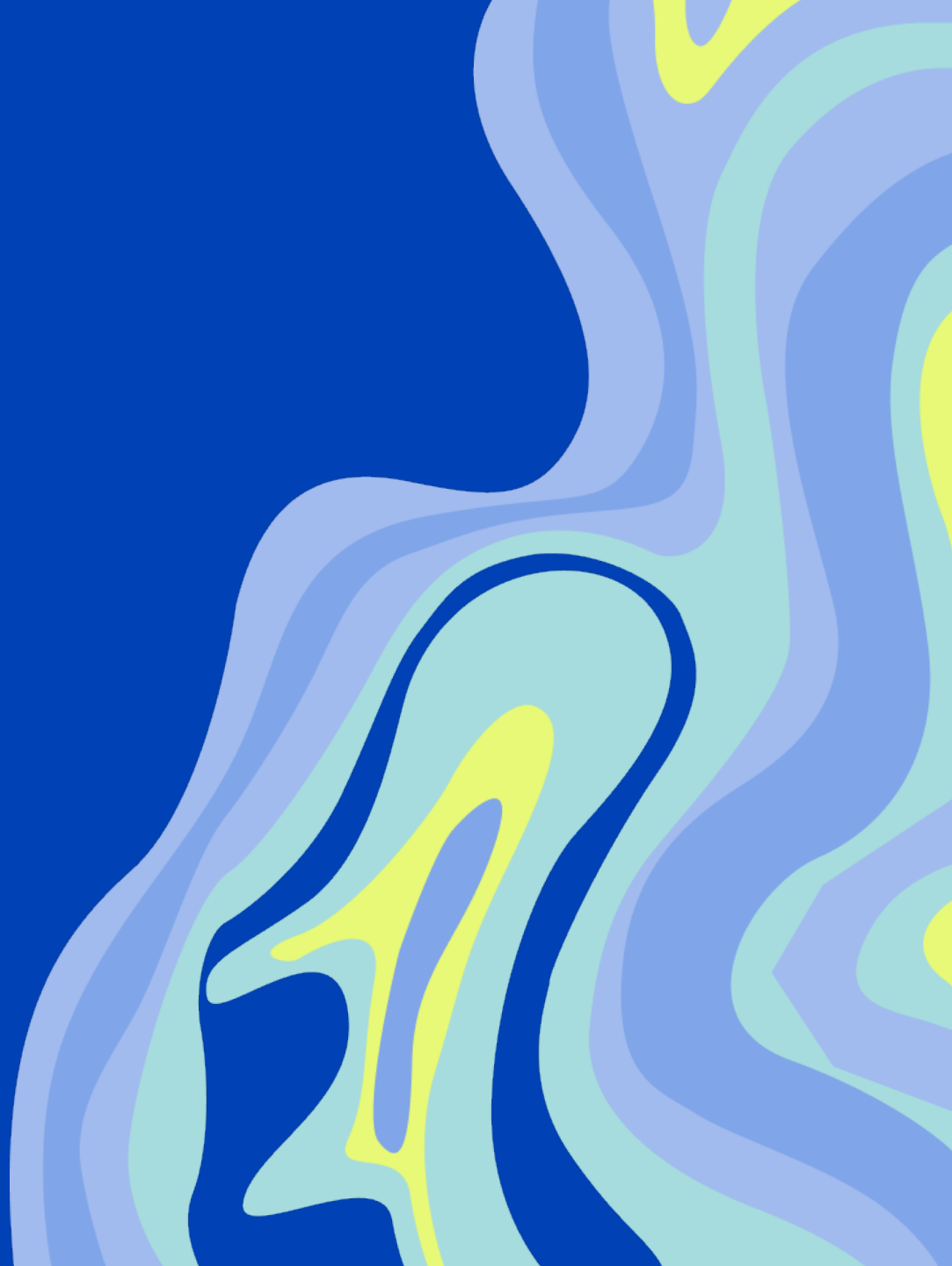
KANDINSKY

ICOMP
INTERNATIONAL CONFERENCE
ON COMPUTATIONAL OPTIMIZATION

Diffusion (and more) text-to-video generation methods

Denis Dimitrov

Managing director in data science, Sber AI
Scientific advisor, AIRI



Agenda

«Time-lapse of a flower blooming:
growth, beauty, and the passage of time»
by Kandinsky Video 1.1

01 Task history

02 Diffusion generation of images from text, Kandinsky - 3

03 Diffusion generation of videos from text

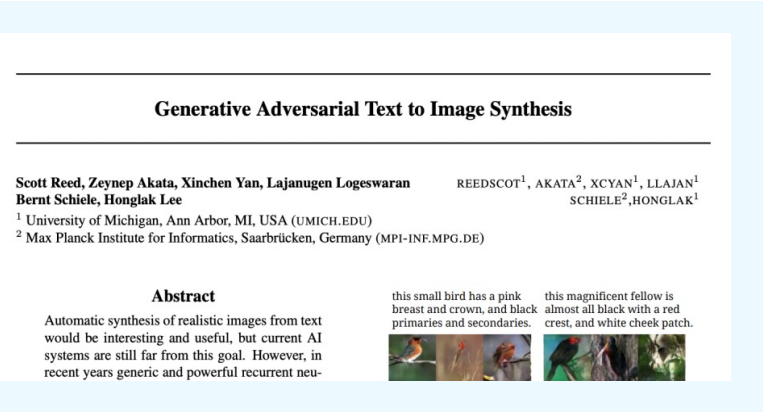
04 Models Application & Conclusions

01

Task history

Brief history of visual generative models

- 2014: the emergence of GAN was a breakthrough in the development of image generation models
- 2016: the first mention of **image generation from textual descriptions** was made at S. Reed et al. (ICML 2016 and NeurIPS 2016)
- Since 2017, no ML/DL/CV conference A/A* passes without publications on the topic of **image** synthesis from textual descriptions
- 2018: one of the first works on **video** generation from text descriptions, Y. Li et al.
- Since 2020 — the most active research and development of **text-to-image** architectures
- Since 2023 — the most active research and development of **text-to-video** architectures



Generative Adversarial Text to Image Synthesis

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran
Bernt Schiele, Honglak Lee

REEDSCOT[1], AKATA[2], XCYAN[1], LLAJAN[1]
SCHIELE[2],HONGLAK[1]

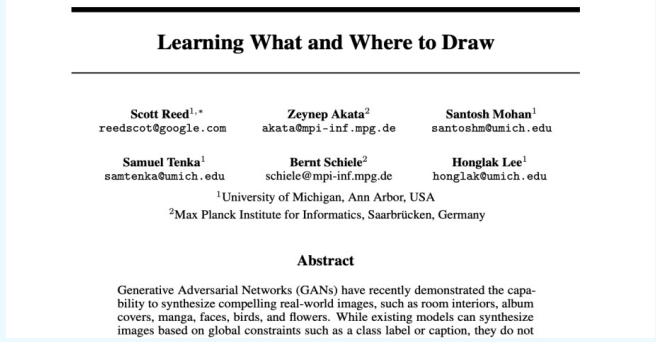
[1] University of Michigan, Ann Arbor, MI, USA (UMICH.EDU)
[2] Max Planck Institute for Informatics, Saarbrücken, Germany (MPI-INF.MPG.DE)

Abstract

Automatic synthesis of realistic images from text would be interesting and useful, but current AI systems are still far from this goal. However, in recent years generic and powerful recurrent neu-

this small bird has a pink breast and crown, and black primaries and secondaries.

this magnificent fellow is almost all black with a red crest, and white cheek patch.



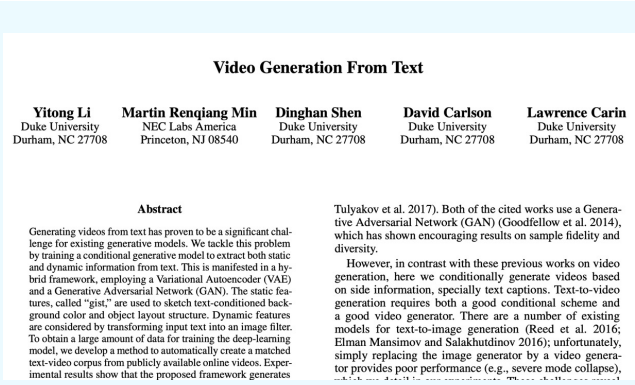
Learning What and Where to Draw

Scott Reed[1,*] reedscot@google.com
Zeynep Akata[2] akata@mpi-inf.mpg.de
Santosh Mohan[1] santoshm@umich.edu
Samuel Tenka[1] samtenka@umich.edu
Bernt Schiele[2] schiele@mpi-inf.mpg.de
Honglak Lee[1] honglak@umich.edu

[1]University of Michigan, Ann Arbor, USA
[2]Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

Generative Adversarial Networks (GANs) have recently demonstrated the capability to synthesize compelling real-world images, such as room interiors, album covers, manga, faces, birds, and flowers. While existing models can synthesize images based on global constraints such as a class label or caption, they do not



Video Generation From Text

Yitong Li, Duke University, Durham, NC 27708
Martin Renqiang Min, NEC Labs America, Princeton, NJ 08540
Dinghan Shen, Duke University, Durham, NC 27708
David Carlson, Duke University, Durham, NC 27708
Lawrence Carin, Duke University, Durham, NC 27708

Abstract

Generating videos from text has proven to be a significant challenge for existing generative models. We tackle this problem by training a conditional generative model to extract both static and dynamic information from text. This is manifested in a hybrid framework, employing a Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN). The static features, called "gist," are used to sketch text-conditioned background color and object layout structure. Dynamic features are considered by transforming input text into an image filter. To obtain a large amount of data for training the deep-learning model, we develop a method to automatically create a matched text-video corpus from publicly available online videos. Experimental results show that the proposed framework generates

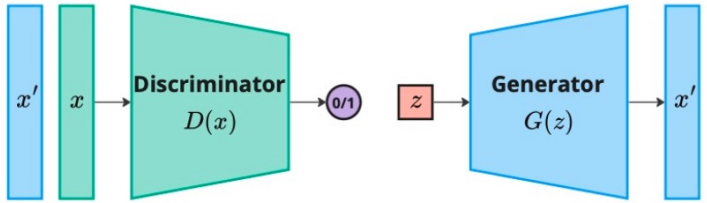
Tulyakov et al. 2017). Both of the cited works use a Generative Adversarial Network (GAN) (Goodfellow et al. 2014), which has shown encouraging results on sample fidelity and diversity.

