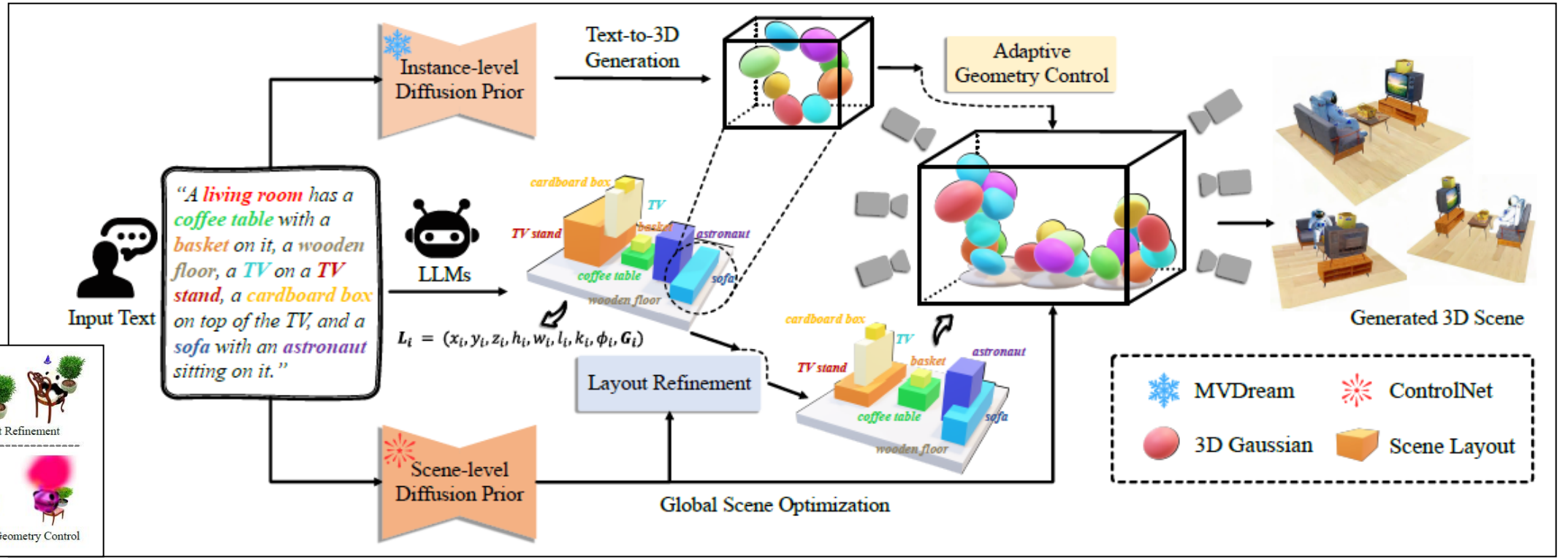
LLM + 3D Gaussian Splatting

1) LLM создает 3D макет
2) строится связанная с ним 3D GS модель
3) одновременно уточняем 3D GS модель и макет сцены

3D модель сцены может редактироваться при помощи текстовых указаний!

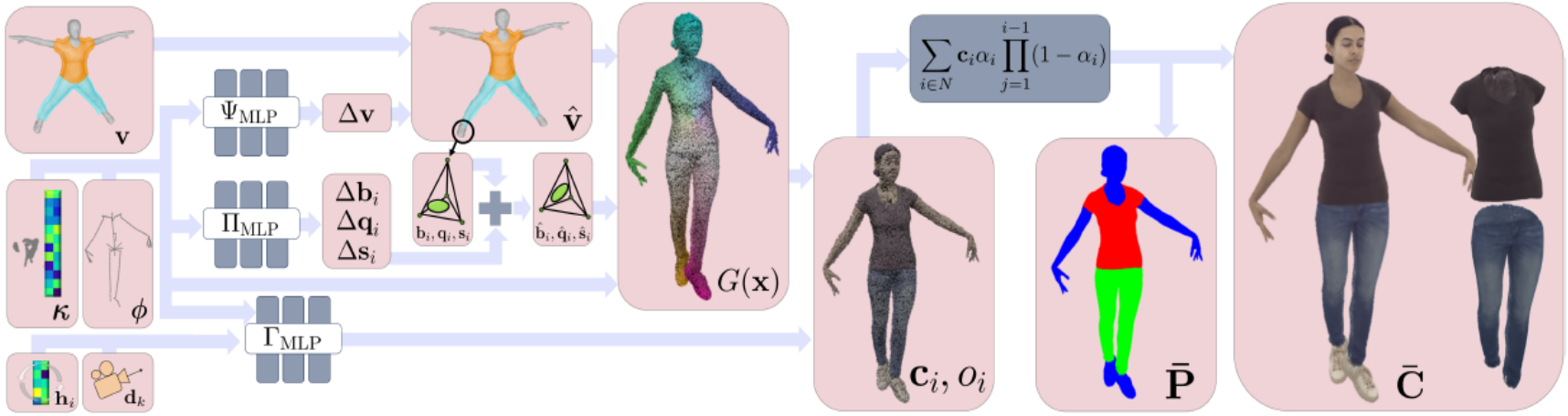
GALA3D outperforms existing methods in text-to-3D scene generation





GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. 2024.

Drivable 3D Gaussian Avatars



We present Drivable 3D Gaussian Avatars (D3GA), the first 3D controllable model for human bodies rendered with Gaussian splats. This work uses the recently presented **3D Gaussian Splatting (3DGS) technique to render realistic humans at real-time framerates**. To deform those primitives, we depart from the commonly used point deformation method of **linear blend skinning (LBS)** and use a **classic volumetric deformation method: cage deformations**.

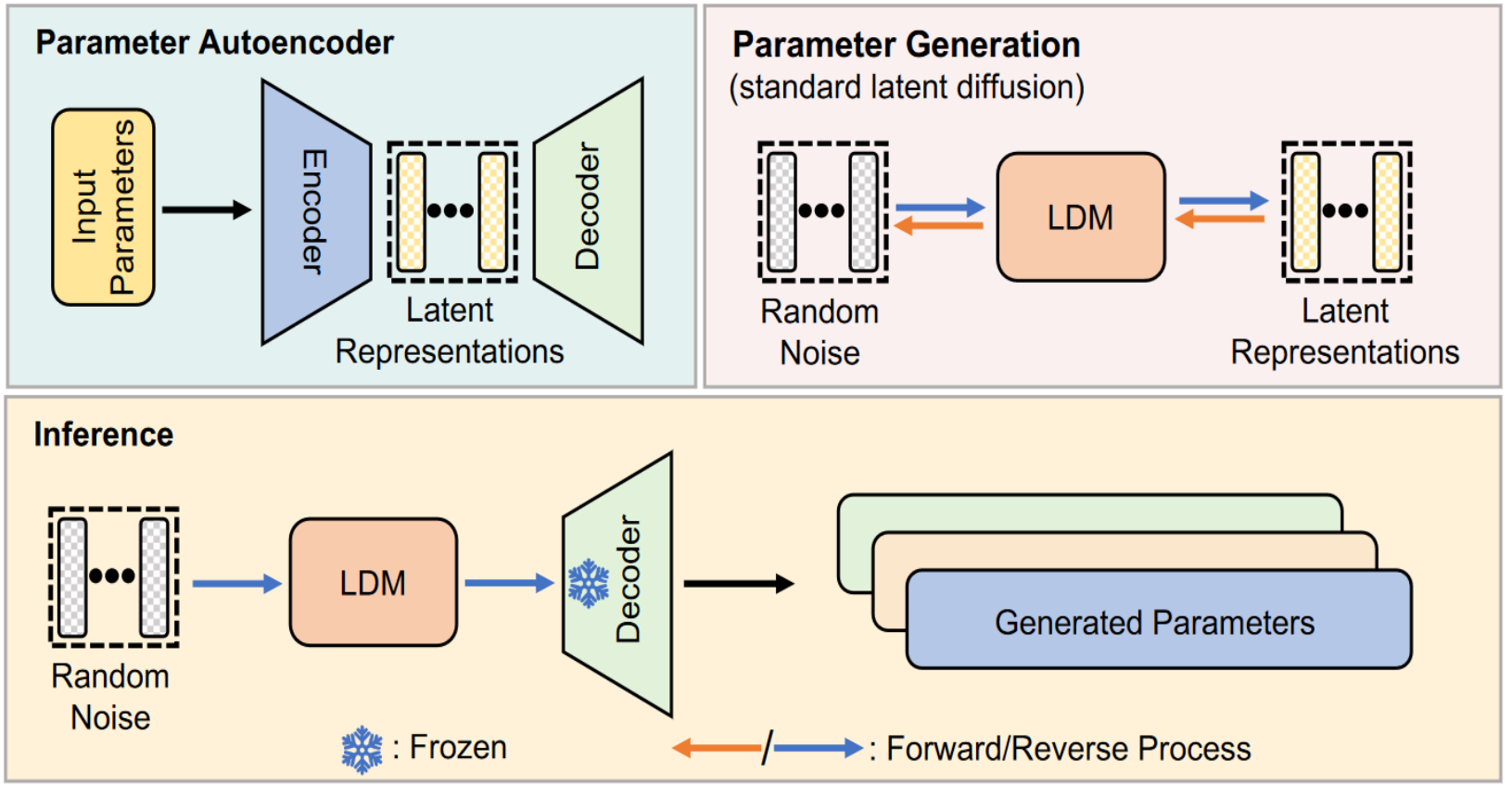
Figure 8. Reposing eight avatars with a pose from another subject

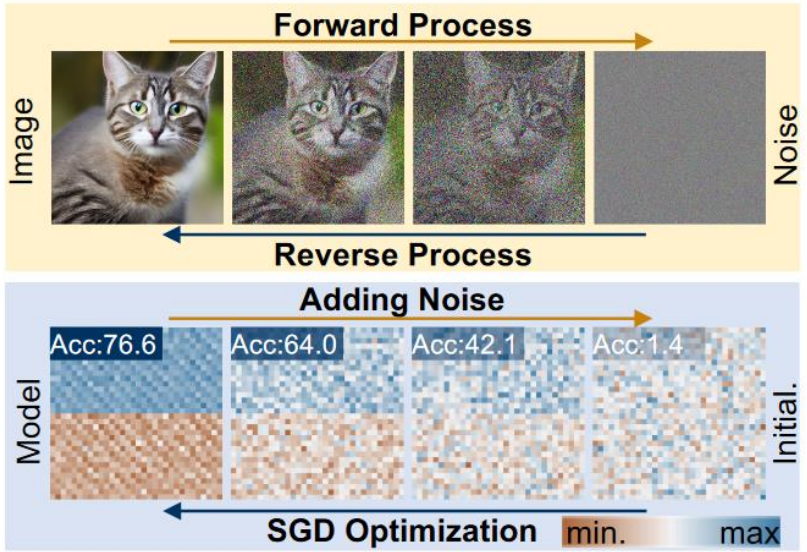
Модель 3D GS также используется для генерации и рендернига **в реальном времени реалистичных 3D аватар** (подвижных моделей тела) людей по их ростовым фотографиям.

Drivable 3D Gaussian Avatars. 2024.

Neural Network Diffusion

Neural Network Diffusion (2024)





The top: illustrates the standard diffusion process in image generation. The bottom: denotes the parameter distribution of batch normalization (BN) during the training CIFAR-100 with ResNet-18. The upper half of the bracket: BN weights. The lower half of the bracket: BN biases.

Parameter autoencoder aims to extract the latent representations and reconstruct model parameters via the decoder. The extracted representations are used to train a standard latent diffusion model (LDM). In the inference, the random noise is fed into LDM and trained decoder to obtain the generated parameters.

Table 1. We present results in the format of 'original / ensemble / p-diff'. Our method obtains similar or even higher performance than baselines. The results of p-diff is average in three runs. **Bold entries** are best results.

Network\Dataset	MNIST	CIFAR-10	CIFAR-100	STL-10	Flowers	Pets	F-101	ImageNet-1K
ResNet-18	99.2 / 99.2 / **99.3**	92.5 / 92.5 / **92.7**	76.7 / 76.7 / **76.9**	75.5 / 75.5 / 75.4	49.1 / 49.1 / **49.7**	60.9 / 60.8 / **61.1**	71.2 / 71.3 / 71.3	78.7 / 78.5 / 78.7
ResNet-50	99.4 / 99.3 / 99.4	91.3 / 91.4 / 91.3	71.6 / 71.6 / **71.7**	69.2 / 69.1 / 69.2	33.7 / 33.9 / **38.1**	58.0 / 58.0 / 58.0	68.6 / 68.5 / 68.6	79.2 / 79.2 / **79.3**
ViT-Tiny	99.5 / 99.5 / 99.5	96.8 / 96.8 / 96.8	86.7 / **86.8** / 86.7	97.3 / 97.3 / 97.3	87.5 / 87.5 / 87.5	89.3 / 89.3 / 89.3	78.5 / 78.4 / 78.5	73.7 / 73.7 / **74.1**
ViT-Base	99.5 / 99.4 / 99.5	98.7 / 98.7 / 98.7	91.5 / 91.4 / **91.7**	99.1 / 99.0 / **99.2**	98.3 / 98.3 / 98.3	91.6 / 91.5 / **91.7**	83.4 / 83.4 / 83.4	84.5 / 84.5 / **84.7**
ConvNeXt-T	99.3 / **99.4** / 99.3	97.6 / 97.6 / 97.7	87.0 / 87.0 / **87.1**	98.2 / 98.0 / 98.2	70.0 / 70.0 / **70.5**	92.9 / 92.8 / **93.0**	76.1 / 76.1 / **76.2**	82.1 / 82.1 / **82.3**
ConvNeXt-B	99.3 / 99.3 / **99.4**	98.1 / 98.1 / 98.1	88.3 / 88.4 / 88.4	98.8 / 98.8 / **98.9**	88.4 / 88.4 / **88.5**	94.1 / 94.0 / 94.1	81.4 / 81.4 / **81.6**	83.8 / 83.7 / **83.9**

We extend our approach to two small architectures: ConvNet-3 on CIFAR-10/100 and MLP-3 on CIFAR-10 and MNIST datasets. These experiments demonstrate the effectiveness and generalization of our approach in synthesizing entire model parameters. However, we can not synthesize the entire parameters of large architectures, such as ResNet, ViT, and ConvNeXt series. It is mainly constrained by the limitation of the GPU memory.

Wang, Kai, et al. Neural Network Diffusion, 2024

На целых маленьких моделях тоже работает

Диффузная модель генерирует обученные веса для CNN или ViT! (*пока не для целой модели*)

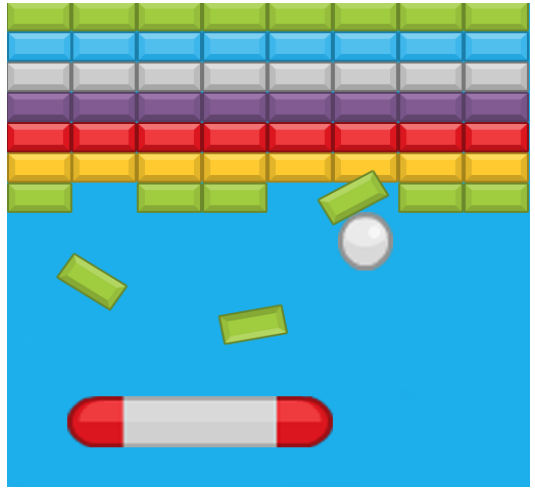
Тенденции и результаты 2020-2024 в области обучения с подкреплением

- Open-Ended Learning
- Трансформеры в RL
- Minecraft RL
- RL встречается с LLM

Deep Reinforcement Learning: обучение с подкреплением

RL до 2020

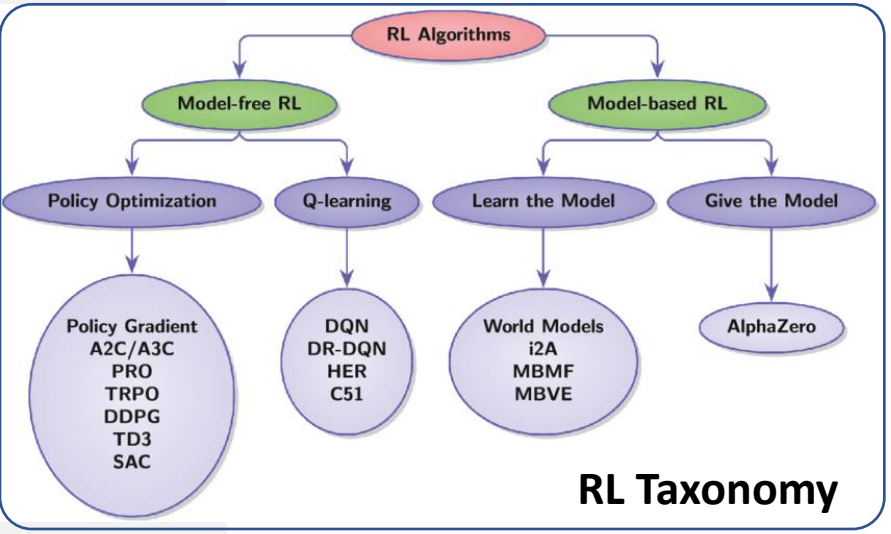
Главная идея: обучение методом проб и ошибок за счет обратной связи от среды



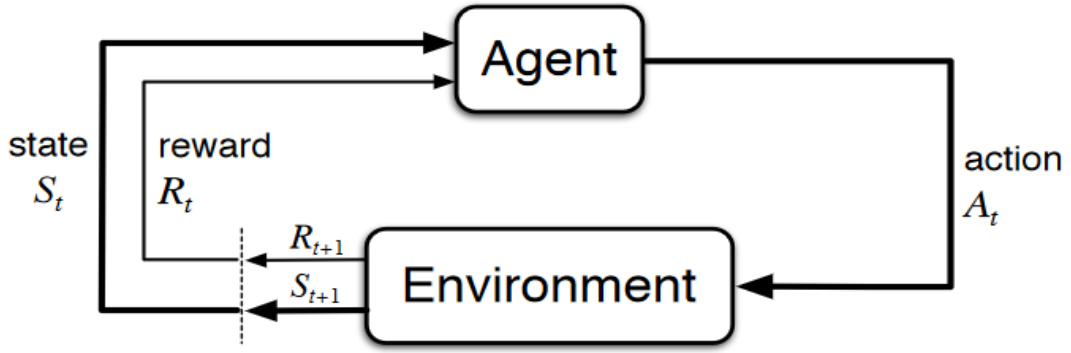
- **Reward $R(t)$:** score you earned at current step
  Вознаграждение (после хода)
- **State S:** current screen
  Состояние (что видим на экране)
- **Action i:** move your board left / right
  Действие (управление/ход)
- **Policy $\pi(s)$:** How to choose your action
  Стратегия (закон управления/хода)
- **Action value function $\hat{Q}(S, i)$:**
  your predicted future total rewards
  Стоимость (выигрыш в будущем)

Задача RL в терминах теории управления:

методом проб и ошибок найти закон управления, оптимизирующий функцию Беллмана



RL Taxonomy



Агент (ОУ) действует в среде, и получает отклик в виде штрафа или вознаграждения

Как научить нейросеть играть в компьютерную игру

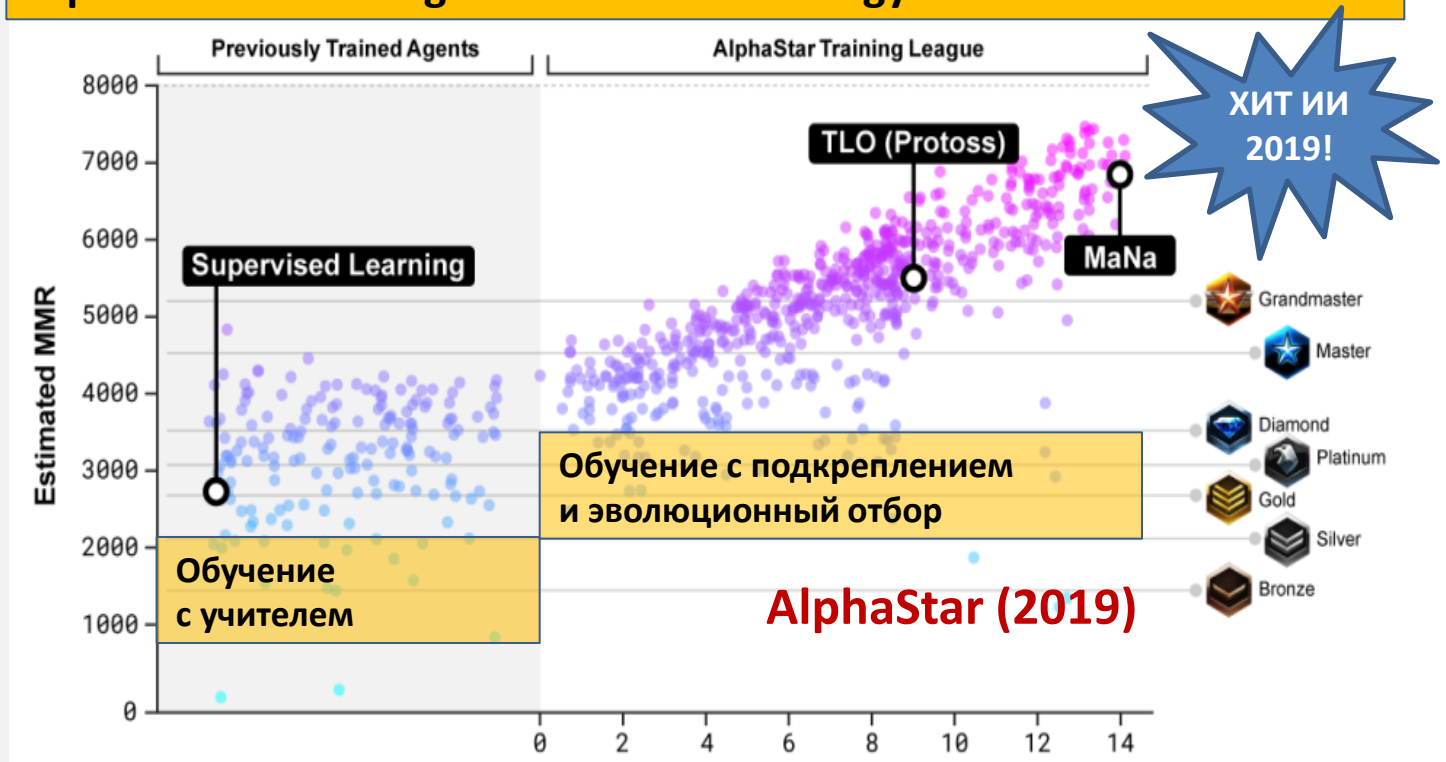
Human-level control through deep reinforcement learning (Minh, et al. Nature, 2015)

AlphaGo, AlphaZero, AlphaStar: что дальше?

RL до 2020

...

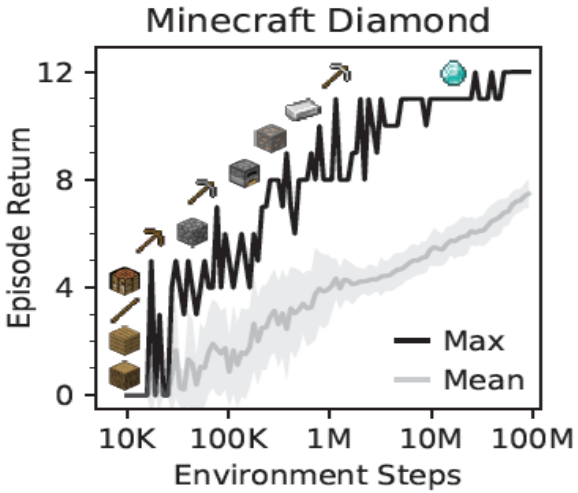
AlphaStar: Mastering the Real-Time Strategy Game StarCraft II



ХИТ ИИ 2019!

https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Item	Reward
Log	1
Planks	2
Stick	4
Crafting table	4
Wooden pickaxe	8
Stone	16
Furnace	32
Stone pickaxe	32
Iron ore	64
Iron ingot	128
Iron pickaxe	256
Diamond	1024




(a) Shelter with Farmland (b) Iron Golem (c) Redstone Circuit (d) Nether Portal

Проблемы и вызовы в RL (2020)

- Нужно большое количество примеров от людей
- Агенты не универсальные (учим под каждую игру)
- Проблема отложенного вознаграждения

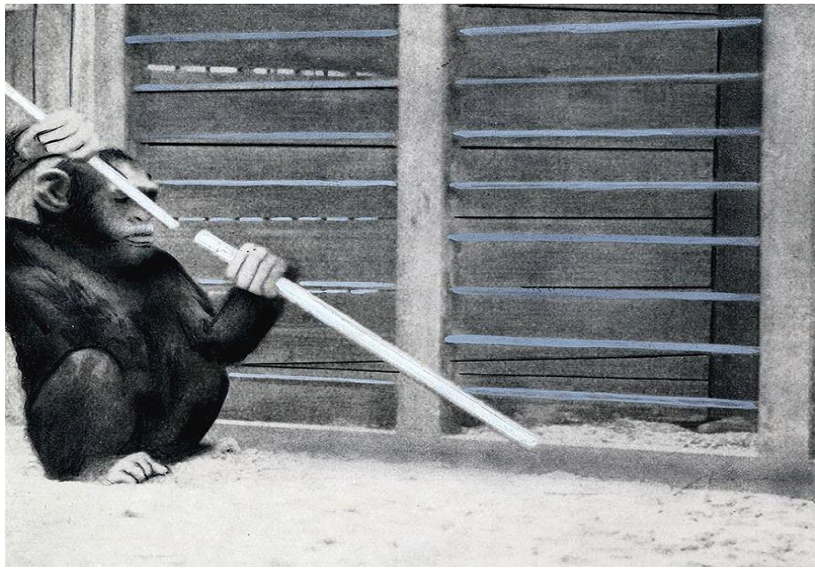
Open-Ended Learning

Нужно большое количество примеров от людей

Агенты не универсальные (учим под каждую игру)

Open-Ended Learning (Открытое обучение)

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?



Открытое обучение учит не задачам, а когнитивному поведению!

Гипотеза: если построить вселенную игровых задач и последовательно обучать ИИ-агентов играть в эти игры, то с каждой новой игрой они будут достигать лучших результатов в этой вселенной и за ее пределами

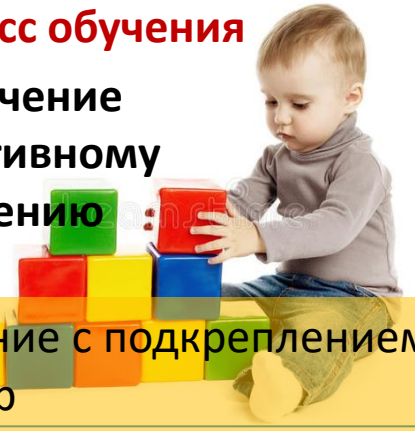
Полученные модели поведения (экспериментирование, использование инструментов, сотрудничество агентов) характерны для когнитивного поведения людей и животных и необходимы для самообучающихся роботов.

Open-Ended Learning Leads to Generally Capable Agents, DeepMind, 2021.

Close-Ended Learning vs. Open-Ended Learning

По сути речь идет о предобученной фундаментальной модели для RL!

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?

	Close-Ended RL	**Open-Ended RL**
Цель обучения:	обучение конкретным навыкам	обучение когнитивному поведению
		
Процесс обучения	**Процесс обучения AlphaStar**  Обучение с учителем Обучение с подкреплением и отбор	**Процесс обучения** **(I) Обучение когнитивному поведению**  Обучение с подкреплением и отбор
Исходные данные и форма их представления	Задача представлена обучающей выборкой. Требуется набрать большой объем примеров действий людей.	Задача представлена текстовым (формальным) описанием (**zero-shot**) и небольшой обучающей выборкой примеров **(few-shot)**. Нужна база знаний предметной области (**ИИ1 + ИИ2!**).

Переход от схемы
1. обучение с учителем
2. обучение с подкреплением
к трехэтапной, где первой фазой будет обучение с подкреплением когнитивному поведению

Обучение с учителем

Обучение с подкреплением и отбор

(II) Изучение задачи

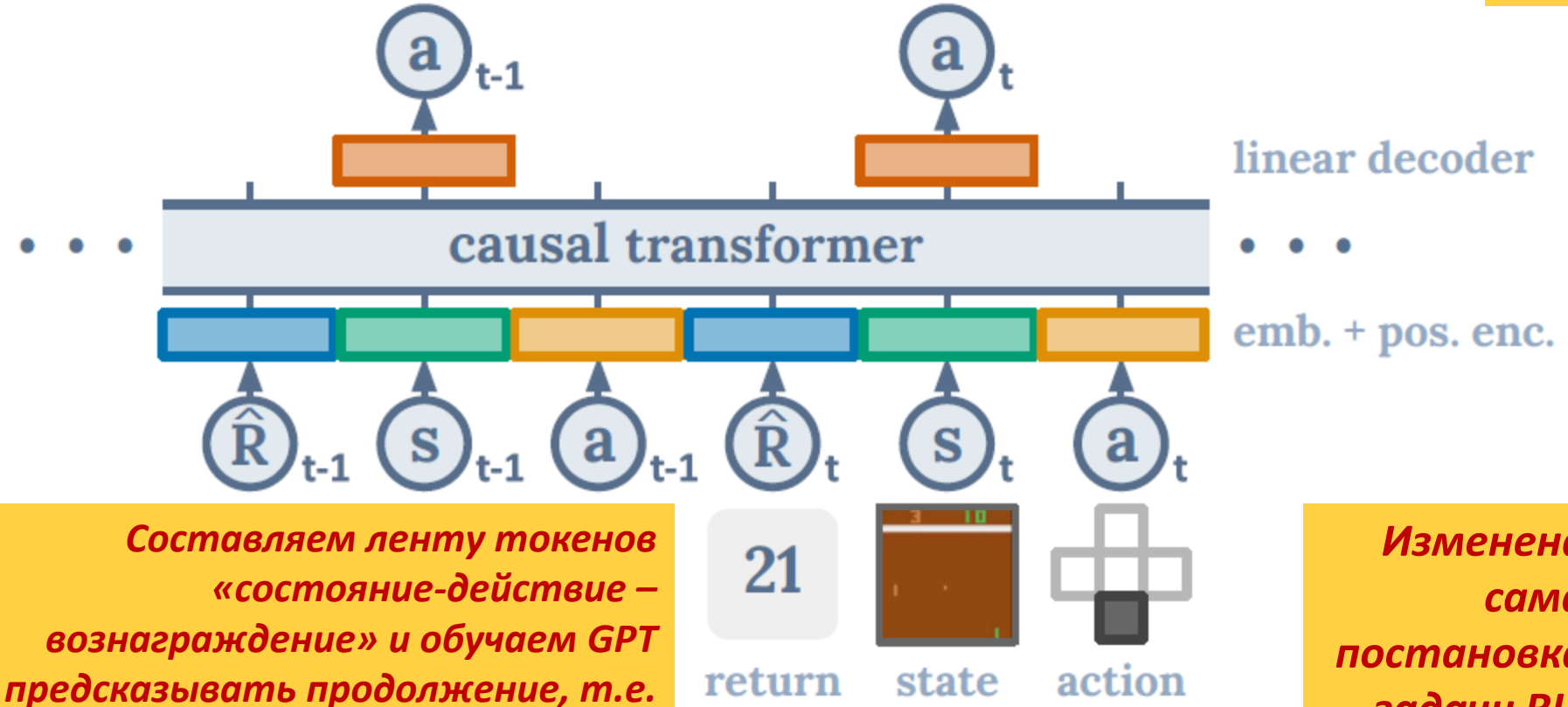
(III) Когнитивный поиск решения

Трансформеры в RL

Агенты не универсальные (учим под каждую игру)

Decision Transformer

2021: Трансформер для обучения с подкреплением



Составляем ленту токенов «состояние-действие – вознаграждение» и обучаем GPT предсказывать продолжение, т.е. «замаскированное» от нас наилучшее будущее решение

Изменена сама постановка задачи RL!

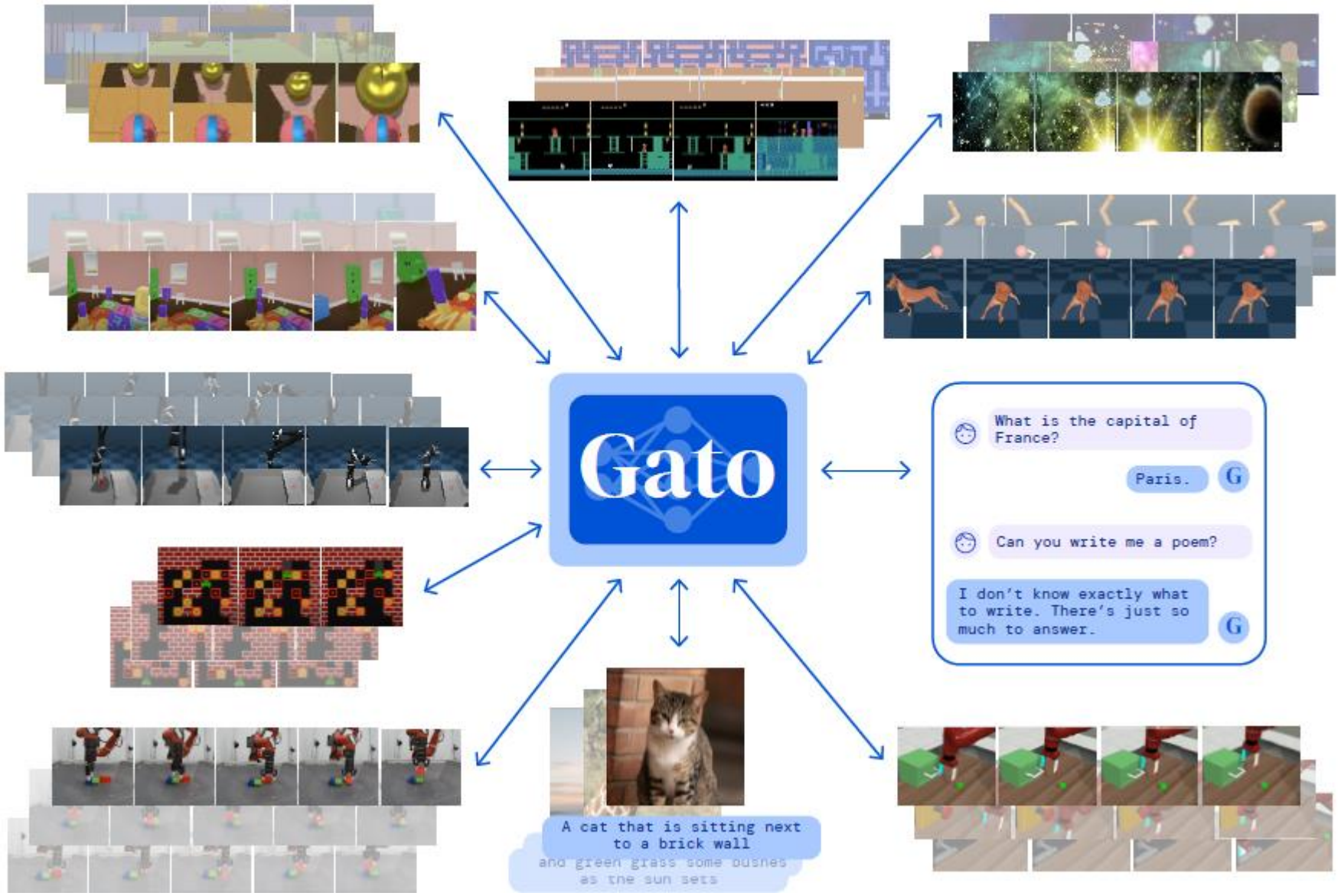
"We study, if **generative trajectory modeling** – i.e. *modeling the joint distribution of the* ***sequence of states, actions, and rewards*** – can serve as a **replacement for conventional RL algorithms**."