However, in contrast with these previous works on video generation, here we conditionally generate videos based on side information, specially text captions. Text-to-video generation requires both a good conditional scheme and a good video generator. There are a number of existing models for text-to-image generation (Reed et al. 2016; Elman Mansimov and Salakhutdinov 2016); unfortunately, simply replacing the image generator by a video generator provides poor performance (e.g., severe mode collapse),

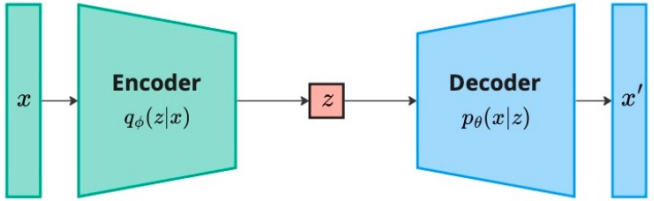
KANDINSKY

Generative architectures

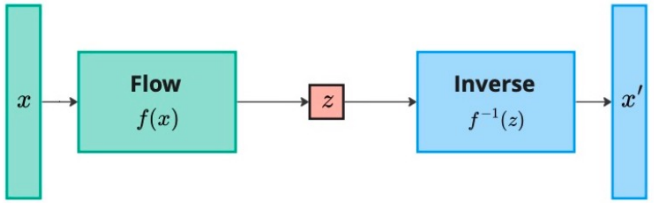
GAN:
Adversarial learning



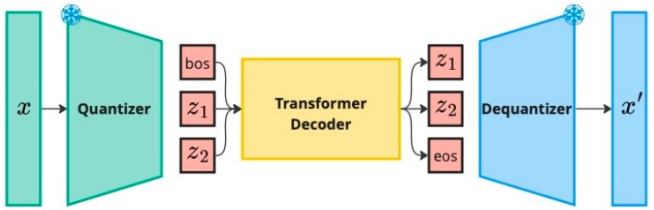
VAE:
ELBO maximization



Flow-based model:
Invertible distribution function

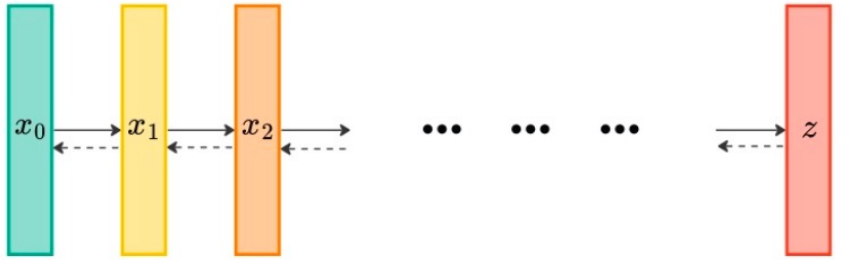


Autoregression:
Token-based generation of quantized image



Diffusion models:
Gradually add Gaussian noise
and then reverse

Flow-matching, rectified flows etc



Advantages

- Show better performance
- No adversarial training
- No mode collapse

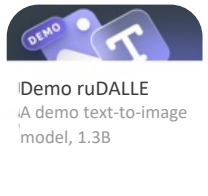
Disadvantages

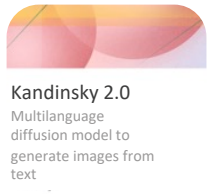
- Large inference time

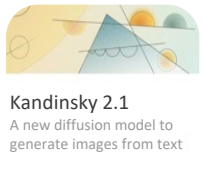
KANDINSKY

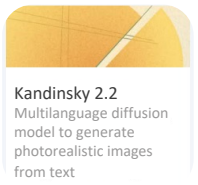
Text-to-visual models of Sber & AIRI

2 Nov/ 14 June 2021/ 2022	23 Nov 2022	4 April 2023	12 July 2023	22 Nov 2023	22 Nov 2023	22 April 2024	22 April 2024	28 May 2024
Malevich/ Kandinsky 1.0	**Kandinsky 2.0**	**Kandinsky 2.1**	**Kandinsky 2.2**	**Kandinsky 3.0**	**Kandinsky Video 1.0**	**Kandinsky 3.1**	**Kandinsky Flash**	**Kandinsky Video 1.1**
1.3B/ 12B	**2.0B**	**3.3B**	**4.8B**	**11.9B**	**17.5B**	**15.5B**	**11.9B**	**17.5B**
Autoregressive image generation from Russian text	Latent diffusion model for generating images from text in 101 languages	Latent diffusion model for generating images from text	Latent diffusion model for generating photorealistic images from text	New generation of latent diffusion models for generating photorealistic images from text	The first model in Russia for generating videos from text	New latent diffusion model for generating photorealistic images from text	New fast latent diffusion model for generating photorealistic images from text	New latent diffusion model for generating videos from text









KANDINSKY

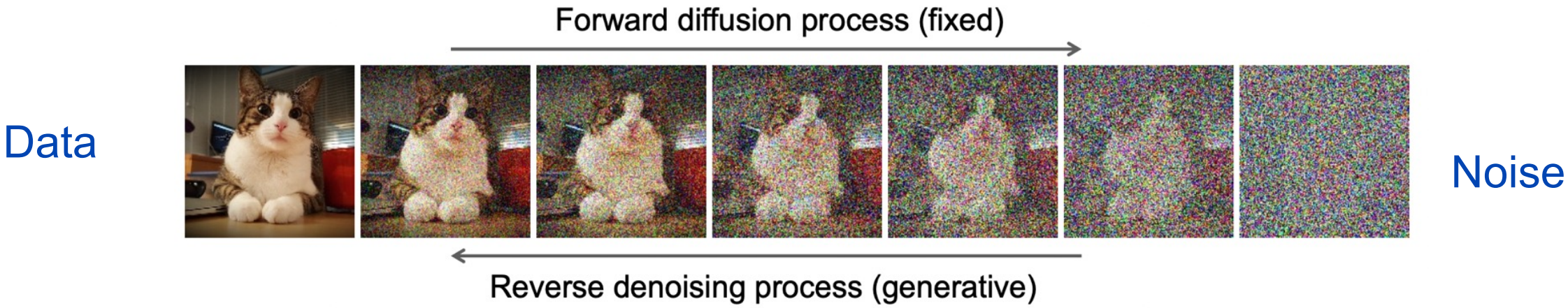
02

Diffusion generation of images from text, Kandinsky - 3

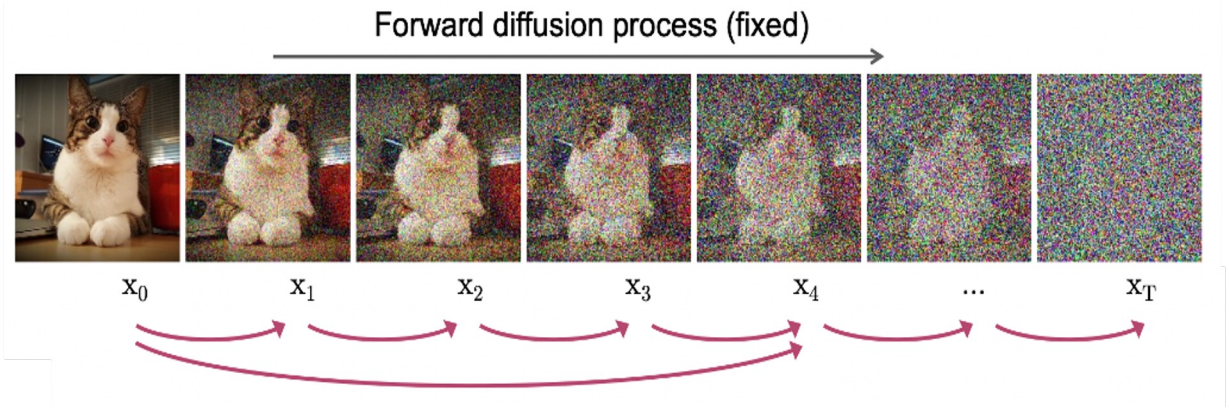
Denoising Diffusion Probabilistic Models (DDPM)

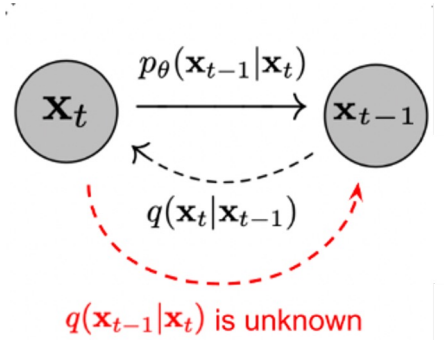
The diffusion process consists of two parts:

- Forward diffusion process – iterative noise addition
- Reverse diffusion process – iterative noise removal



Forward diffusion process





Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$:

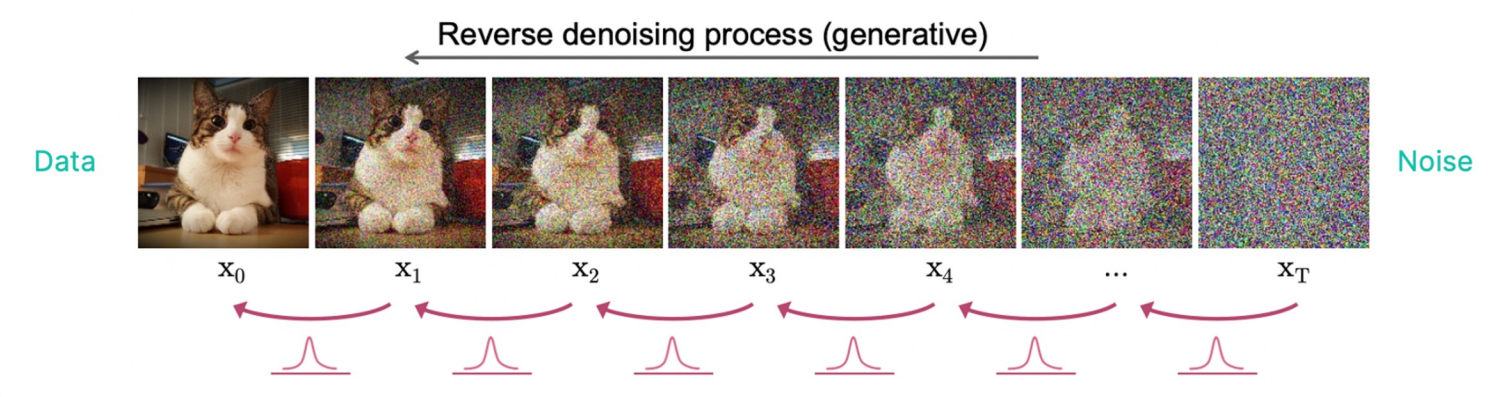
$$
\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \\
&= \cdots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \text{ , where}
\end{aligned}
$$

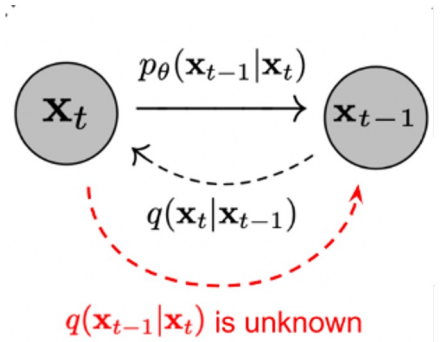
$\boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\bar{\boldsymbol{\epsilon}}_{t-2}$ merges to Gaussians

$$\Rightarrow \quad \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\,\epsilon \text{ , where } \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Training of a reverse diffusion process



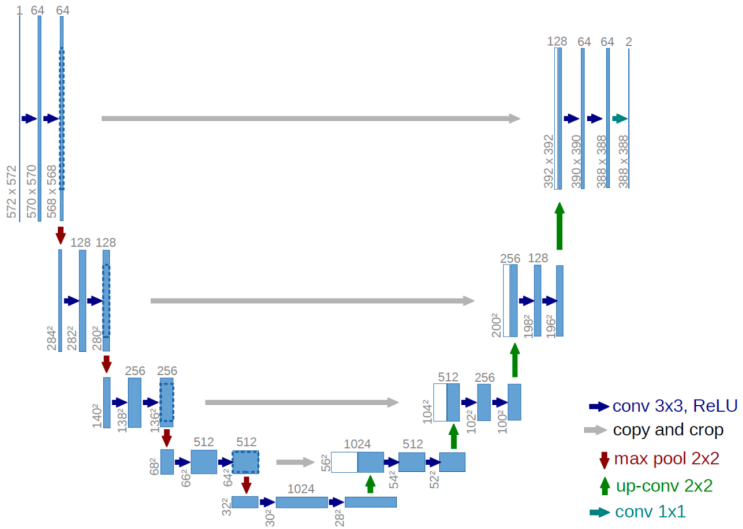


$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \underbrace{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2\mathbf{I})$$

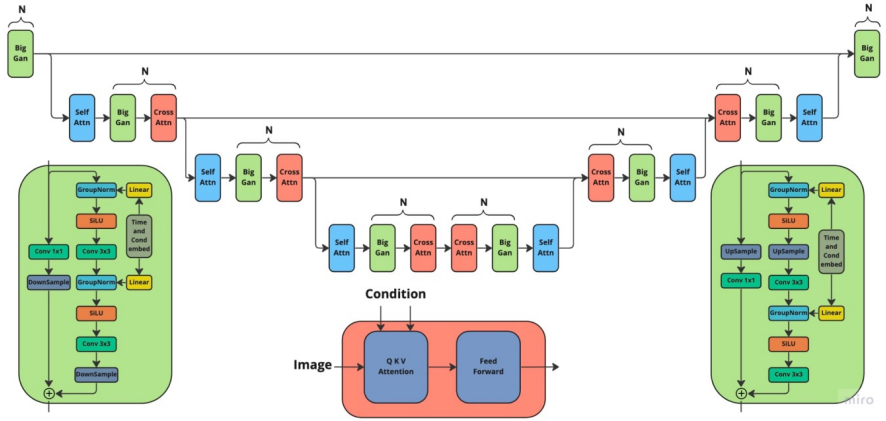
$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