Decision Transformer architecture. States, actions, and returns are fed into modality-specific linear embeddings and a positional episodic timestep encoding is added. Tokens are fed into a GPT architecture which predicts actions autoregressively using a causal self-attention mask.

Decision Transformer: Reinforcement Learning via Sequence Modeling, *Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, Igor Mordatch*, 2021.

GATO: универсальный агент

Трансформер для всех задач, включая управление



Inspired by progress in large-scale language modeling, we apply a similar approach towards building a **single generalist agent beyond the realm of text outputs**. The agent, which we refer to as Gato, works as a multi-modal, multi-task, multi-embodiment generalist policy. **The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens.**

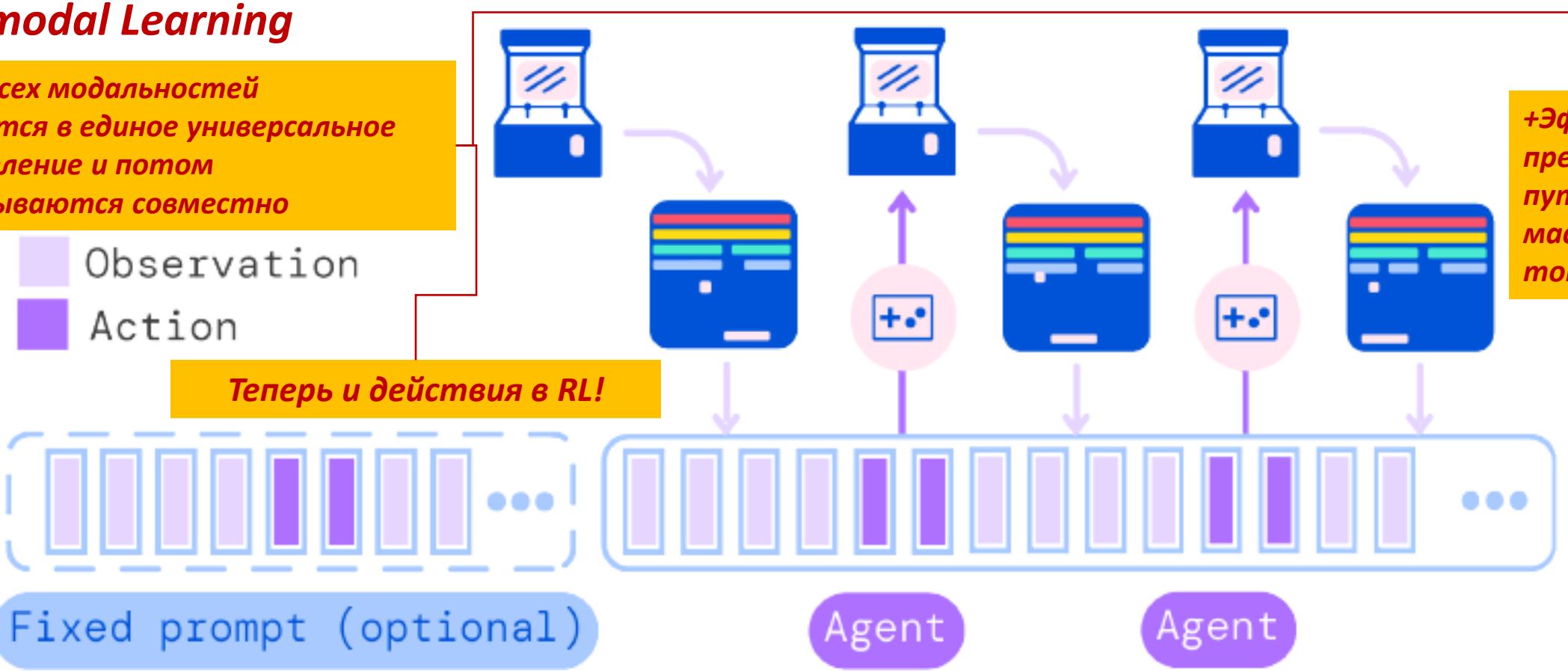
A Generalist Agent, DeepMind, 2022.

GATO: универсальный агент

Трансформер для всех задач, включая управление

Multi-modal Learning

Данные всех модальностей переводятся в единое универсальное представление и потом обрабатываются совместно



+Эффективное предобучение путем маскирования токенов

Running Gato as a control policy. Gato consumes a sequence of interleaved tokenized observations, separator tokens, and previously sampled actions to produce the next action in standard autoregressive manner. The new action is applied to the environment – a game console in this illustration, a new set of observations is obtained, and the process repeats.

Each batch mixes subsequences approximately uniformly over domains (e.g. Atari, MassiveWeb, etc.)!

A Generalist Agent, DeepMind, 2022.

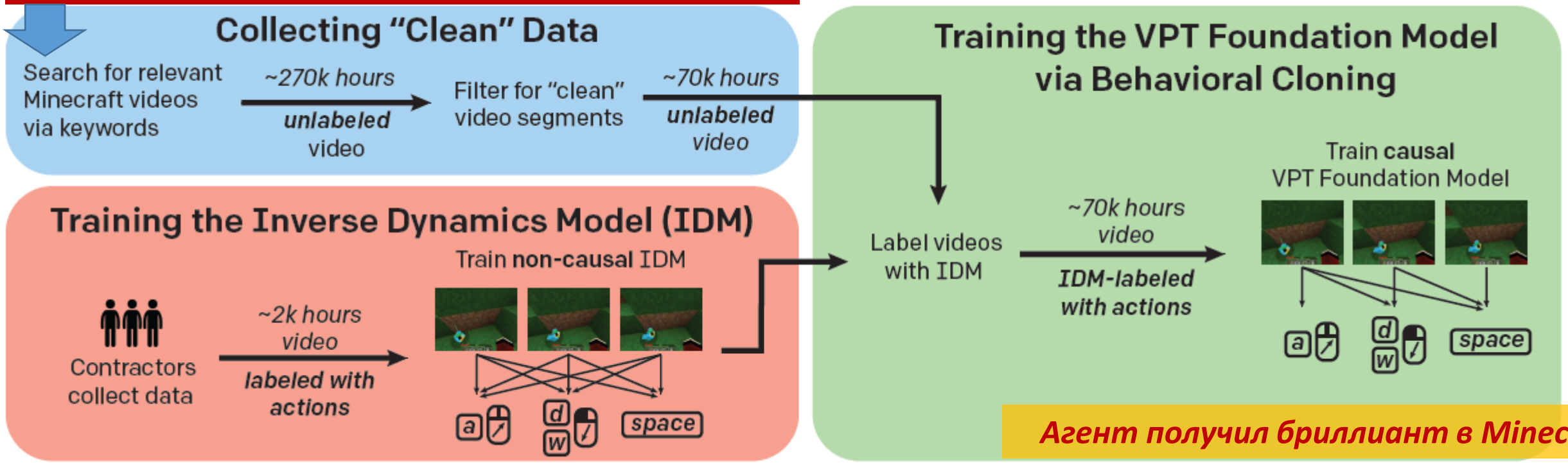
Minecraft RL

Проблема отложенного вознаграждения

Video PreTraining (VPT)

Трансформер, который учится на неразмеченных примерах видео с игрой

Нужно большое количество примеров от людей



Агент получил бриллиант в Minecraft

Для многих областей, требующих выполнения сложных последовательностей действий (робототехника, видеоигры,…), общедоступные данные не содержат меток, необходимых для обучения решению задач. Здесь предложена парадигма **предварительного обучения по данным из Интернета на последовательное принятие решений с помощью имитации обучения, в ходе котрой агенты учатся действовать, просматривая онлайн-видео без маркировки**.

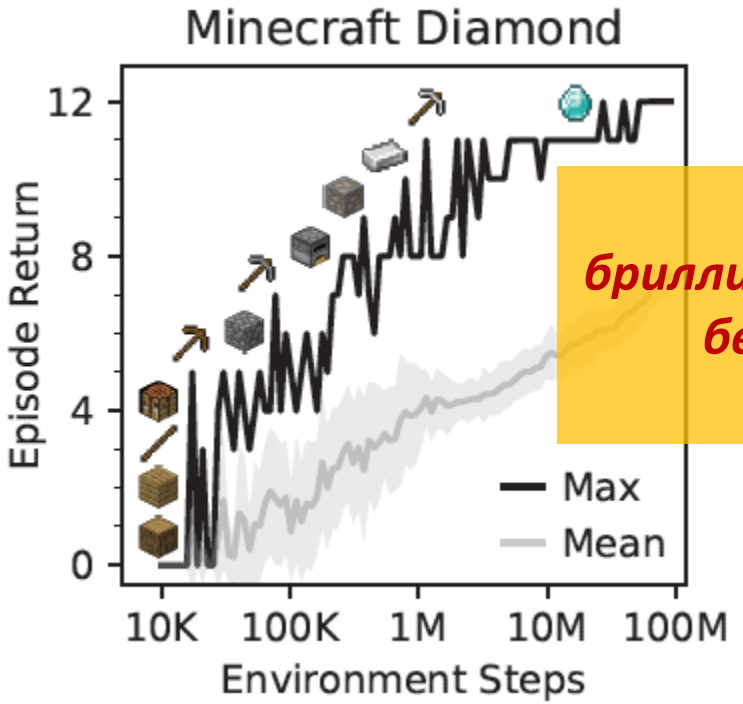
В частности, показано, что **с небольшим количеством размеченных данных** можно обучить модель обратной динамики, достаточно точную, чтобы **разметить огромный неразмеченный источник онлайн-данных** (**онлайн-видео людей, играющих в Minecraft**), **на основе которых затем можно выучить общую модель поведения (игры)**.

Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos, OpenAI, 2022.

DreamerV3: World Model is all you need

Обучение с подкреплением на основе модели мира

Чтобы победить в игре, где вознаграждение отложено на много ходов, нужно уметь предсказывать будущее на много ходов!



Агент получил бриллиант в Minecraft без предобучения на видео!

TOP AI 2022!

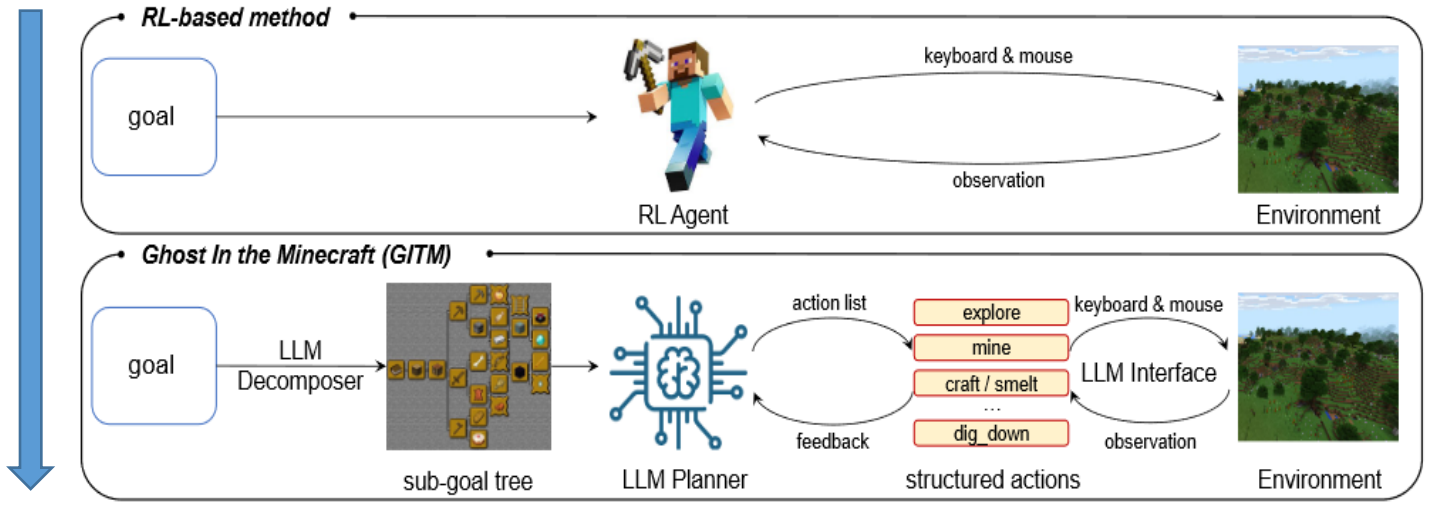
Multi-step predictions on Minecraft. The model receives the first 5 frames as context input and the predicts 45 steps into the future given the action sequence and without access to intermediate images

Mastering Diverse Domains through World Models, DeepMind, 2022.

GITM: Large Language Model is all you need

Программирование на LLM для извлечения и применения знаний

RL как градиентный спуск: учим веса НС методом проб и ошибок



RL без градиентов: LLM, глядя в описание игры (Wiki Minecraft) выполняет автоматическое иерархическое планирование (построение дерева решений), которое в цикле проб и ошибок находит путь к заданной цели в игре

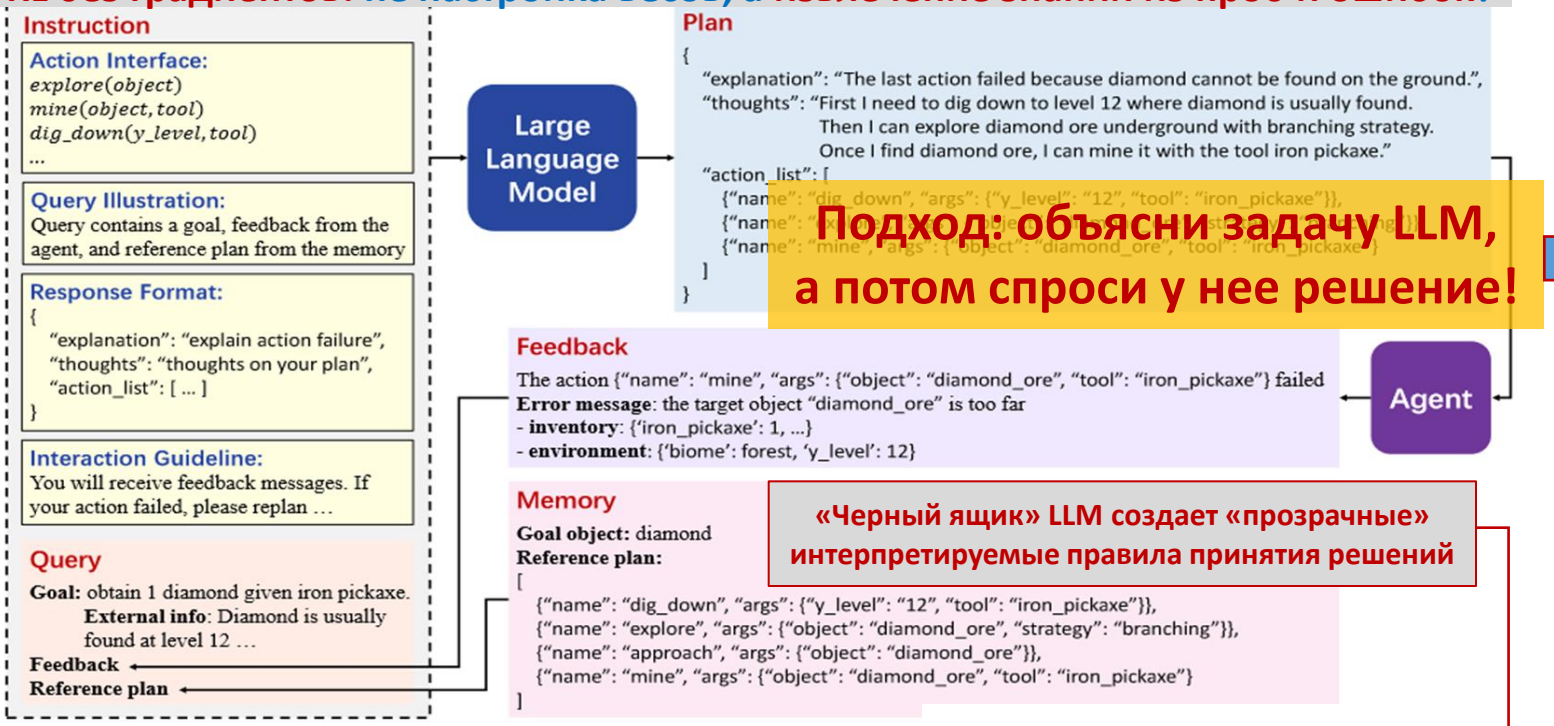
Стратег* разбивает задачу на части, Тактик* планирует действия, Игрок* играет, Наблюдатель* описывает опыт. => И так в цикле до достижения цели
***Все они – запросы к GPT-3.5!**

Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory, 2023.

GITM: Large Language Model is all you need

Программирование на LLM для извлечения и применения знаний

RL без градиентов: не настройка весов, а извлечение знаний из проб и ошибок!

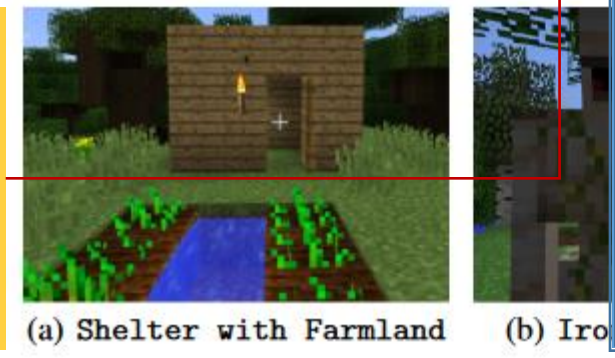


Подход: объясни задачу LLM, а потом спроси у нее решение!

«Черный ящик» LLM создает «прозрачные» интерпретируемые правила принятия решений

Стратег* разбивает задачу на части, Тактик* планирует действия, Агент* играет, Наблюдатель* описывает опыт. => И так в цикле до достижения цели

***Все они – запросы к GPT-3.5!**



(a) Shelter with Farmland (b) Iro

TOP AI 2023!

Пример новой парадигмы программирования в ИИ: Программирование на LLM для извлечения и применения знаний.

Частично напоминает программирование на Прологе. Отличие в том, что база знаний формируется автоматически в процессе обучения большой языковой модели.

А вот составление запросов и использование ответов – дело «программистов на LLM».

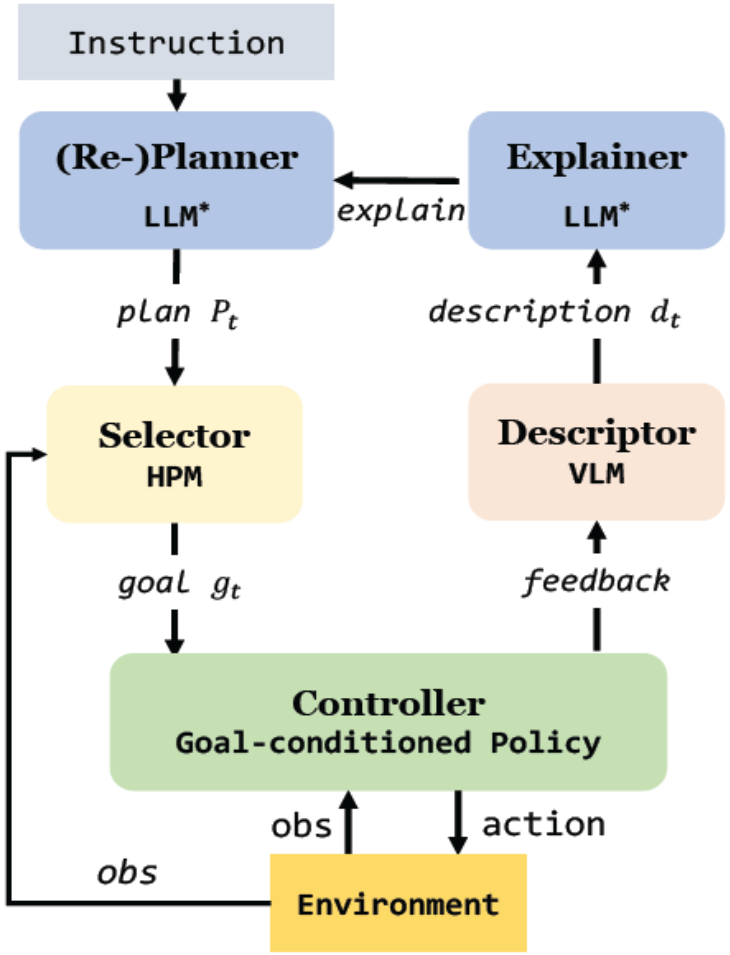
Прогноз: в ближайшем будущем программисты на LLM во многом заменят программистов на Python.

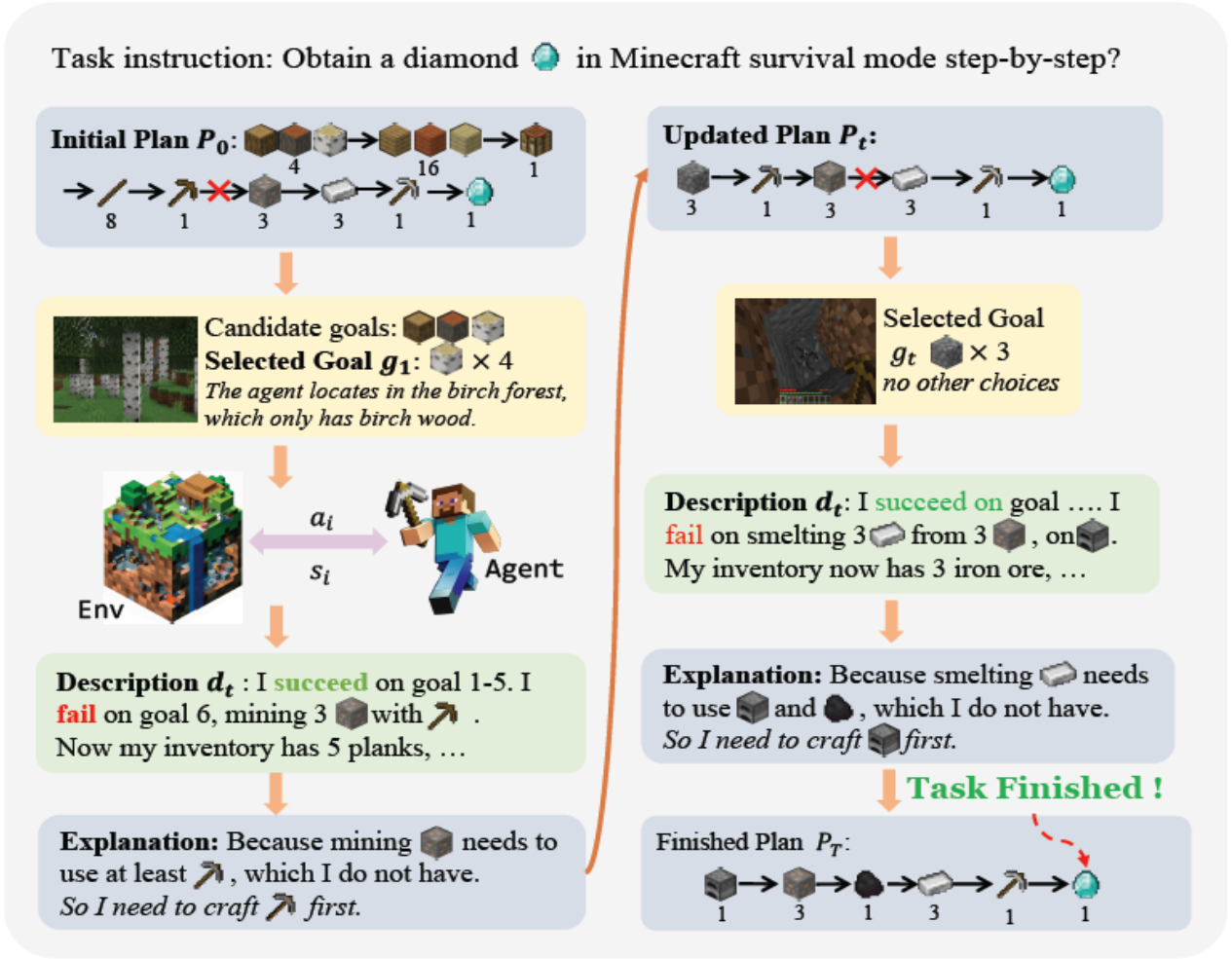
(Инженерия запросов)

Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory, 2023.

Describe, Explain, Plan and Select: Open-World Multi-Task Agents

Программирование на LLM для извлечения и применения знаний





Mine oak wood | Mine birch wood | Craft acacia planks | Craft crafting table | Craft stick | Mine iron ore | Mine coal | Craft furnace | Mine diamond

Mine acacia wood | Craft oak planks | Craft birch planks | Craft wood pickaxe | Craft stone pickaxe | Mine cobblestone | Smelt iron ingot | Craft iron pickaxe

DEPS facilitates better error correction on initial **LLM-generated plan** by integrating **description of the plan execution process** and providing **self-explanation of feedback** when encountering failures during the extended planning phases. Furthermore, it includes a **goal selector**, which is a trainable module that ranks parallel candidate sub-goals based on the estimated steps of completion, consequently refining the initial plan.

1st zero-shot multi-task agent: 70+ Minecraft tasks, x2 overall performances

Describe, Explain, Plan and Select: Interactive Planning with LLM Enables Open-World Multi-Task Agents. 2023.

RL & LLM

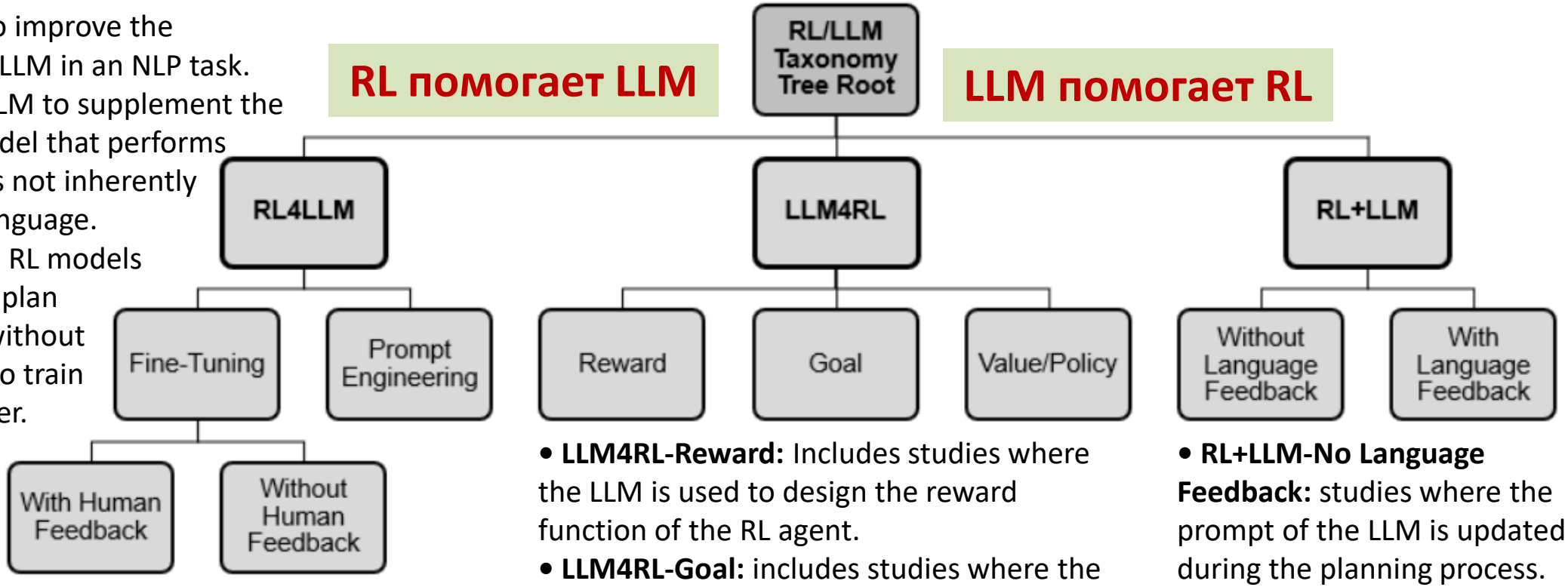
(*что они могут вместе*)

RL/LLM Taxonomy Tree

1. RL4LLM: use RL to improve the performance of the LLM in an NLP task.
2. LLM4RL: use an LLM to supplement the training of an RL model that performs a general task that is not inherently related to natural language.
3. RL+LLM: combine RL models with LLM models to plan over a set of skills, without using either model to train or fine-tune the other.

RL помогает LLM

LLM помогает RL



- **RL4LLM-Fine-tuning:** Encompasses studies where RL is used to perform model fine-tuning, which involves tweaking the model parameters, until the model achieves the desired performance. This subclass can be further refined according to the presence or absence of human feedback.
- **RL4LLM-Prompt Engineering:** Includes studies where RL is used to iteratively update the prompt of the LLM, until the model achieves the desired performance.

- **LLM4RL-Reward:** Includes studies where the LLM is used to design the reward function of the RL agent.
- **LLM4RL-Goal:** includes studies where the LLM is utilized for goal setting, which applies to goal-conditioned RL settings.
- **LLM4RL-Policy:** includes studies where the LLM represents the policy function to be learned, or directly assists its training or pretraining.

- **RL+LLM-No Language Feedback:** studies where the prompt of the LLM is updated during the planning process.
- **RL+LLM-With Language Feedback:** studies where the prompt of the LLM stays fixed throughout the planning process.

RL и LLM вместе достигают нового качества

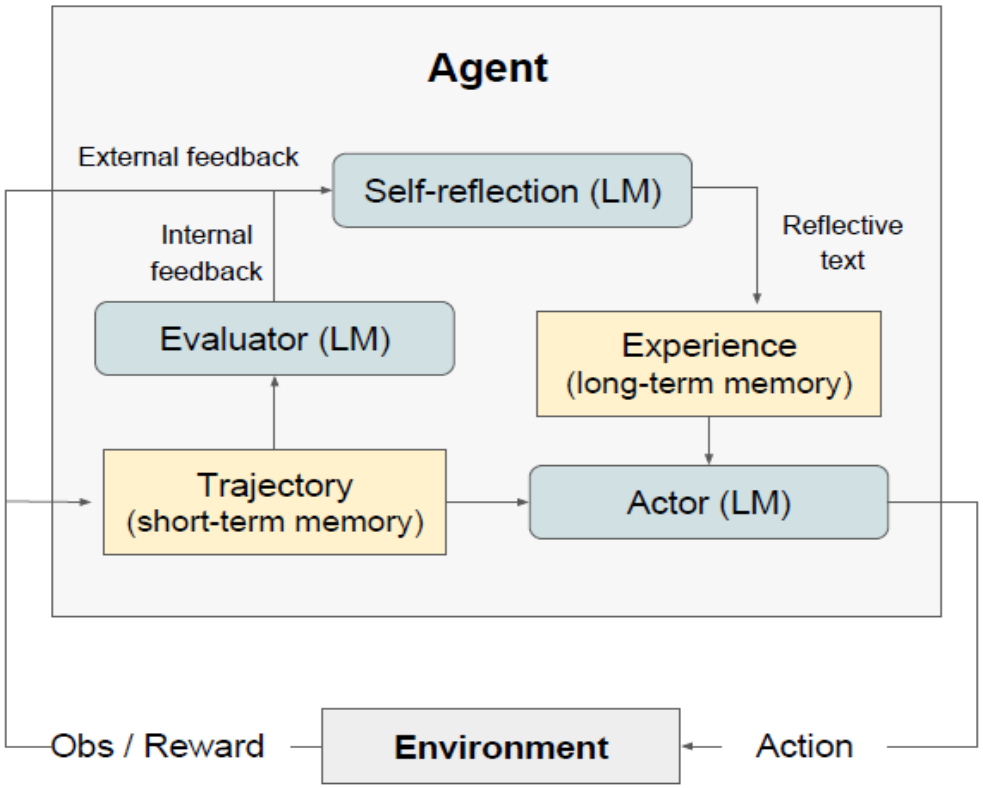
THE RL/LLM TAXONOMY TREE: REVIEWING SYNERGIES BETWEEN REINFORCEMENT LEARNING AND LARGE LANGUAGE MODELS, 2024

LLM для RL

(*вербальный, контекстный и символьный RL*)

Reflexion: Verbal Reinforcement Learning

LLM4RL



Algorithm 1 Reinforcement via self-reflection

Initialize Actor, Evaluator, Self-Reflection: M_a, M_e, M_{sr}
Initialize policy $\pi_\theta(a_i|s_i), \theta = \{M_a, mem\}$
Generate initial trajectory using π_θ
Evaluate τ_0 using M_e
Generate initial self-reflection sr_0 using M_{sr}
Set $mem \leftarrow [sr_0]$
Set $t = 0$
while M_e not pass or $t <$ max trials **do**
  Generate $\tau_t = [a_0, o_0, \ldots a_i, o_i]$ using π_θ
  Evaluate τ_t using M_e
  Generate self-reflection sr_t using M_{sr}
  Append sr_t to mem
  Increment t
end while
return

This is akin to **how humans iteratively learn to accomplish complex tasks in a few-shot manner** – by **reflecting on their previous failures** in order to form **an improved plan of attack for the next attempt**.