KANDINSKY

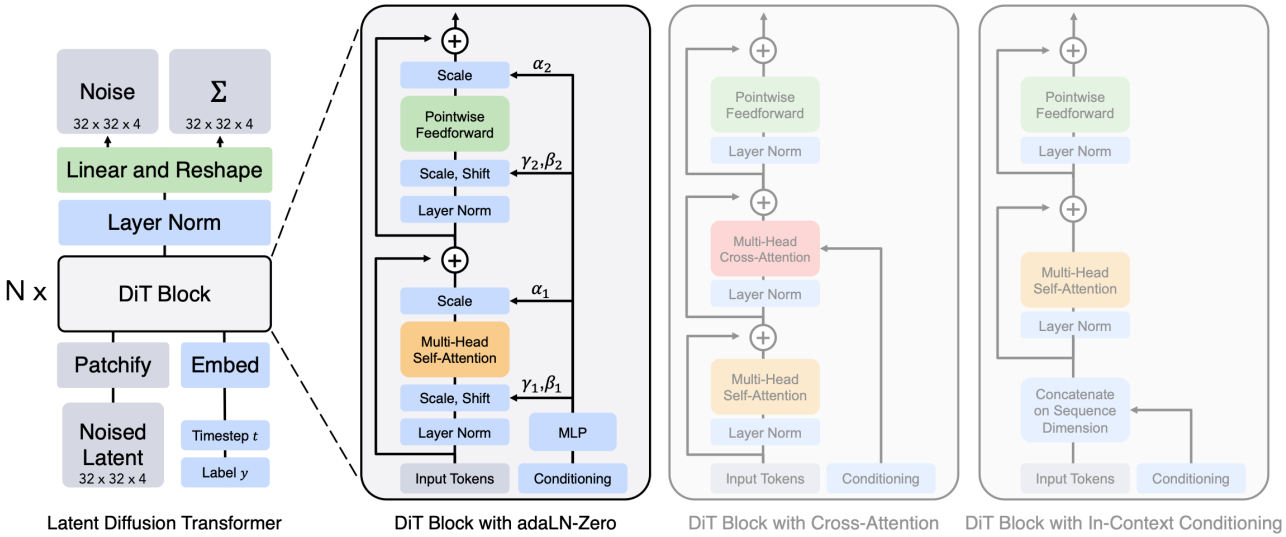
Base architectures – U-Net, DiT



Unconditional U-Net architecture



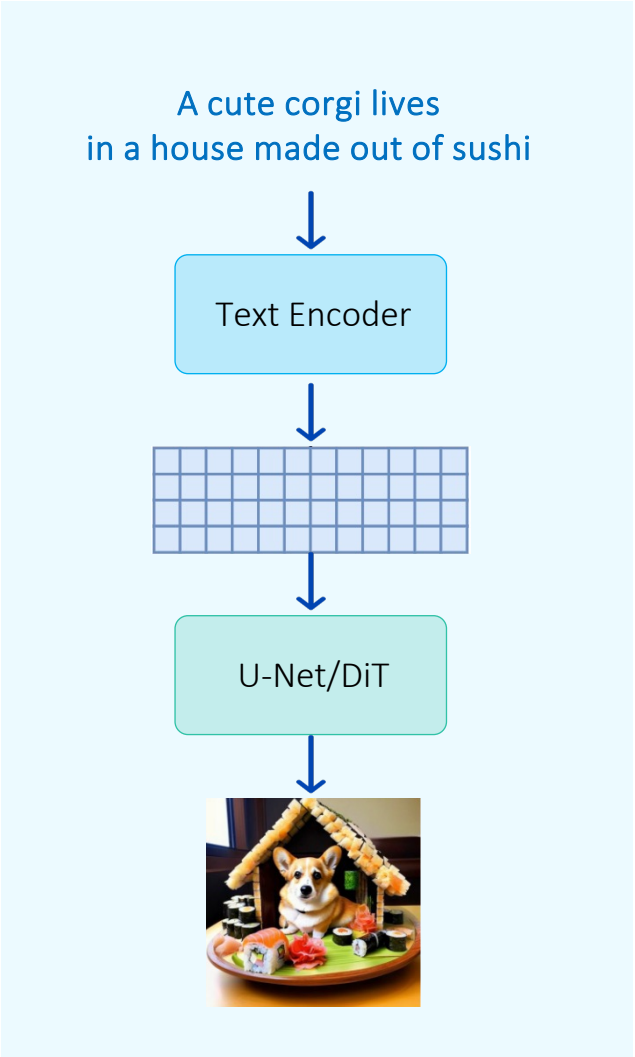
Conditional U-Net architecture

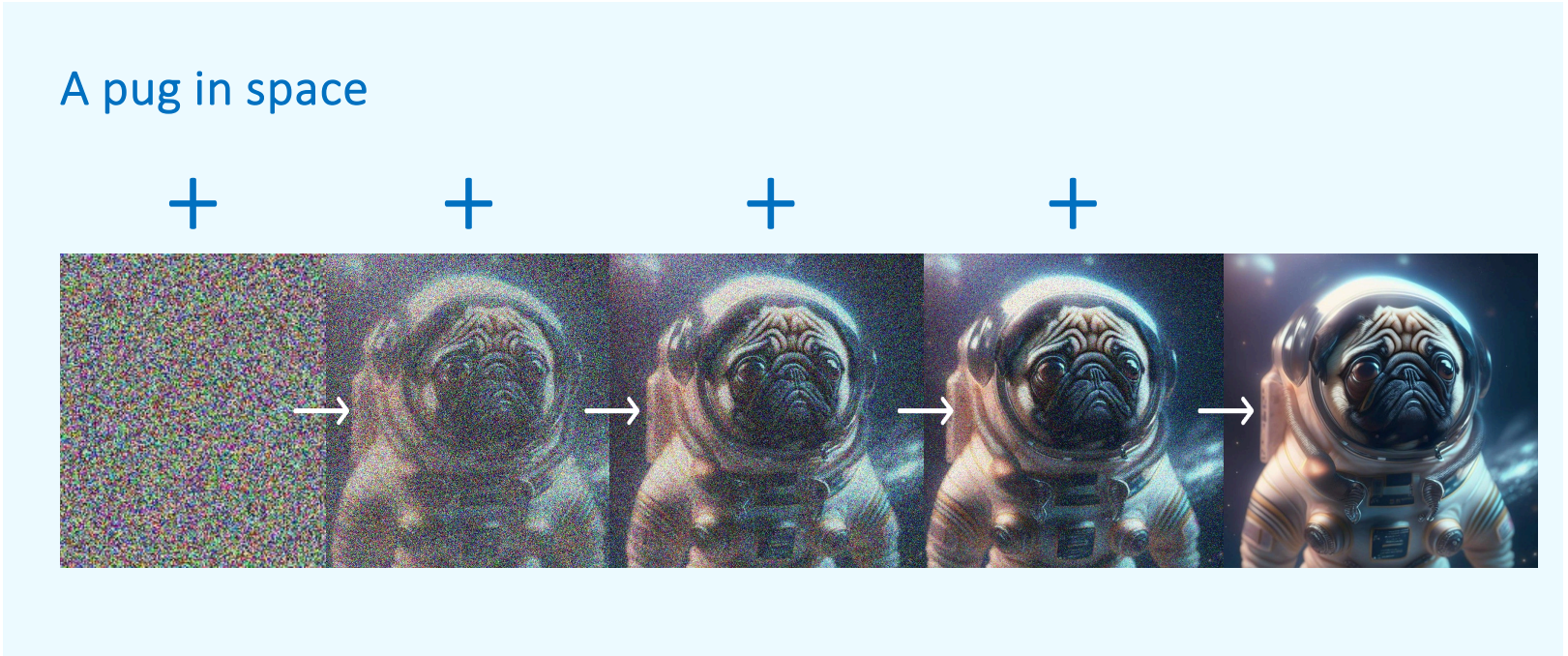


The Diffusion Transformer (DiT) architecture

KANDINSKY

General inference scheme of diffusion text-to-image models





KANDINSKY

Why DiT?

Architectures

U-Net	Mostly applied to image generation tasks
Transformer	▪ Compatible with LLM distributed learning frameworks ▪ Clear scaling ability ▪ More memory efficient than U-Net ▪ Sora-like models use this approach

Training type

Diffusion	▪ High quality of output generations ▪ Sora-like models use exactly this approach
Autoregression	LLM architectures can be used

Kandinsky 3.0 / 3.1

Inference pipeline

A cute corgi lives
in a house made out of sushi

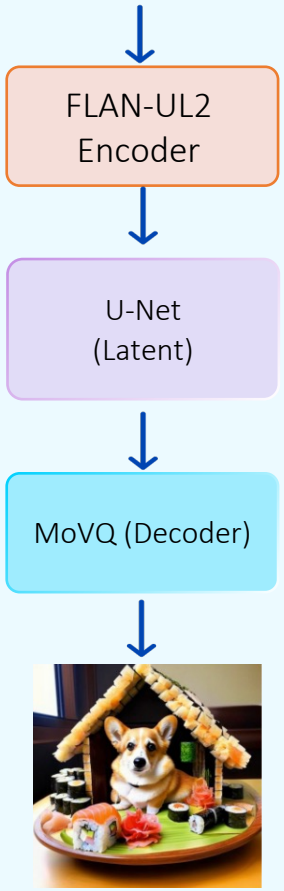
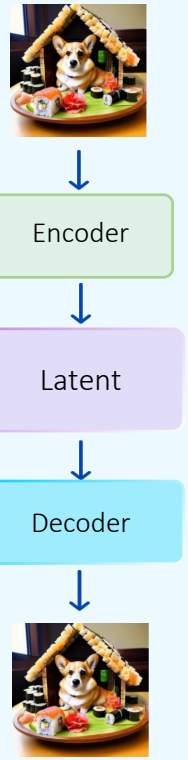
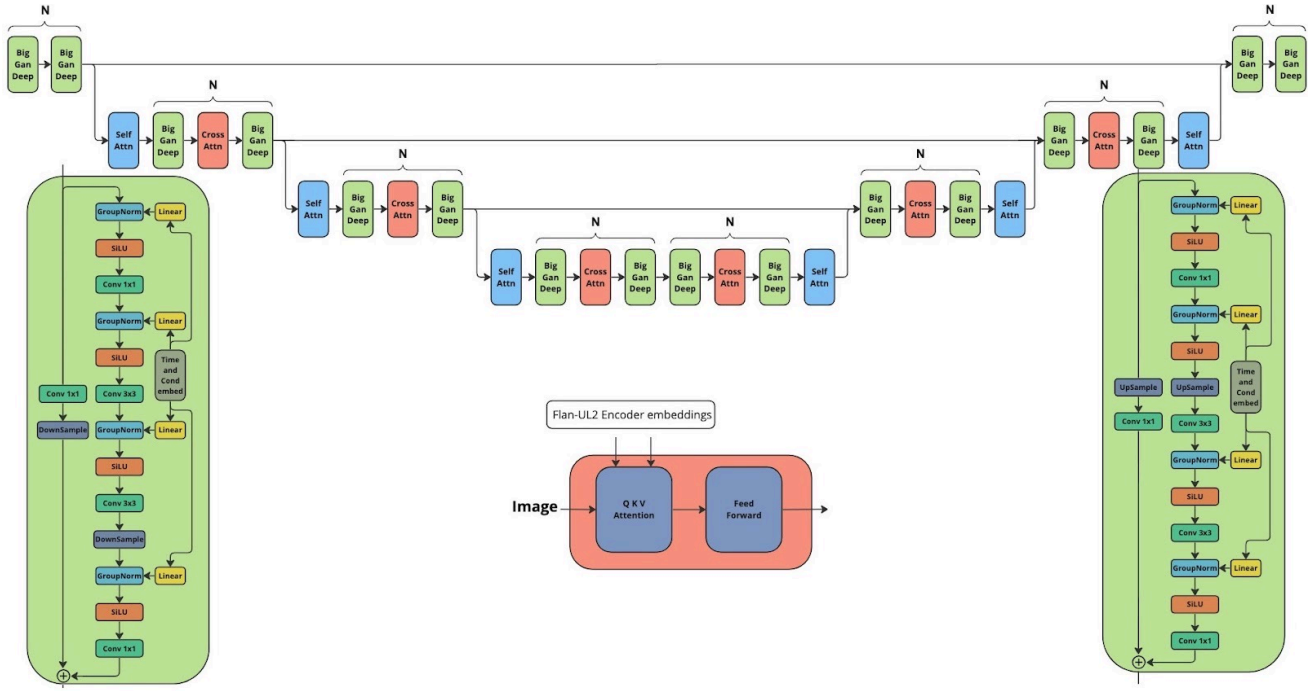


Image-VAE
(MoVQ)



Proprietary U-Net architecture, learning with cosine scheduling



KANDINSKY

Kandinsky 3.0 training

Amount of data for training (including filters):