Попробовать, определить, в чем ошибка, запомнить, найти новый план с учетом опыта

Self-reflection (LLM) generates verbal self-reflections to provide valuable feedback for future trials. Given a **sparse reward signal**, such as a binary success status (success/fail), the **current trajectory**, and its **persistent memory mem**, the **self-reflection model generates nuanced and specific feedback**, which is more informative than scalar rewards, is then **stored in the agent's memory** (mem).

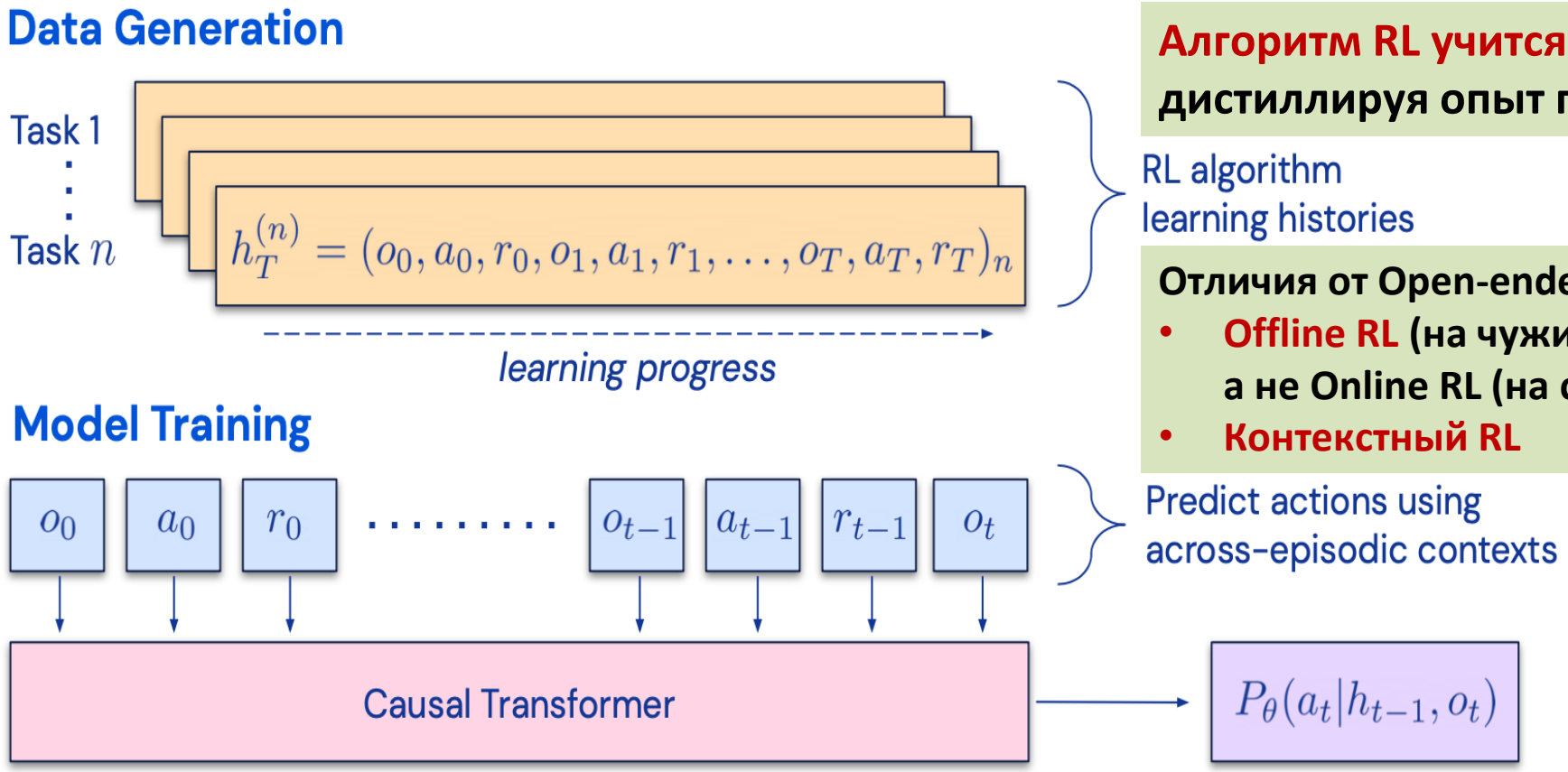
RL для NN: оперантное научение (как для бессловесных животных)
RL для LLM: скажи, в чем ошибка! (как человеку)

Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al. 2023.

In-context Reinforcement Learning with Algorithm Distillation

LLM4RL

Здесь также речь идет о создании предобученной фундаментальной модели для RL!



Алгоритм RL учится учиться, дистиллируя опыт прошлых RL

RL algorithm learning histories

Отличия от Open-ended-learning

- **Offline RL (на чужих ошибках), а не Online RL (на своих)**
- **Контекстный RL**

Predict actions using across-episodic contexts

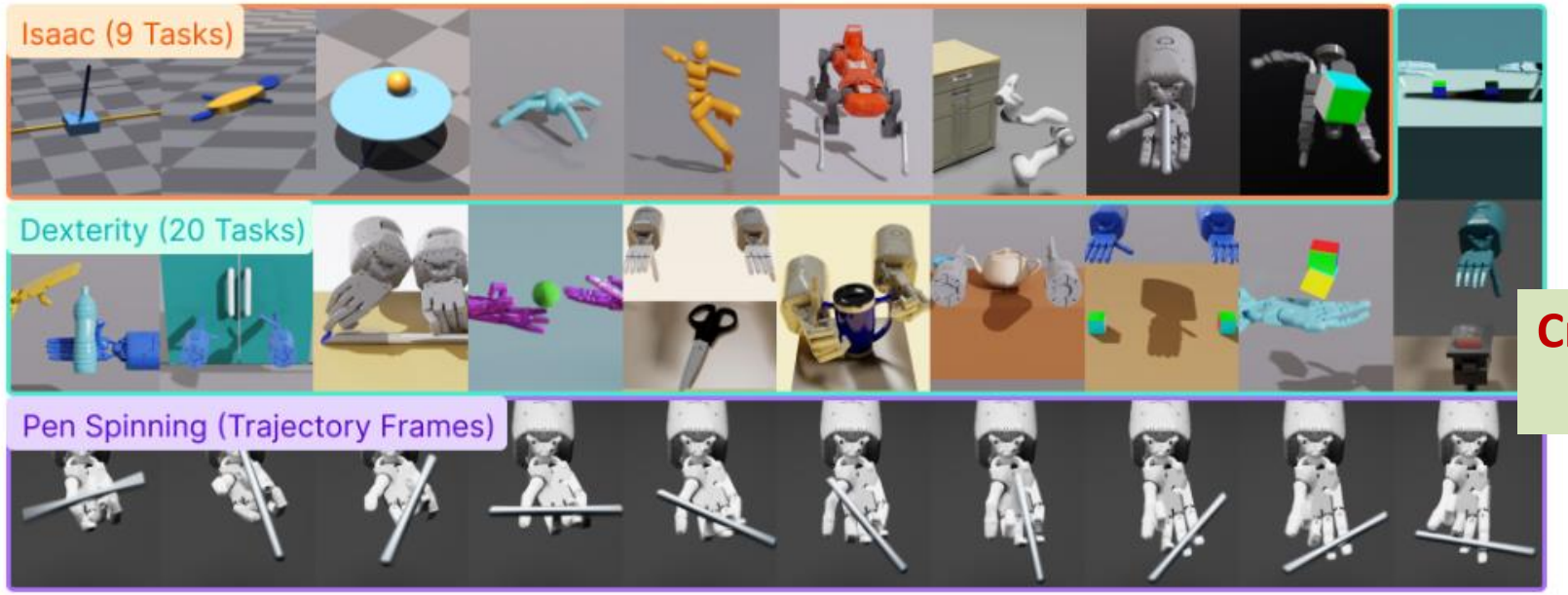
AD is the first method to demonstrate **in-context reinforcement learning via sequential modeling of offline data with an imitation loss**

Algorithm Distillation (AD) has two steps – (i) a **dataset of learning histories is collected from individual single-task RL** algorithms solving different tasks; (ii) a **causal transformer predicts actions from these histories using across-episodic contexts**. Since the RL policy improves throughout the learning histories, by predicting actions accurately AD learns to output an improved policy relative to the one seen in its context. AD models state-action-reward tokens.

Laskin et al. "In-context Reinforcement Learning with Algorithm Distillation" (2023)

Eureka: Human-Level Reward Design via Coding Large Language Models

LLM4RL



Символьный RL

Reward Term	Formulation
Dexterity	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation	$2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))$
AllegroHand	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation difference	$1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Maximize orientation difference	$-1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Ant	
Torso height	$-\|h - h_t\|$
Torso velocity	$-\|\|v_{xy}\|_2 - v_t\|$
Angle to target	$-\|\theta - \theta_t\|$
Anymal	
Minimize difference	$\exp -(x - x_t)^2$
BallBalance	
Ball position	$1/(1 + \|p - p_t\|_2)$
Ball velocity	$1/(1 + \|v - v_t\|_2)$
Cartpole	
Pole angle	$-(\theta - \theta_t)^2$
Pole velocity	$-\|v - v_t\|$
Cart velocity	$-\|v - v_t\|$
FrankaCabinet	
Minimize hand distance	$-\|p_1 - p_2\|_2$
Maximize hand distance	$\|p_1 - p_2\|_2$
Drawer extension	$-\|p - p_t\|$
Humanoid	
Torso height	$-\|h - h_t\|$
Torso velocity	$-\|\|v_{xy}\|_2 - v_t\|$
Angle to target	$-\|\theta - \theta_t\|$
Quadcopter	
Quadcopter position	$1/(1 + \|p - p_t\|_2^2)$
Upright alignment	$1/(1 + \|1 - n_z\|^2)$
Positional velocity	$1/(1 + \|v - v_t\|_2^2)$
Angular velocity	$1/(1 + \|\omega - \omega_t\|_2^2)$
ShadowHand	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation difference	$1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Maximize orientation difference	$-1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$

Evolution-driven **U**niversal **RE**ward **K**it for **A**gent (EUREKA)

Algorithm 1 EUREKA

Require: Task description l, environment code M, coding LLM LLM, fitness function F, initial prompt prompt
Hyperparameters: search iteration N, iteration batch size K
for N iterations **do**
  // Sample K reward code from LLM
  $R_1, ..., R_k \sim \text{LLM}(l, M, \text{prompt})$
  // Evaluate reward candidates
  $s_1 = F(R_1), ..., s_K = F(R_K)$
  // Reward reflection
  prompt := prompt : Reflection(R^n_{best}, s^n_{best}), where $best = \arg\max_k s_1, ..., s_K$
  // Update Eureka reward
  $R_{\text{Eureka}}, s_{\text{Eureka}} = (R^n_{\text{best}}, s^n_{\text{best}})$, if $s^n_{\text{best}} > s_{\text{Eureka}}$
Output: R_{Eureka}

1. **Human-level performance on reward design**.
2. **Solves dexterous manipulation tasks that were previously not feasible by manual reward engineering**.
3. **Enables a new gradient-free in-context learning approach to reinforcement learning from human feedback (RLHF)** that can generate human-aligned reward functions based on various forms of human inputs.

Автоматический подбор reward для RL с использованием LLM рефлексии!

Eureka: Human-Level Reward Design via Coding Large Language Models, 2023

RL & LLM

за пределы обучающей выборки!

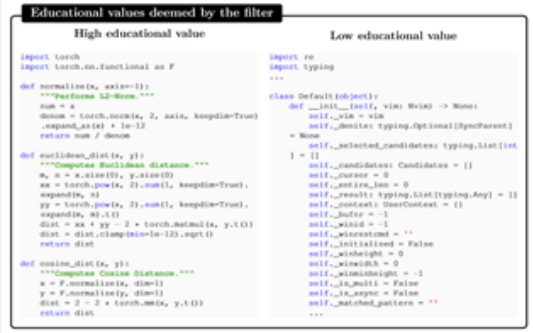
Surpassing the training data: getting more from LLMs at inference time

RL4LLM

Пределы обучения LLM:

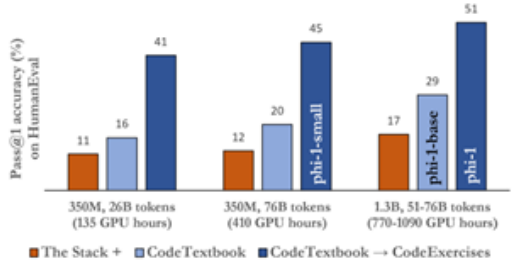
- Предобученная LLM пытается продолжать текст как **в среднем по интернету**
- При помощи фильтрации данных (**textbooks**) и **RLHF** мы пытаемся сдвинуть выученное LLM распределение **к лучшему из интернета**
- Таким путем в лучшем случае **LLM может достичь лучшего уровня человеческих ответов**
- **Можно ли превысить этот уровень?**

Phi-1: Textbooks Are All You Need (обучаться только на хороших примерах)



Our training relies on three main datasets:
- *A filtered code-language dataset*, which is a subset of The Stack and StackOverflow, obtained by using a language model-based classifier (consisting of about 6B tokens).
- A *synthetic textbook dataset* consisting of <1B tokens of GPT-3.5 generated Python textbooks.
- A *small synthetic exercises dataset* consisting of ~180M tokens of Python exercises and solutions.

The standard code datasets form a large and diverse corpus covering broad range of topics and use cases. One can only imagine how frustrating and inefficient it would be for a human learner to try to acquire coding skills from these datasets, as they would have to deal with a lot of noise, ambiguity, and incompleteness in the data. We hypothesize that these issues also affect the performance of language models, as they reduce the quality and quantity of the signal that maps natural language to code. We conjecture that language models would benefit from a training set that has the same qualities as a good "textbook": it should be clear, self-contained, instructive, and balanced.

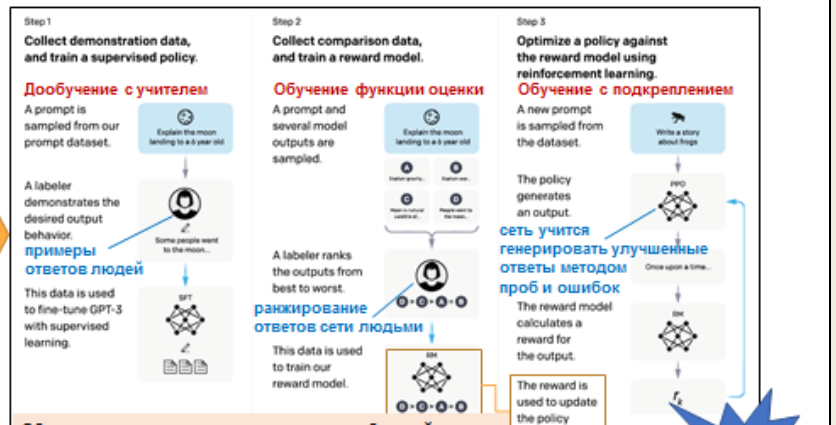


Textbooks Are All You Need, Microsoft Research, 2023

ChatGPT (InstructGPT): как добиться от GPT полезных и релевантных ответов при помощи обучения с подкреплением

We want language models to be **helpful** (they should help the user solve their task), **honest** (they shouldn't fabricate information or mislead the user), **harmless** (they should not cause physical, psychological, or social harm to people or the environment).