256x256	1.1B
384x384	780M
512x512	450M
768x758	224M
Mixed	280M

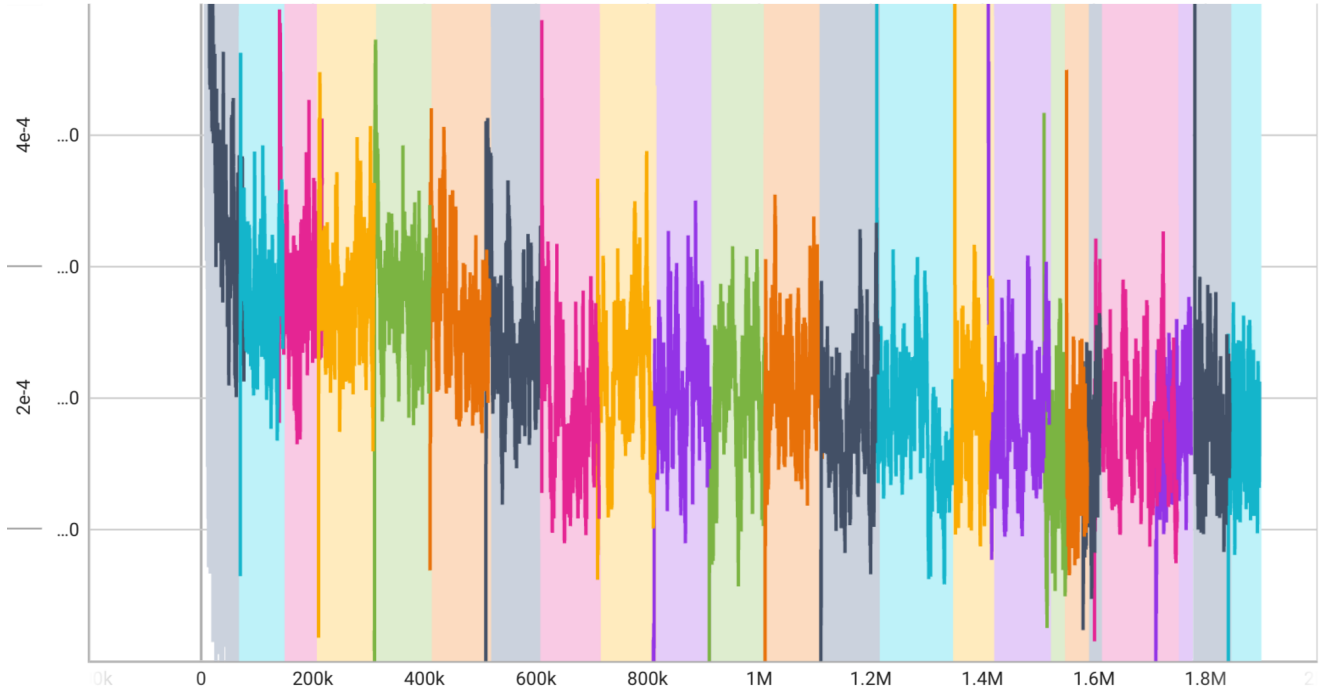
Datasets:

Not so good:

- LAION 800M
- COYO 700M
- TIGER 80M

Good:

- Depositphotos 100M
- RussianThemed 236K



Training loss

KANDINSKY

«Oriental tea party, baklava, tea, cups, honey, hospitality, harmonious serving, oriental sweets, anime style»

«Image in the golden circle of the Kul Sharif Mosque, Kazan, photorealism, maximum detail»

«Image of a black headdress with oriental gold ornament, gold thread embroidery, detailing»

Kandinsky Flash

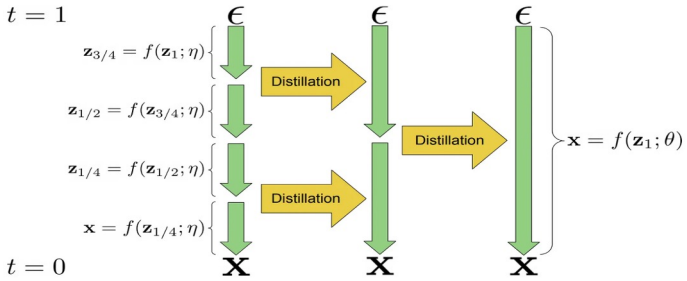
Kandinsky 3.0

Kandinsky Flash

A Pikachu with an angry expression and red eyes, with lightning around it, hyper realistic style

A panda is playing a guitar

A Pomeranian is sitting on the Kings throne wearing a crown. Two tiger soldiers are standing next to the throne