Не всегда понимает, чего от нее хотят

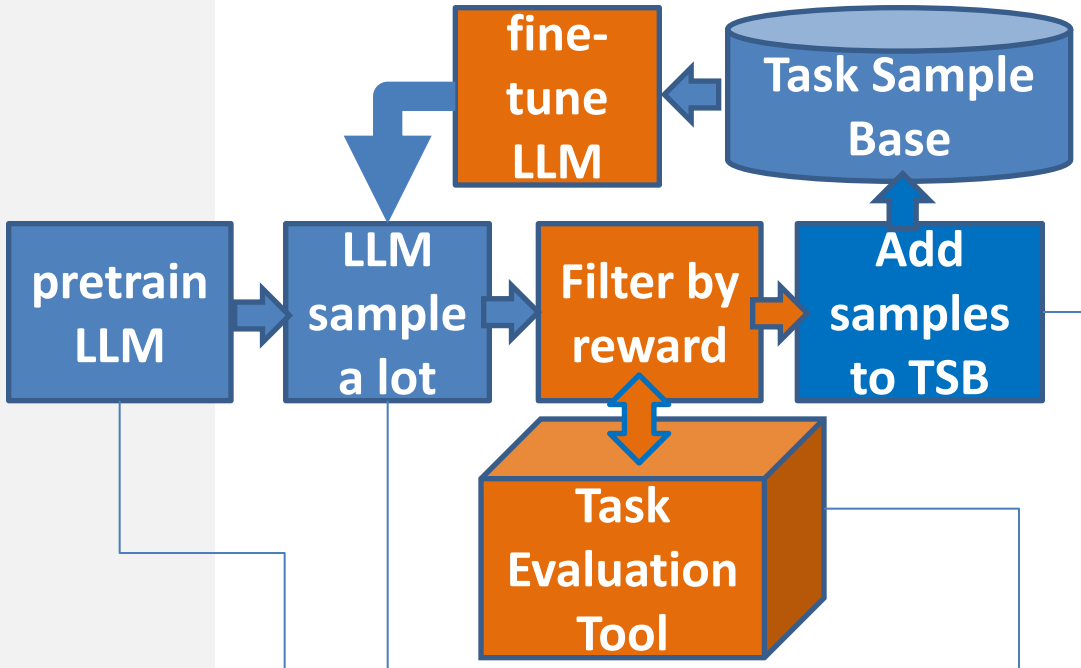


Training language models to follow instructions with human feedback, OpenAI, 2022

Alexander Novikov, Surpassing the training data: getting more from LLMs at inference time, 2024

Surpassing the training data: getting more from LLMs at inference time

RL4LLM



В цикле: {генерируем, фильтруем, дообучаем}

Отличия от классического RL:

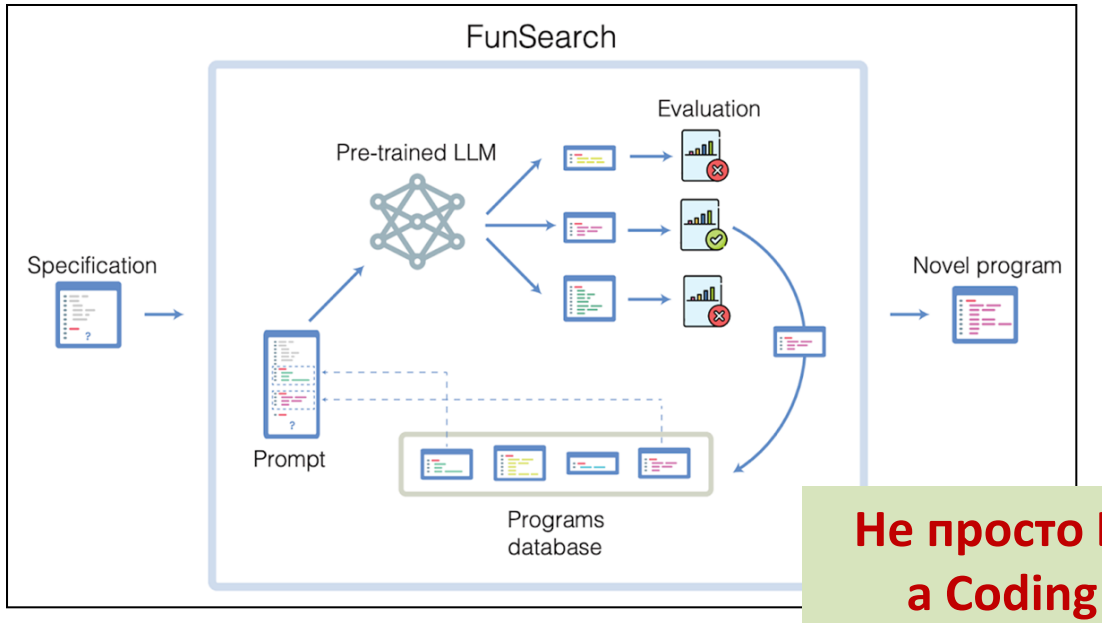
- **Стартуем уже с «хорошей» предобученной LLM**
- Учитывая, что модель большая, **нужен разумный компромисс между временем и ресурсами, затраченными на генерацию, оценку и дообучение**.

За счет дообучения распределение решений LLM постепенно смещается в сторону желаемой области

Alexander Novikov, Surpassing the training data: getting more from LLMs at inference time, 2024

Surpassing the training data: getting more from LLMs at inference time

RL4LLM

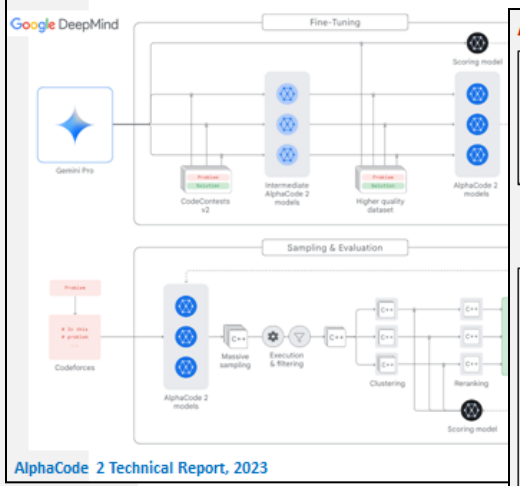


Не просто LLM, а Coding LLM

RL ищет алгоритм, а не действие:

- **Малое изменение программы ведет к большому изменению поведения**
- Могут возникать и в дальнейшем многократно использоваться **библиотеки алгоритмов** («ноу-хау», рецептов, функций)

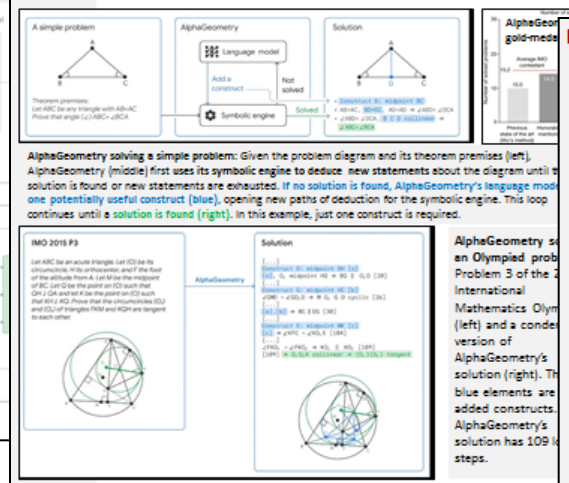
AlphaCode 2: 1M попыток + фильтр = расширение пространства решений



AlphaCode 2 Technical Report, 2023

AlphaGeometry: Neuro-symbolic Olympiad-level AI system for geometry

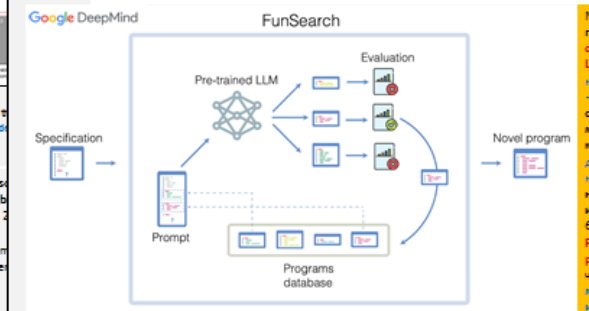
Автоматическое программирование для математических задач



Solving olympiad geometry without human demonstrations, 2024

FunSearch: Mathematical discoveries from program search with LLMs

Автоматическое программирование для математических задач



Мы представляем FunSearch (сокращение от «поиск в пространстве функций»), эволюционную процедуру, основанную на объединении предварительно обученного LLM с систематическим оценщиком. Применяя FunSearch к центральной проблеме экстремальной комбинаторики — проблеме предельного множества — мы обнаруживаем новые конструкции больших предельных множеств, выходящие за рамки самых известных, как в конечномерных, так и в асимптотических случаях. Мы демонстрируем универсальность FunSearch, применяя его к алгоритмической задаче, онлайн-упаковке корзин, и находим новые эвристики, которые улучшают широко используемые базовые показатели. В отличие от большинства подходов к компьютерному поиску, FunSearch ищет программы, которые описывают, как решить проблему, а не каково ее решение. Помимо того, что найденные программы являются эффективной и масштабируемой стратегией, они, как правило, более интерпретируемы, что обеспечивает обратную связь между экспертами в предметной области

The FunSearch process:

- The LLM is shown a selection of the best programs it has generated so far (retrieved from the programs database), and asked to generate an even better one.
- The programs proposed by the LLM are automatically executed, and evaluated.
- The best programs are added to the database, for selection in subsequent cycles.
- The user can at any point retrieve the highest-scoring programs discovered so far.

Mathematical discoveries from program search with large language models, 2024

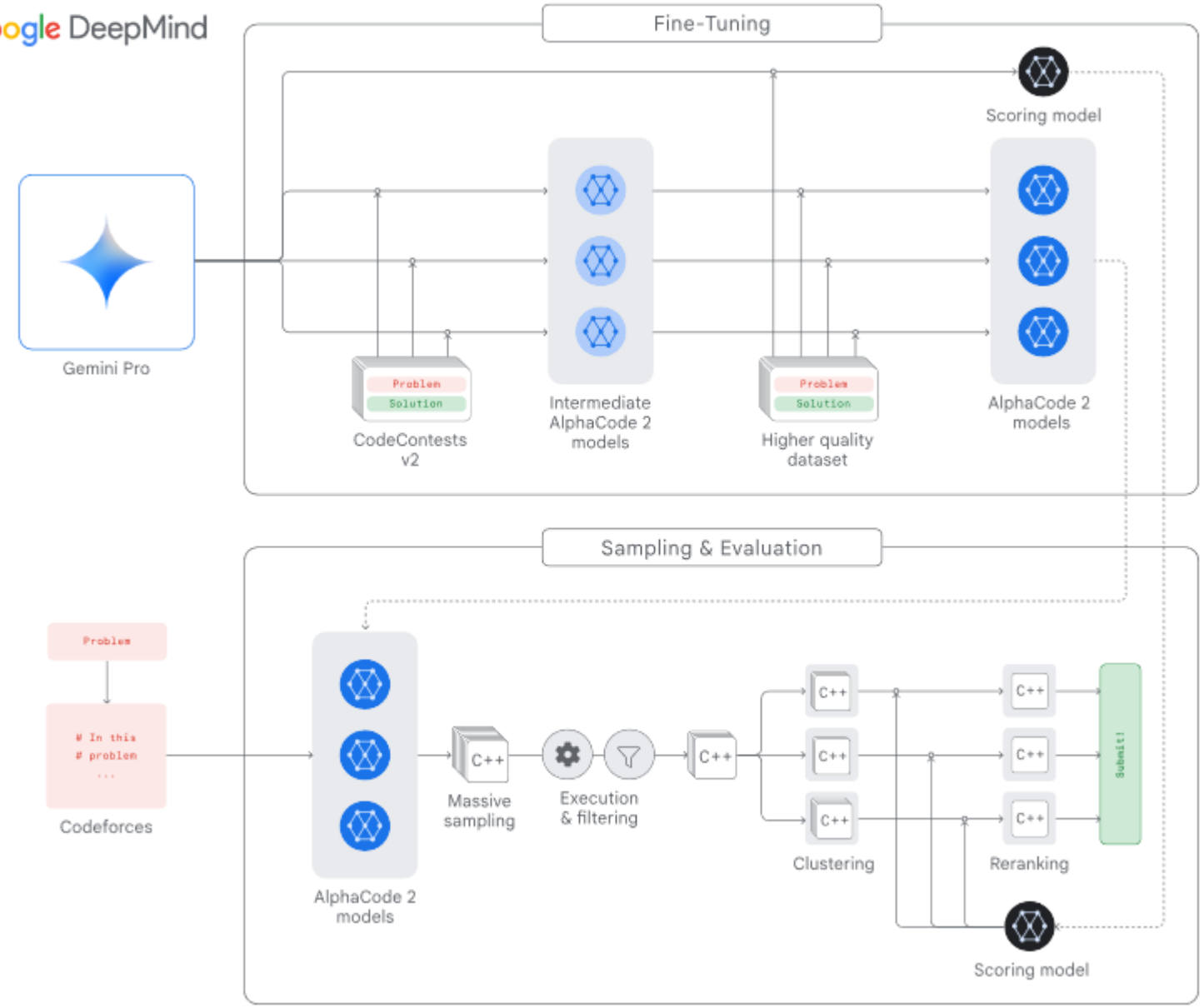
Алгоритм решения любой задачи можно описать как программу (инструкцию) на специальном или естественном языке: *Поиск решения как программы оказывается универсальным научным навыком!*

За счет дообучения на новой синтетике распределение решений LLM постепенно смещается в сторону желаемой области

Alexander Novikov, Surpassing the training data: getting more from LLMs at inference time, 2024

AlphaCode 2: Generate and filter 1M code solutions for each problem

Google DeepMind

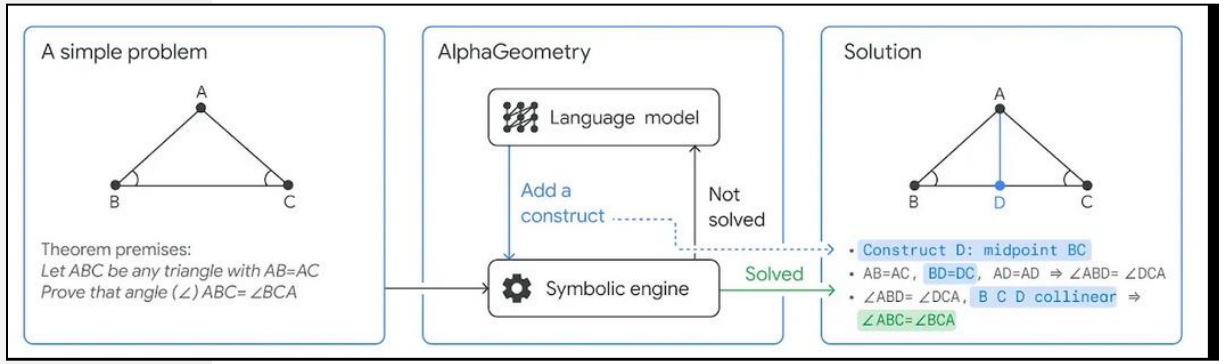


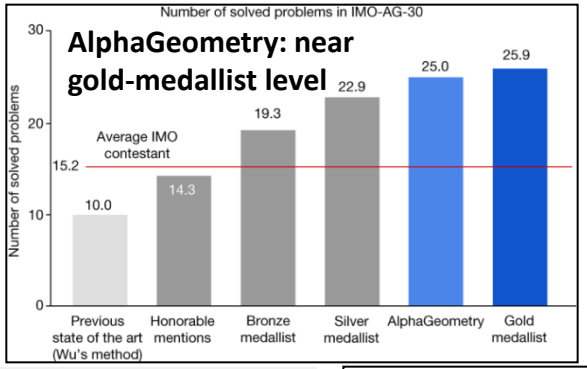
- A family of policy models which **generate code samples for each problem**;
- A sampling mechanism that encourages **generating a wide diversity of code samples** to search over the space of possible programs;
- A **filtering mechanism to remove code samples** that do not comply with the problem description;
- A **clustering algorithm that groups semantically similar code samples**, allowing us to avoid redundancies;
- A **scoring model** which we use to surface the **best candidate out of each of the 10 biggest code samples clusters**.

We generate up to a million code samples per problem, using a **randomized temperature** parameter to encourage diversity.
We execute each code sample and filter out all which do not produce the expected output.