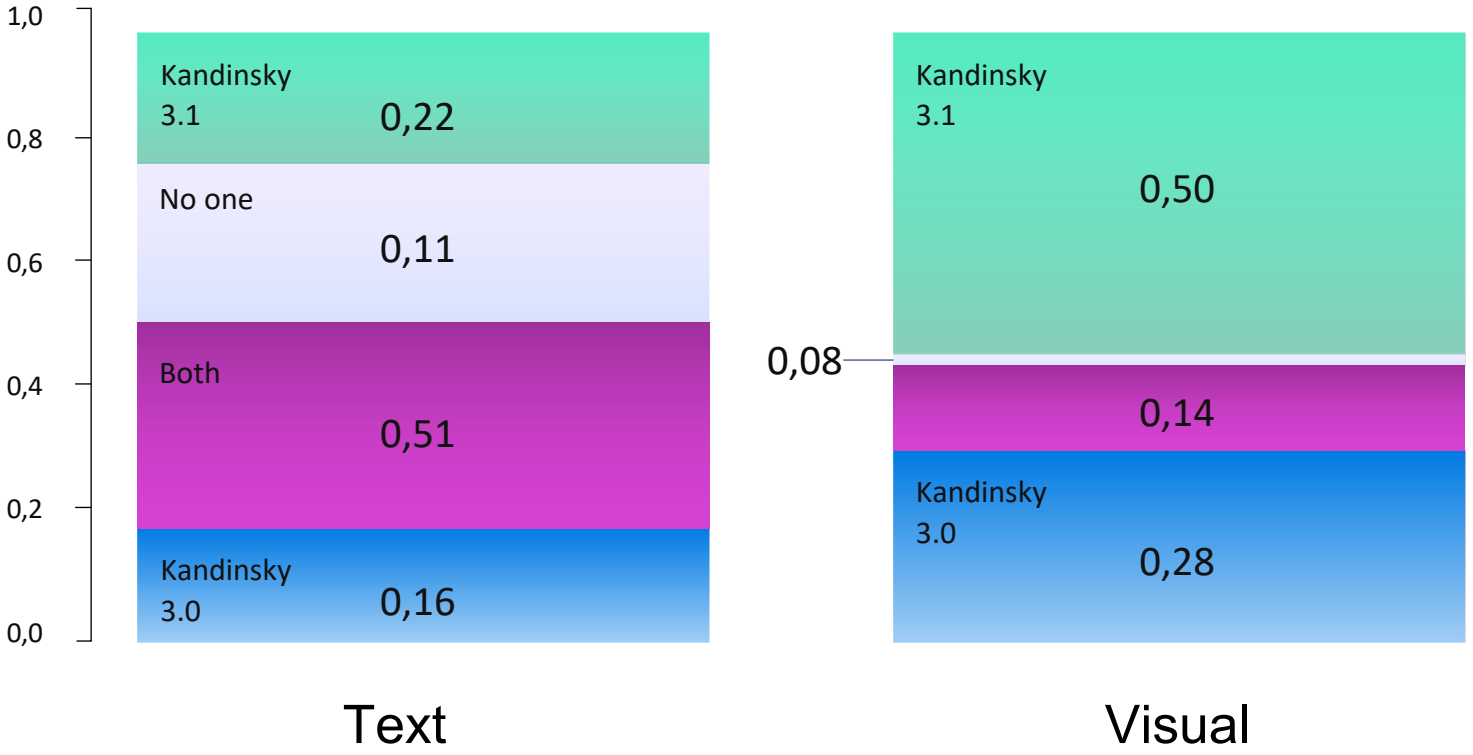
Speedup by more than x10

0.4s for generation of 1 image

KANDINSKY

Kandinsky 3.1 = Kandinsky 3.0 + Kandinsky Flash

Kandinsky 3.1 vs Kandinsky 3.0

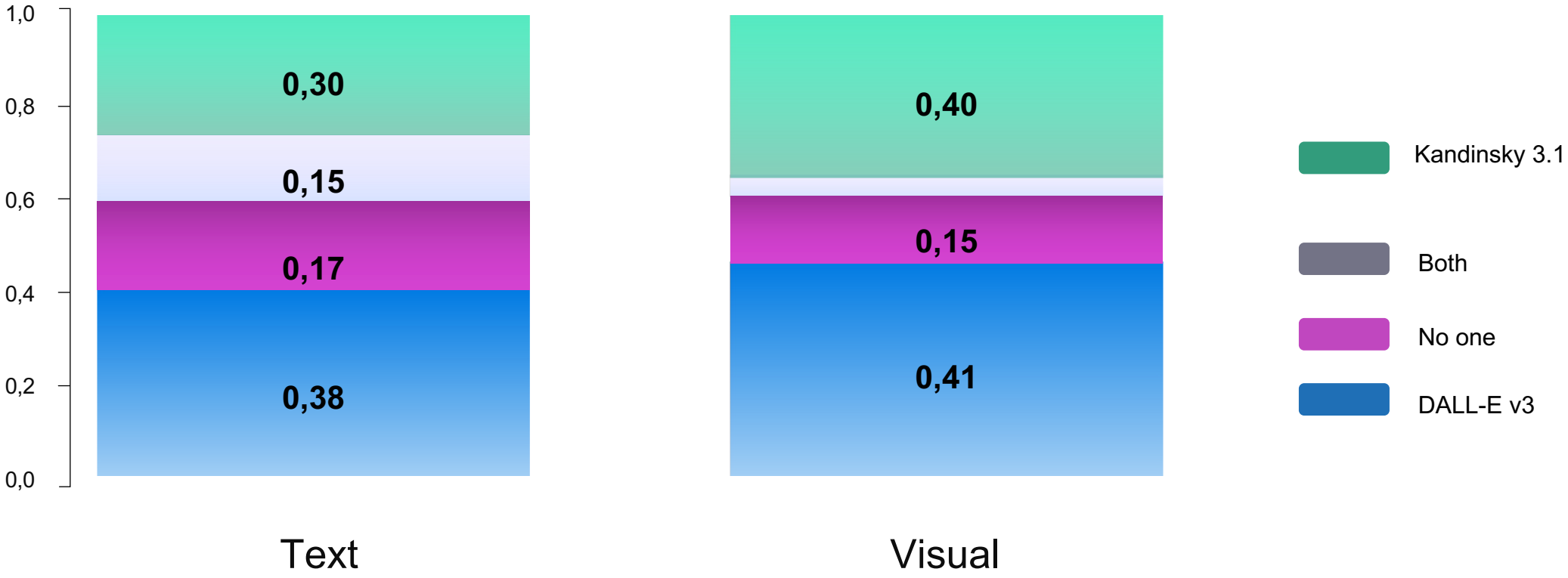


30 annotators
23 812 for text
24 741 for visual

KANDINSKY

Kandinsky 3.1 vs DALL-E 3

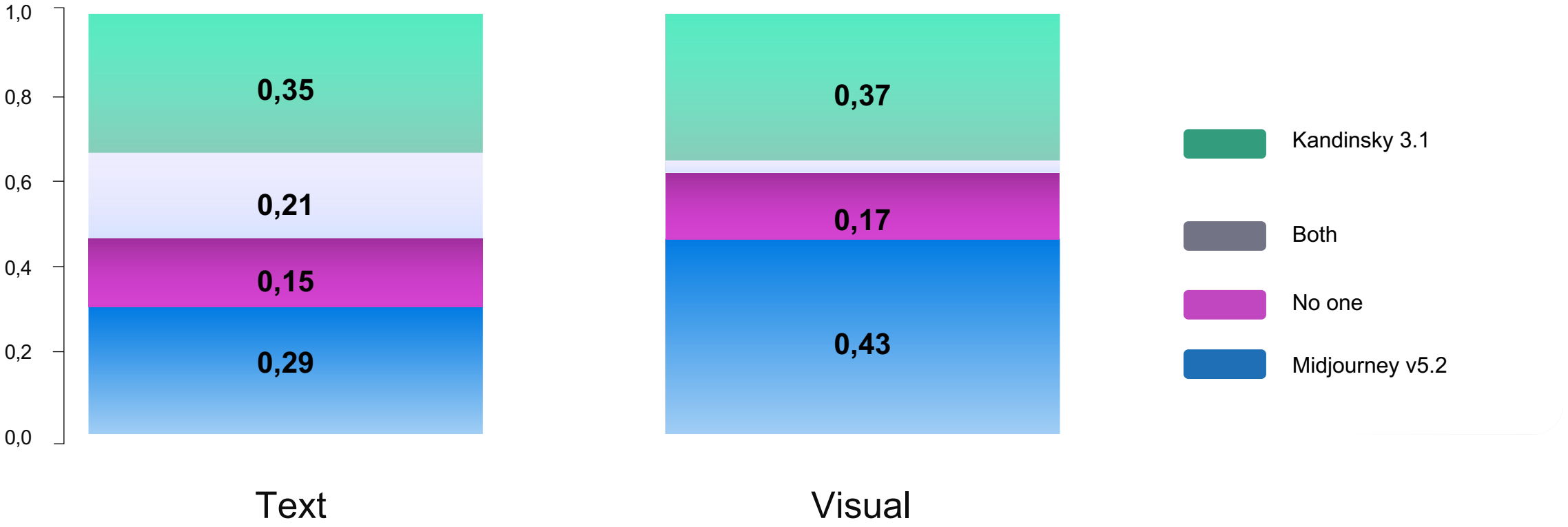
Kandinsky 3.1 vs DALL-E v3



KANDINSKY

Kandinsky 3.1 vs Midjourney v5.2

SBS, Kandinsky 3.1 vs Midjourney v5.2



KANDINSKY

03

Diffusion generation of videos from text

The main challenges in the field of video generation, which are tackled by the world's most advanced AI Labs

To develop a high-quality Sora-level long video generation model:
1) Up to 1 minute duration;
2) **1920x1080** resolution, **24 fps**
(Full HD 1080p or even better format)

At the same time, it is necessary to train intermediate versions of the model with a lower spatial resolution of 1280x720 (HD or 720p) and a shorter duration of 10 seconds per video generation. This will demonstrate the ability to scale the developed architecture.

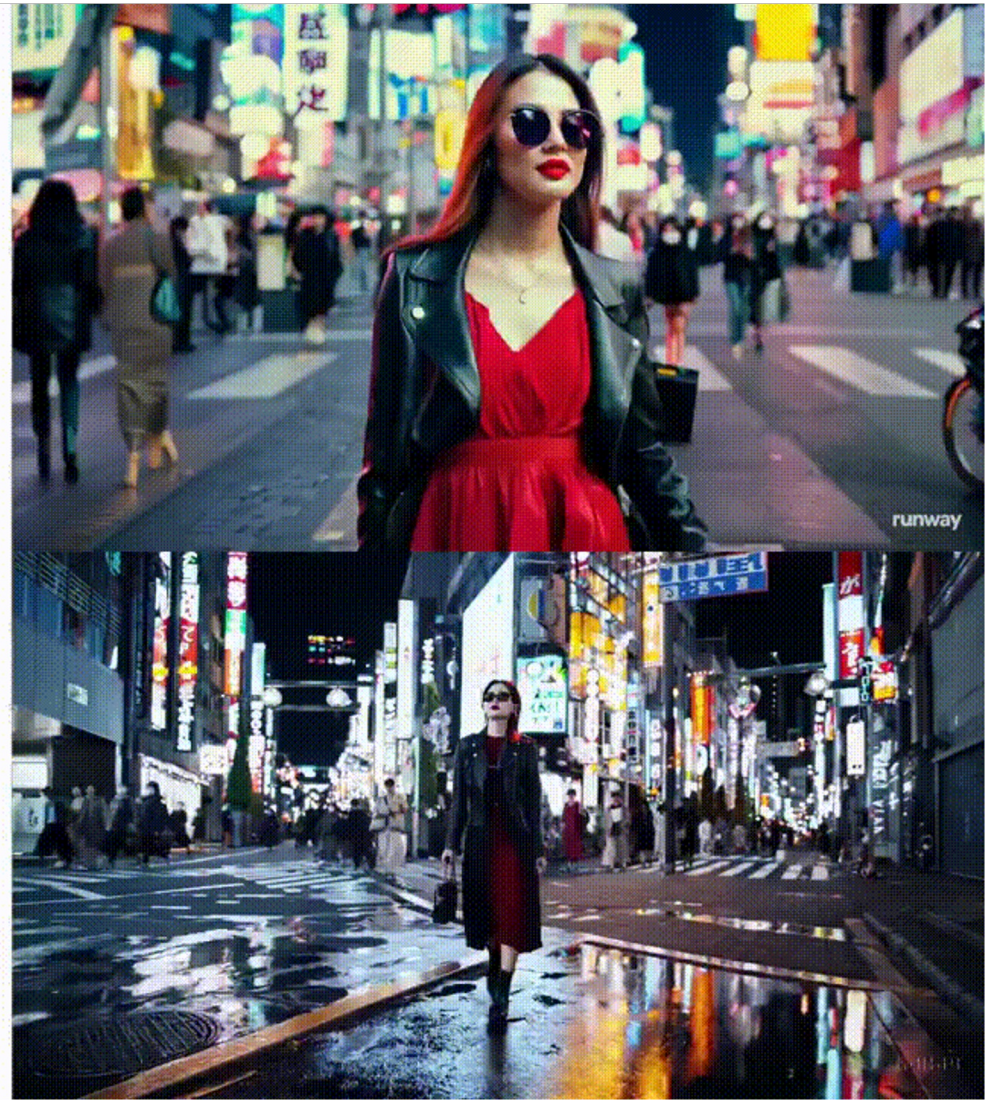
The final model should support the ability to generate video by text, by the initial frame, by the initial video fragment, as well as the ability to supplement an arbitrary "incomplete" video fragment (including solving the problem of merging two videos). In addition, this model must adapt to the tasks of instructional image and video editing, video-inpainting, as well as to any other reasonable additional conditions (ControlNet, IP Adapter).

runway

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

@ProperPrompter

Sora



KANDINSKY

VGM development

In the field of video generation (VGM), there is competition similar so that of LLM

On February 15, 2024, Open AI announced its video generation model called Sora, which significantly surpasses all previous models. Since then, more than 5 companies have presented their video generation models of comparable quality

Company	Model	Country	Status	Date	Computing resources	Data (presumably)	Architecture	Output
Open AI	Sora	USA	Preview	15.02.2024	Cluster ~2 000—10 000 H100	**Hundreds of millions** of pairs of «text-video», «text-image», as well as just images and videos	DiT	1080p video up to 60 seconds
ShengShu	Vidu	China	Preview	27.04.2024			U-ViT	1080p video up to 16 seconds
DeepMind, Google	Veo	USA	Preview	14.05.2024			DiT	1080p video up to 60 seconds
Luma Labs	Dream Machine	USA	Open demo	12.06.2024			DiT (?)	1080p video up to 120 seconds
Runway AI	Gen-3	USA	Open demo	17.06.2024			DiT (?)	1080p video up to 10 seconds
Kuaishou Technology	Kling	China	Open demo	24.07.2024			DiT (?)	1080p video up to 120 seconds
…								
MovieGen	**Meta***	**USA**	**Preview**	**04.10.2024**	**6144 H100**	**Around 400 millons**	**DiT**	**768p up to 16 seconds (16 fps)**

KANDINSKY

Difficulties of the video generation task

- New field without established architectures
- No high-quality Open Source solutions
- The computational complexity is higher than in the LLM field

	Large Language Models (LLM)	Video Generation Models (VGM, VPT, GWM)
Start of intensive model development	2017	2023
Established architectures that are explicitly described (or coded) and have been verified to scale	GPT, BERT, T5 etc	— (most likely it would be a special version of DiT)
Open Source	**LLaMA** (Meta), **Mistral** (Mistral AI), **Gemma** (Google, DeepMind), **Grok** (X.ai) etc	**Open-Sora**, **Open-Sora-Plan** (community enthusiasts) **CogVideoX** (Zhipu AI)
Model context length	**128k** (GigaChat)	**864k** (for 10 sec video with 720p resolution)

VGM development

SORA

Промпт: Reflections in the window of a train traveling through the Tokyo suburbs.