AlphaCode 2 Technical Report, 2023

AlphaGeometry: Neuro-symbolic Olympiad-level AI system for geometry

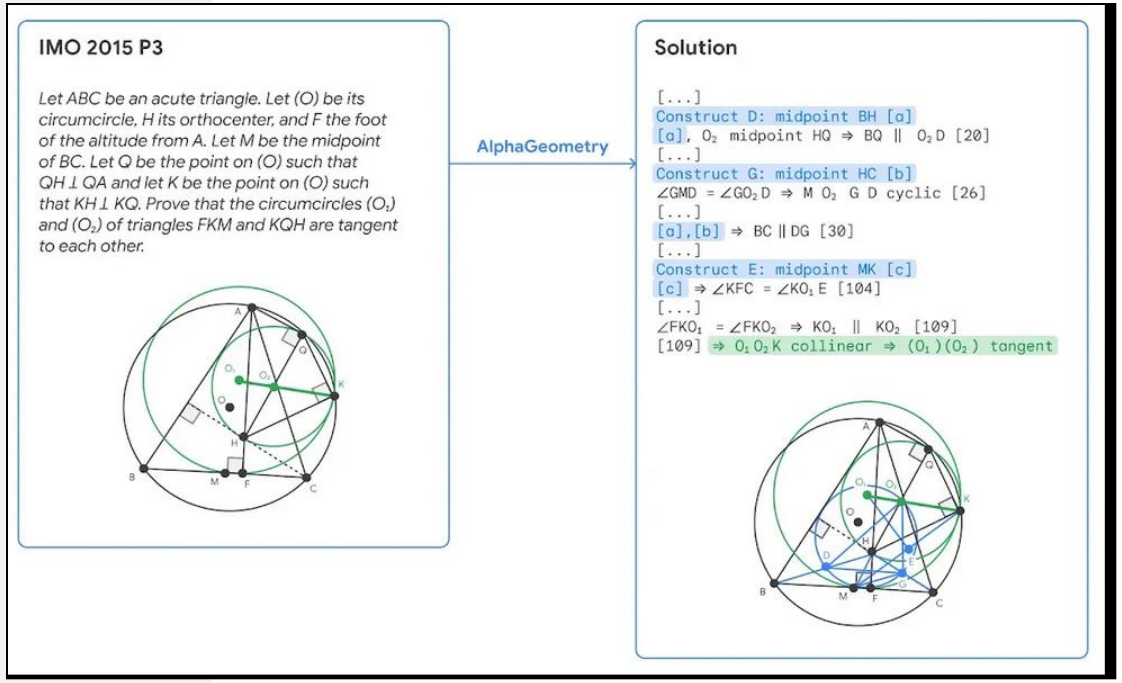




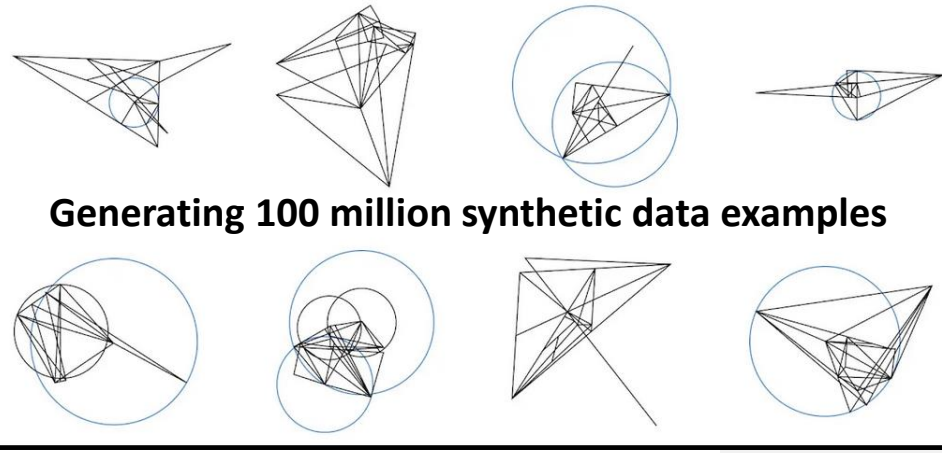
AlphaGeometry method to **generate and train language models on purely synthetic data** provides a **general guiding framework for mathematical domains** with the same data-scarcity problem.

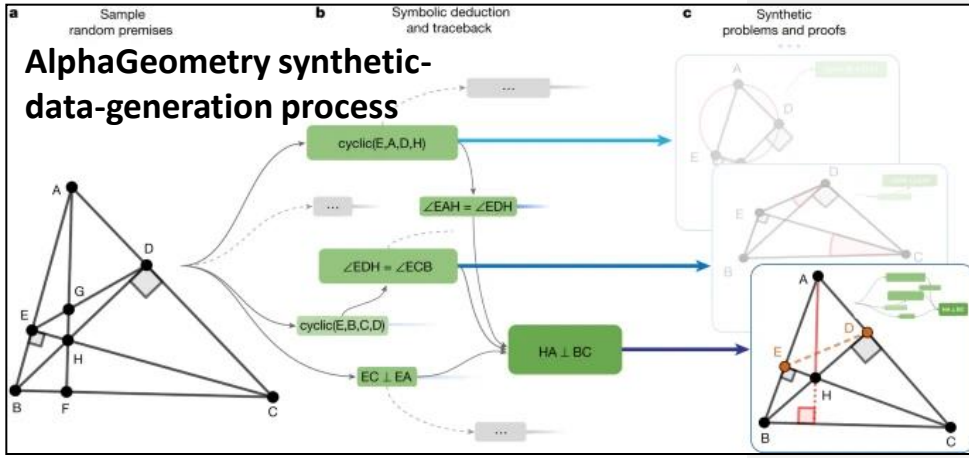
Google DeepMind

AlphaGeometry solving a simple problem: Given the problem diagram and its theorem premises (left), AlphaGeometry (middle) first **uses its symbolic engine to deduce new statements** about the diagram until the solution is found or new statements are exhausted. **If no solution is found, AlphaGeometry's language model adds one potentially useful construct (blue),** opening new paths of deduction for the symbolic engine. This loop continues until a **solution is found (right)**. In this example, just one construct is required.



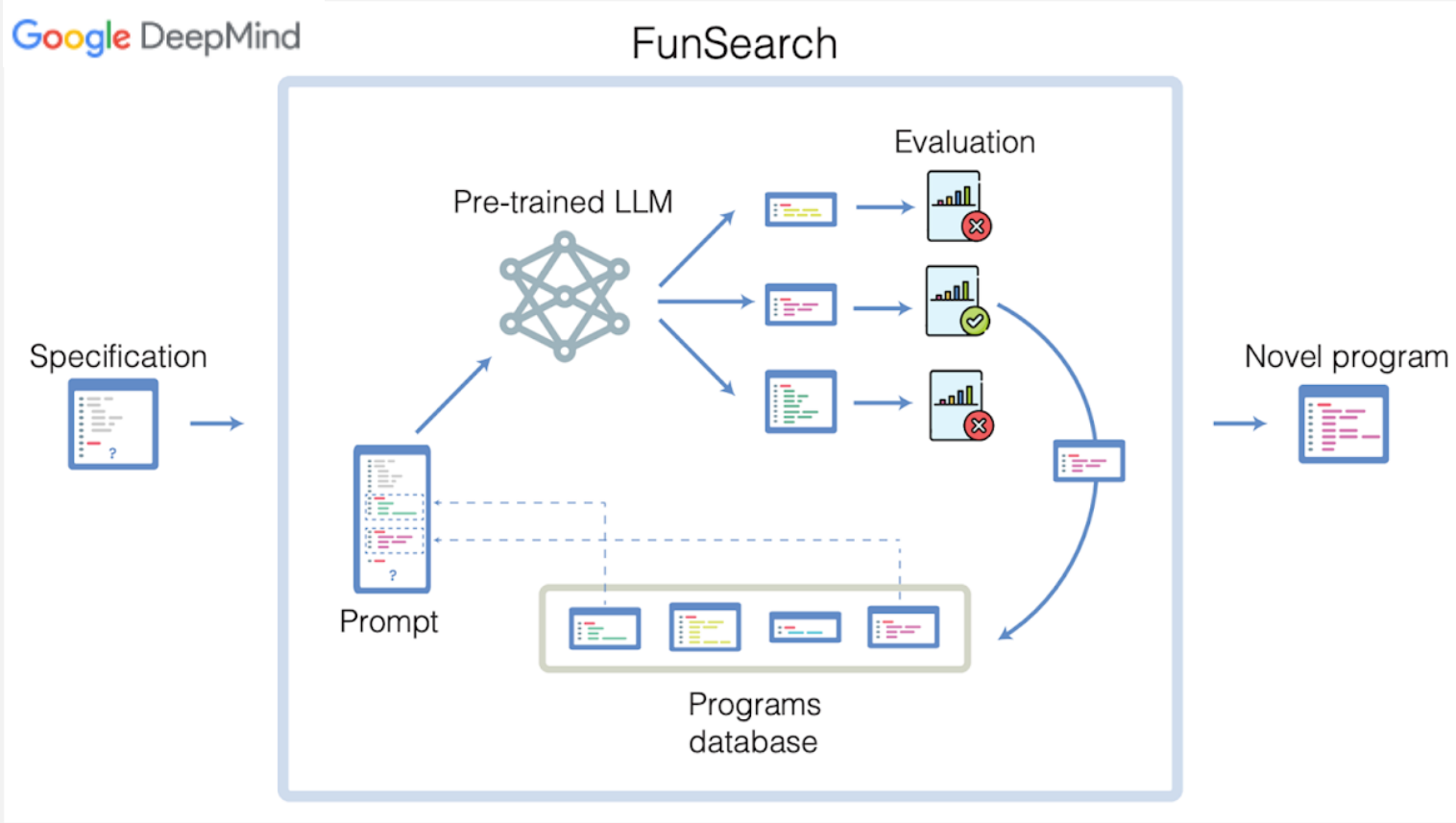
AlphaGeometry solving an Olympiad problem: Problem 3 of the 2015 International Mathematics Olympiad (left) and a condensed version of AlphaGeometry's solution (right). The blue elements are added constructs. AlphaGeometry's solution has 109 logical steps.





Solving olympiad geometry without human demonstrations, 2024

FunSearch: Mathematical discoveries from program search with LLMs



Мы представляем **FunSearch** (сокращение от «**поиск в пространстве функций**»), **эволюционную процедуру, основанную на объединении предварительно обученного LLM с систематическим оценщиком**. Применяя FunSearch к центральной проблеме экстремальной комбинаторики — проблеме предельного множества **— мы обнаруживаем новые конструкции больших предельных множеств, выходящие за рамки самых известных, как в конечномерных, так и в асимптотических случаях**. Мы демонстрируем универсальность FunSearch, применяя его к алгоритмической задаче, онлайн-упаковке корзин, и **находим новые эвристики, которые улучшают широко используемые базовые показатели.** В отличие от большинства подходов к компьютерному поиску, **FunSearch ищет программы, которые описывают, как решить проблему, а не каково ее решение**. Помимо того, что **найденные программы являются эффективной и масштабируемой стратегией**, они, как правило, **более интерпретируемы**, что обеспечивает обратную связь между экспертами в предметной области

The FunSearch process:

- The LLM is shown a selection of the best programs it has generated so far (retrieved from the programs database), and asked to generate an even better one.
- The programs proposed by the LLM are automatically executed, and evaluated.
- The best programs are added to the database, for selection in subsequent cycles.
- The user can at any point retrieve the highest-scoring programs discovered so far.

Алгоритм решения любой задачи можно описать как программу (инструкцию) на специальном или естественном языке: *Поиск решения как программы оказывается универсальным научным навыком!*

Mathematical discoveries from program search with large language models, 2024

Перспективы, проблемы и вызовы 2024

Наблюдение: большое объединение уже началось (DL)

Deep Learning (глубокое обучение)

CV	NLP	RL	GM
2011 CNN			
	Transformer	Deep RL	GAN
			Diffusion models
Multi-modal	Multi-modal	Multi-modal	

2023 **Универсальные модели**

Это сильный аргумент в пользу того, что можно ожидать замедления или завершения текущего этапа «революции в ИИ», связанного с нейросетями

LLM уже обучены на большей части информации, накопленной человечеством

Обобщение, экстраполяция: Ю.Визильтер, 03.2023 (сугубо IMHO)

Прогноз: большое объединение в ближайшие годы (DL)

DL

CNN

Deep RL

GAN

Transformer, LLM

Diffusion models

Multi-modal

2023 **Универсальные глубокие модели**

Обобщение, экстраполяция: Ю.Визильтер, 03.2023 (сугубо IMHO)

Прогноз: большое объединение в ближайшие годы (DL, ML)

Машинное обучение

DL	ML
CNN	Деревья решений
Deep RL	Байесовские решения
GAN	Логистическая регрессия
Transformer, LLM	Функции потенциалов
Diffusion models	Ансамбли
Multi-modal	

2023 **Универсальные глубокие модели**

2023 **Универсальные модели**

Обобщение, экстраполяция: Ю.Визильтер, 03.2023 (сугубо IMHO)

Прогноз: большое объединение в ближайшие годы (ИИ)

Искусственный интеллект

ИИ-2 (машинное обучение)

DL

CNN
Deep RL
GAN
Transformer, LLM
Diffusion models
Multi-modal

ML

Деревья решений
Байесовские решения
Логистическая регрессия
Функции потенциалов
Ансамбли

ИИ-1

Формальные системы
Символьное программирование
Логическое программирование
Базы знаний
Экспертные системы
Онтологии
Семантические сети

2023 **Универсальные глубокие модели**

2023 **Универсальные модели**

2025 **Универсальный гибридный искусственный интеллект**

Процедуры ML, ИИ-1, вычислительные алгоритмы, математическое моделирование и др. становятся *инструментами*, которыми пользуются LLM-агенты и AGI

AI for Robotics:

- **Open-world** (работа в новых местах и ситуациях)
- **Open-task** (с новыми задачами)
- **Open-Ended Learning** (когнитивное поведение, умение учиться)
- **Retrieval-based** (способность активно добывать информацию из различных источников)
- **Transparent** (прозрачность)
- **Explainable** (объяснимость)

Обобщение, экстраполяция: Ю.Визильтер, 03.2023 (сугубо IMHO)

Прогноз: большое объединение в ближайшие годы (ИИ+ЕИ = И^2)

Искусственный интеллект

ИИ-2 (машинное обучение)

DL

- CNN
- Deep RL
- GAN
- Transformer, LLM
- Diffusion models
- Multi-modal

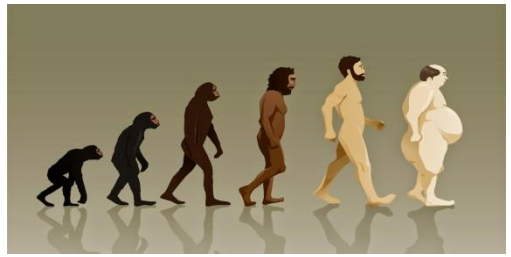
ML

- Деревья решений
- Байесовские решения
- Логистическая регрессия
- Функции потенциалов
- Ансамбли

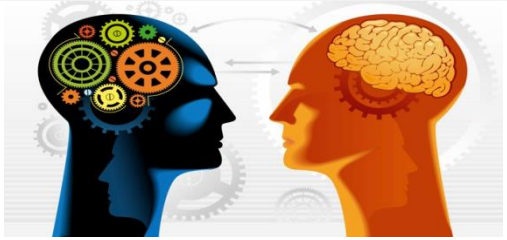
ИИ-1

- Формальные системы
- Символьное программирование
- Логическое программирование
- Базы знаний
- Экспертные системы
- Онтологии
- Семантические сети

Естественный интеллект



Человеко-машинные сети как усилитель интеллекта



2023 Универсальные глубокие модели

2023 Универсальные модели

2025 Универсальный гибридный искусственный интеллект

2030 Универсальный гибридный человеко-машинный интеллект

Обобщение, экстраполяция: Ю.Визильтер, 03.2023 (сугубо IMHO)

Вопросы, проблемы, вызовы

- Есть ли у нас сегодня теоретическое понимание того, как работают и обучаются нейронные сети? Достаточно ли нам этого понимания? Следует ли уделять больше внимания теоретическим основам нейросетевых технологий и машинного обучения?
- Следует ли ожидать дальнейшего бурного развития нейросетей, или потенциал этой технологий уже близок к полной реализации?
- Лежат ли по-прежнему в основе успешных нейросетевых решений сбор и обработка больших данных?
- Мешает ли сегодня недостаток данных для обучения решению практических задач? Как бороться с этой проблемой?
- Можно ли использовать на практике системы с постоянным дообучением (life-long learning) и активным самообучением?
- Остались ли ещё нерешенные научные проблемы в области компьютерного зрения?
- Какие нейросетевые модели будут востребованы для обработки сенсорных данных - сверточные, трансформерные, гибридные, какие-то новые?
- Большие языковые модели (LLM) действительно способны рассуждать, или они просто воспроизводят то, что запомнили?

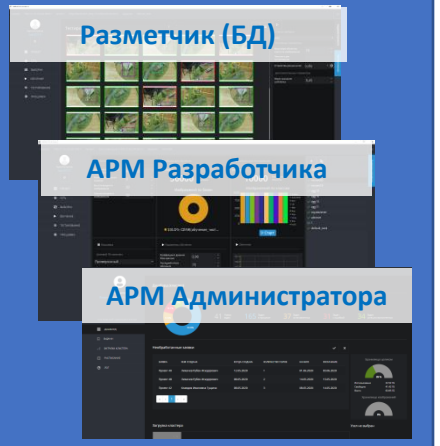
- Что будет основным инструментом разработчика интеллектуальных систем в ближайшие годы - машинное обучение или инженерия запросов?
- Каковы перспективы гибридных систем, сочетающих символьный ИИ с машинным обучением?
- Когда обучение нейросетей с подкреплением начнёт показывать серьёзные практические результаты за пределами компьютерных игр?
- Каковы перспективы автоматического программирования при помощи LLM?
- Насколько появление LLM и генеративных агентов повлияет на развитие науки в ближайшие 5-10 лет?
- Обладают ли генеративные нейросети творческими способностями, или это только их грубая имитация?
- Насколько важна проблема непрозрачности нейросетевых решений? Каковы пути решения этой проблемы?
- Насколько важна с практической точки зрения уязвимость нейросетей? Как бороться с атаками на нейросети?
- Как обеспечить доверие и безопасность при практическом применении нейросетевых решений, особенно в автономной робототехнике?

- Вытеснит ли новая аппаратно-программная технология нынешний нейросетевой ИИ на основе графических карт и тензорных ускорителей? Если да, то когда, и с каким типом вычислителя: нейроморфным, оптическим, квантовым, ещё каким-либо другим?
- Важно ли сегодня иметь полный отечественный стек аппаратных и программных технологий для реализации и обучения нейросетей?
- Следует ли практическим разработчикам ориентироваться на отечественные решения, или лучше полностью опираться на открытый стек средств обучения и использования нейросетевых моделей?
- Какие нейросетевые технологии наиболее сильно повлияют на нашу жизнь в ближайшие 5 лет? 10 лет? 15 лет?
- Можно ли считать, что общий или сильный ИИ (AGI)уже создан? Если нет, то когда его можно ожидать?
- Сейчас многие задачи ИИ уже решаются superhuman (немного лучше уровня человека): возможно ли создание нейросетевого superhuman ИИ, уровень решений которого будет настолько превосходить человеческий, что его замыслы, цели и решения будут людям совершено непонятны?

Унифицированная платформа нейросетевой разработки «Платформа-ГНС» (2018-2023)

Интегрированная среда

- **Работа с данными**
- **Формирование и обучение ГНС**
- **Аппаратная реализация ГНС**



Прикладные пакеты

- Готовые типовые решения
- Интегрированные решения
- Пользовательские решения

Типовые задачи

Компьютерное зрение

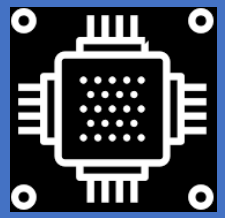
Анализ сигналов

Средства встраивания решений PlatformAPI

Унифицированная платформа

- **Сертифицируемый на НДВ исходный код**
- Импорт/экспорт, работа с фреймворками
- **Готовые решения для типовых задач**
- **Поддержка отечественного АО и ОС**
- Контроль доступа к данным и проектам
- **Низкие требования к квалификации ИТР**

Средства аппаратной реализации

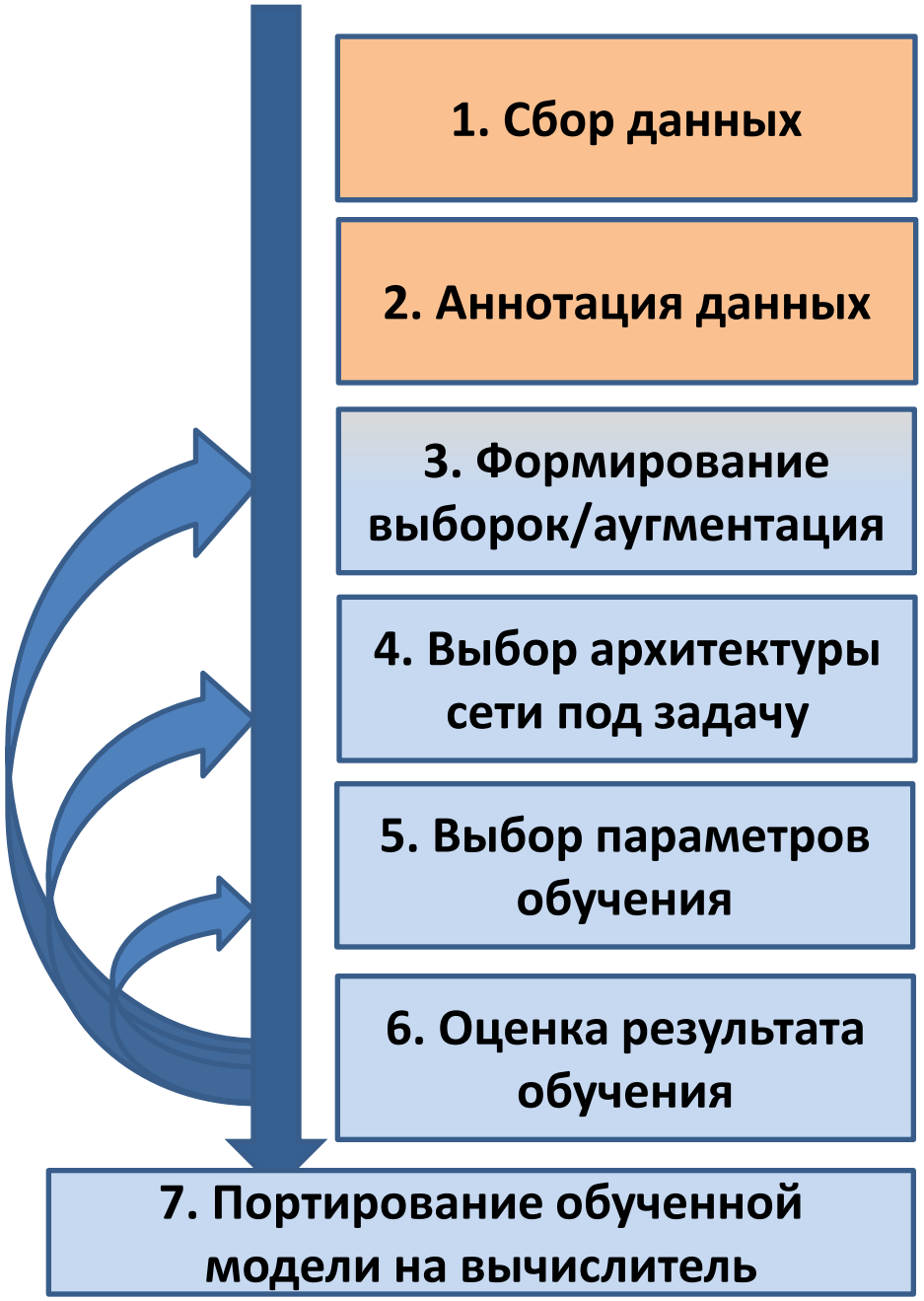


Библиотека машинного обучения (фреймворк) PlatLib

- **Динамические графы**
- **Распределенное обучение**
- **Унификация с PyTorch**

ГосНИИАС поддерживает и развивает полный отечественный программно-аппаратный стек технологий для ИИ

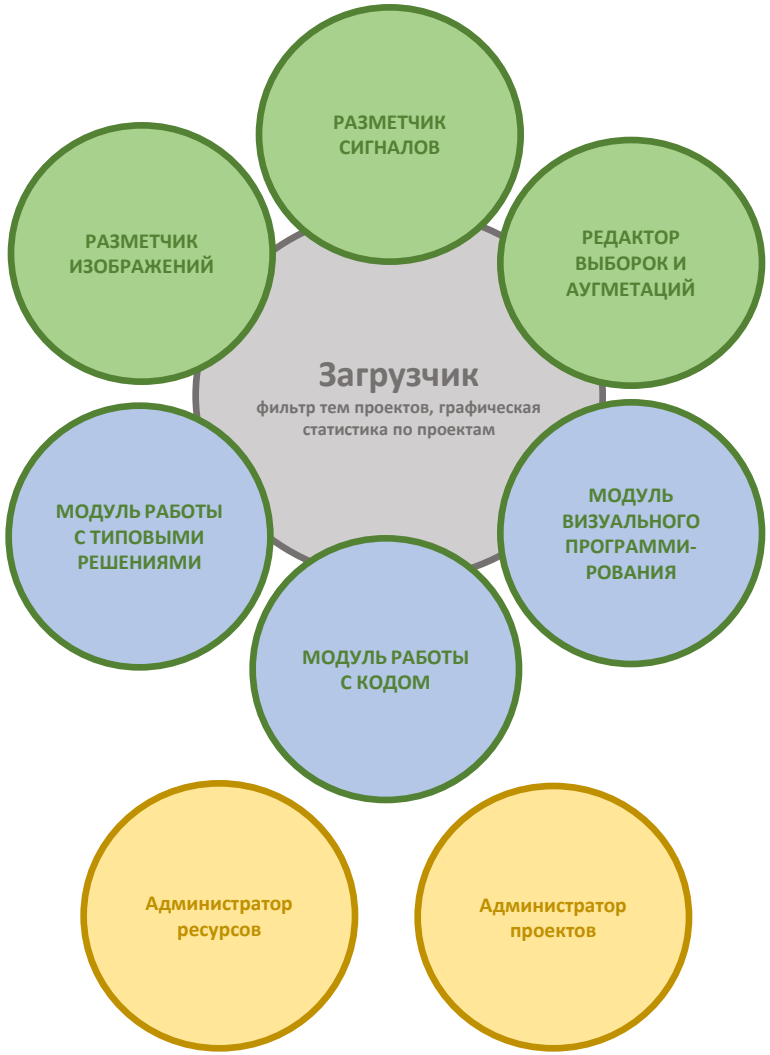
Процесс создания конечного решения. Полный цикл разработки в Платформе-ГНС



Средства аппаратной реализации: Целевые аппаратные платформы

Направление	Поддержка к настоящему моменту	Интегрирование существующих решений или разработка	Состояние Работ	Осуществлено практическое тестирование
CPU x86-64	Да	Разработка	Завершены	Да
CPU ARM (Байкал-М)	Да	Разработка	Завершены	Да
CPU + NVidia GPU (CC 6.0+)	Да	Интегрирование	Завершены	Да
CPU + AMD GPU (gfx9)	Да	Интегрирование	Завершены	Да
NeuroMatrix (НТЦ «Модуль»)	Да	Разработка	Завершены	Да
ПЛИС UltraScale+ (Xilinx)	Да	Разработка	Завершены	Да
Robodeus (НПЦ «Элвис»)	Да	Интегрирование	Завершены	Да
CPU Эльбрус (АО «МЦСТ»)	Да	Разработка	Завершены	Да
СнК IVA (IVA technologies)	Да	Интегрирование + разработка	Завершены	Да

Средства аппаратной реализации: Целевые аппаратные платформы



Сравнительное тестирование отечественных аппаратных средств в задачах ИИ

Решение для российского ОПК:
программа сравнительного тестирования от ГосНИИАС

- ✓ Тестирование вычислителей отечественных производителей проводится на базе Платформы-ГНС
- ✓ На первом этапе рассматриваются широко распространённые задачи: классификация, обнаружение и сегментация
- ✓ В тестировании учувствуют классические архитектуры ГКНС

Платформа ГНС ФГУП «ГосНИИАС»

Набор задач для тестирования (ver. 02.2022)

Задача классификации изображений

БД для обучения и тестирования	ILSVRC 2014
Используемые архитектуры	✓ ResNet 18 ✓ ResNet 34 ✓ ResNet 50 ✓ MobileNet v1
Используемые метрики качества	Accuracy, Top5
Форма входных данных	3x224x224

Задача обнаружения объектов

БД для обучения и тестирования	MS COCO
Используемые архитектуры	✓ SSD-ResNet 34 ✓ SSD-Mobilenet v2
Используемые метрики качества	mAP
Форма входных данных	3x512x512

Задача семантической сегментации

БД для обучения и тестирования	CityScapes
Используемые архитектуры	✓ UNetVGG16
Используемые метрики качества	mIOU
Форма входных данных	3x512x512

Решение для российского ОПК:
программа сравнительного тестирования от ГосНИИАС

- ✓ **«Датацентр»** – категория подразумевает отсутствие ограничений на потребляемую мощность.
- ✓ **«Бортовое исполнение»*** – категория подразумевает ограничение в энергопотреблении вычислителя до 100 Ватт
- ✓ **«Бортовое исполнение (20 Ватт)»*** – категория подразумевает ограничение в энергопотреблении вычислителя до 20 Ватт.

*размер пакета (batch) = 1

Платформа ГНС ФГУП «ГосНИИАС»

Состав тестируемых аппаратных платформ (ver. 02.2022)

Датацентр

Xeon Silver 4112	Intel (2017)
Эльбрус 8 С	АО «МЦСТ» (2014)
1892ВМ248 «RoboDeus»	АО НПЦ «ЭЛВИС» (2021)

Бортовое исполнение

K1945ВМ028 IVA H TPU	IVA Technologies (ГК «ХайТэк») (2021)
Jetson AGX Xavier	Nvidia (2020)

Бортовое исполнение (20 Ватт)

1879ВМ8Я NM6408	АО НТЦ «Модуль» (2017)
1892ВМ248 «RoboDeus»	АО НПЦ «ЭЛВИС» (2020)
Jetson Nano	Nvidia (2020)
NUC, Pentium Silver J5005	Intel (2019)
NUC, UHD Graphics 605	Intel (2019)

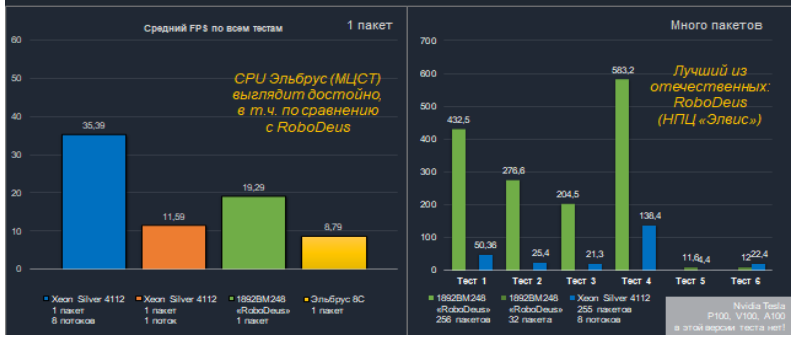
Тестирование отечественных вычислителей ver. 02.2022

Тестирование считается успешным, если пройден 5% порог качества по соответствующей метрике

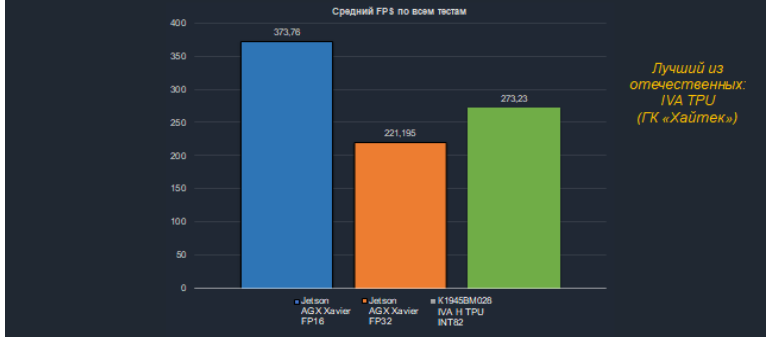
	CNN	Тест	NM6408	Эльбрус-8С	Robodeus	IVA H TPU	Xeon Silver 4112	NUC, Pentium Silver J5005	NUC, UHD Graphics 605	Nvidia Jetson AGX Xavier	Jetson Nano
Класси-фикация	ResNet18	№1	✓	✓	✓	✓	✓	✓	✓	✓	✓
	ResNet34	№2	✓	✓	✓	✓	✓	✓	✓	✓	✓
	ResNet50	№3	✓	✓	✓	✗	✓	✓	✓	✓	✓
	MobileNet-v1	№4	✓	✓	✓	✗	✓	✓	✓	✓	✓
Обнару-жение	SSD-ResNet34	№5	✓	✓	✓	?	✓	✓	✓	✓	✓
	SSD-MobileNet-v2	№6	✓	✓	✓	✓	✓	✓	✓	✓	✓
Сегмен-тация	Unet-VGG16	№7	✓	✗	?	?	✓	✓	✓	✓	✓

✓ - тест пройден ✗ - тест не пройден по качеству ? - данные не предоставлены

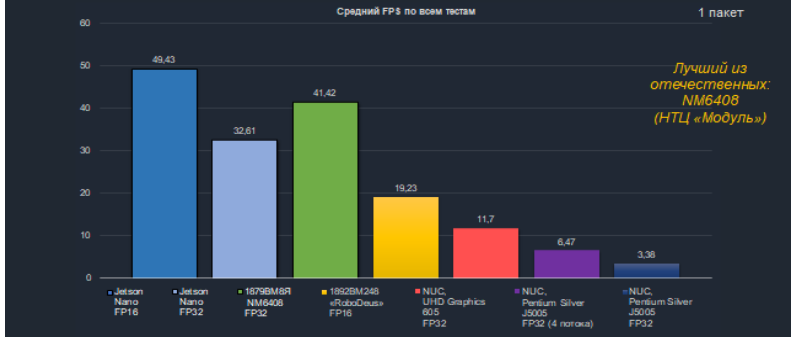
Результаты сравнительного тестирования (ver. 02.2022) для категории «Датацентр»



Результаты сравнительного тестирования (ver. 02.2022) для категории «Бортовое исполнение»



Результаты сравнительного тестирования (ver. 02.2022) для категории «Бортовое исполнение (20 Ватт)»



Платформа-ГНС и PLAT: Перспективы развития (дорожная карта)

Версия 1.7

Новые типы данных I

Автоматическое обучение I

Интеллектуальные помощники

Версия 2.0 «Автомат»

Облачное решение

Абстрактное типовое решение

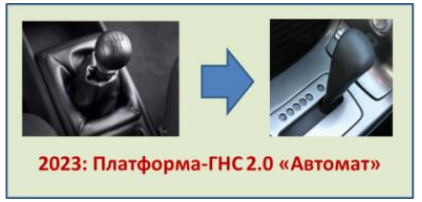
Автоматическое обучение II

Новые типы данных II

Новые прикладные пакеты RL, NLP,...

Большие и универсальные модели

Доступна для тестирования и использования с октября 2023



Документация, презентации, видеоматериалы

2022

2023

2024

АКТУАЛЬНЫЕ ТЕНДЕНЦИИ И РЕЗУЛЬТАТЫ В ОБЛАСТИ КОМПЬЮТЕРНОГО ЗРЕНИЯ, ГЕНЕРАЦИИ ДАННЫХ И ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ (2020-2024)

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН, директор по направлению – руководитель научного комплекса «Искусственный интеллект и техническое зрение» ФАУ «ГосНИИАС», *viz@gosniias.ru*

Спасибо за внимание!

Расширенная версия доклада

Семинар НИУ ВШЭ по высокопроизводительным вычислениям

Москва, 04.06.2024