KLING

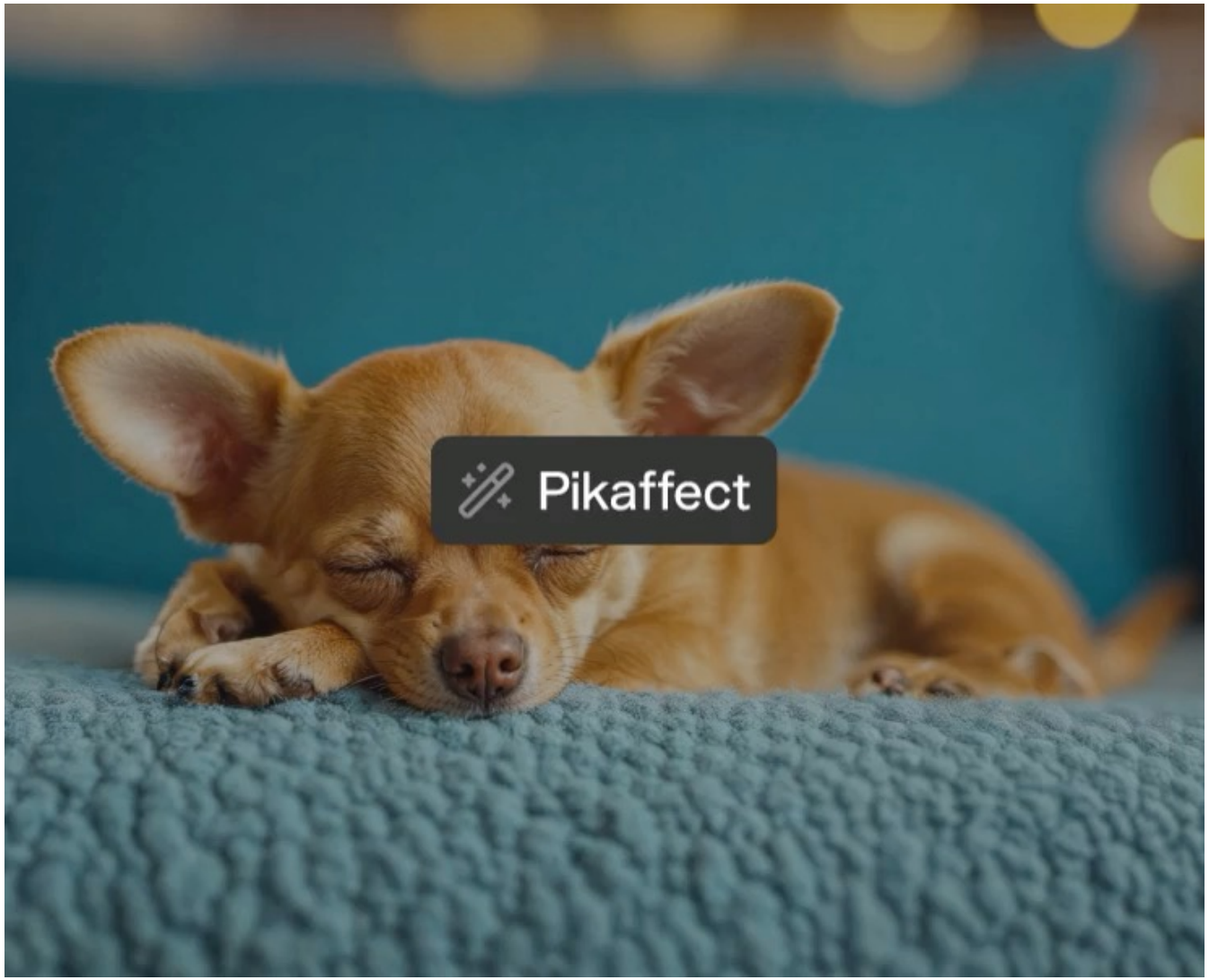
SORA

KANDINSKY

Prompt: Drone view of waves crashing against the rugged cliffs along Big Sur's Garay Point beach. The crashing blue waters create white-tipped waves, while the golden light of the setting sun illuminates the rocky shore. A small island with a lighthouse sits in the distance, and green shrubbery covers the cliff's edge. The steep drop from the road down to the beach is a dramatic feat, with the cliff's edges jutting out over the sea. This is a view that captures the raw beauty of the coast and the rugged landscape of the Pacific Coast Highway.

Source: OpenAI, Sora

VGM development



Pika 1.5 (Pikaffects)

But it's not perfect yet....

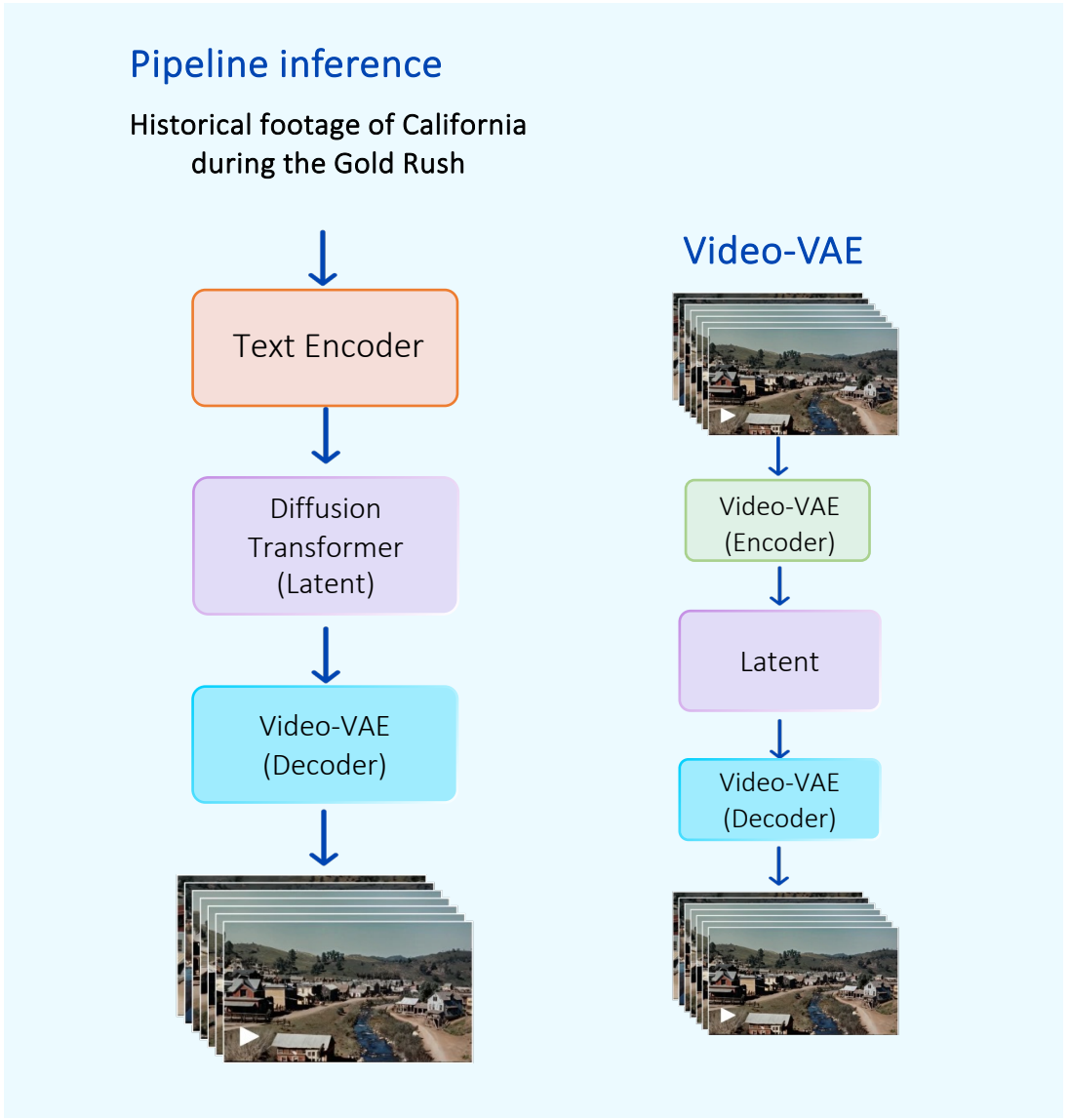
We're sharing our research progress early to start working with and getting feedback from people outside of OpenAI and to give the public a sense of what AI capabilities are on the horizon.

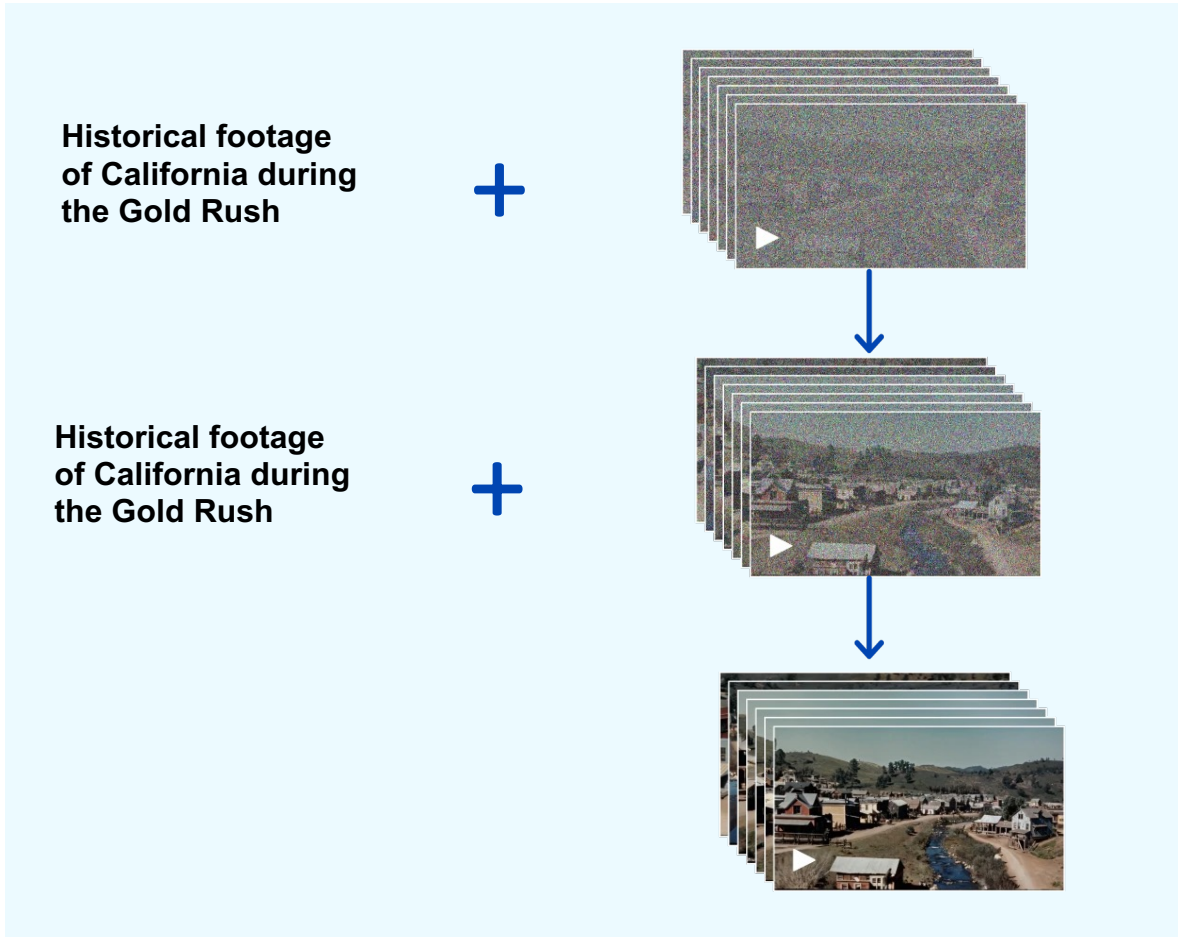
Промпт: Five gray wolf pups frolicking and chasing each other around a remote gravel road, surrounded by grass. The pups run and leap, chasing each other, and nipping at each other, playing. Weakness: Animals or people can spontaneously appear, especially in scenes containing many entities.

https://openai.com/index/sora/

https://openai.com/index/video-generation-models-as-world-simulators/

General inference scheme of diffusion text-to-video models

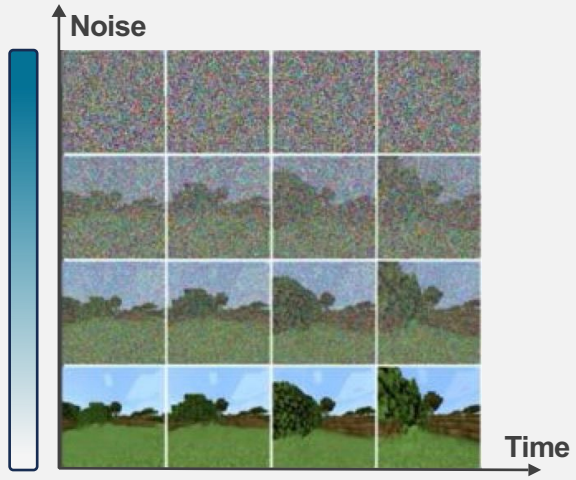




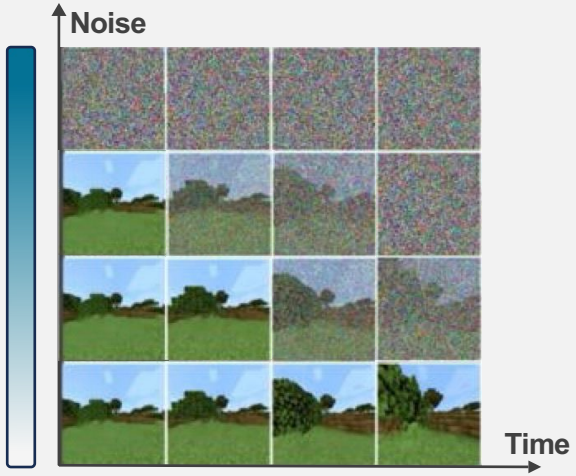
KANDINSKY

Training approaches

Current algorithm:



Autoregressive algorithm:



Different degrees of noise level	Don't make some patches noisy	Completely noising some patches
		

KANDINSKY

Loss functions

Problem

Now diffusion is taught with MSE Loss, which penalizes the image pixel by pixel and does not focus on important details.

Diffusion Model with **MSE Loss** (No CFG)

Diffusion Model with **Self-Perceptual Loss** (No CFG)

The solution is to use more advanced loss functions:

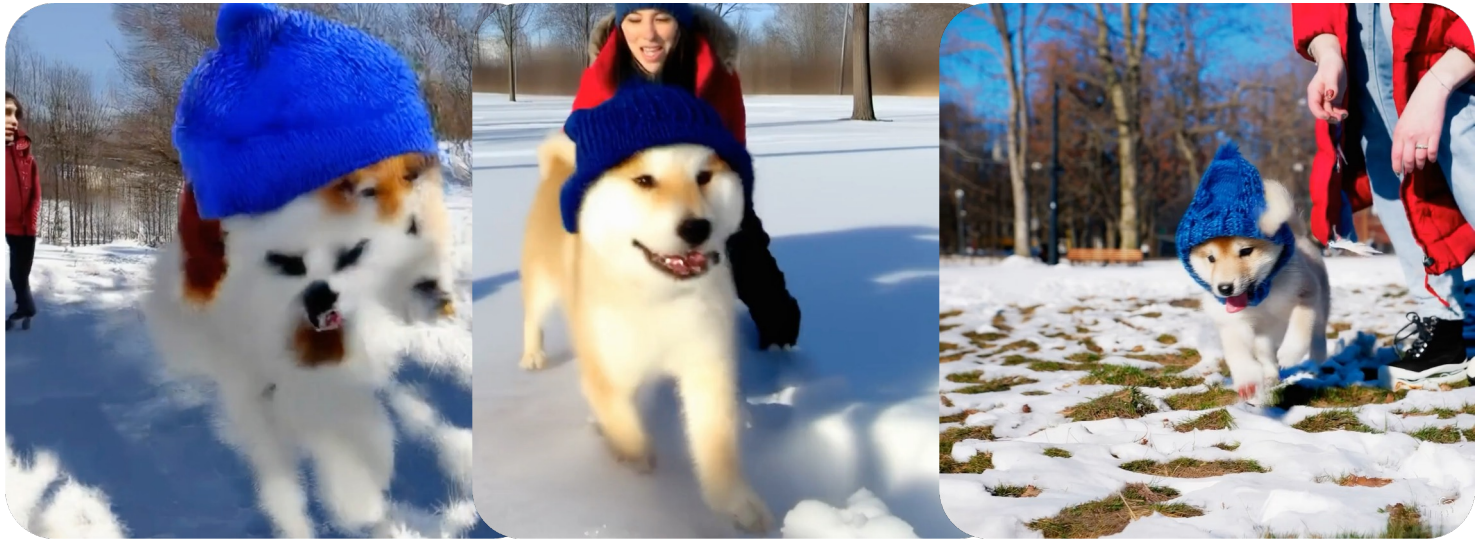
- → Perceptual loss (there are studies on images)
- → Adversarial loss (there is an insight from Kandinsky 3.0 Flash – the discriminator allows to increase the detail)
- → Physics – inspired loss (still at the hypothesis level)

Kandinsky 3.0

Kandinsky 3.1 Flash

KANDINSKY

Compute (Sora)



Base compute | 4x compute | 32x compute

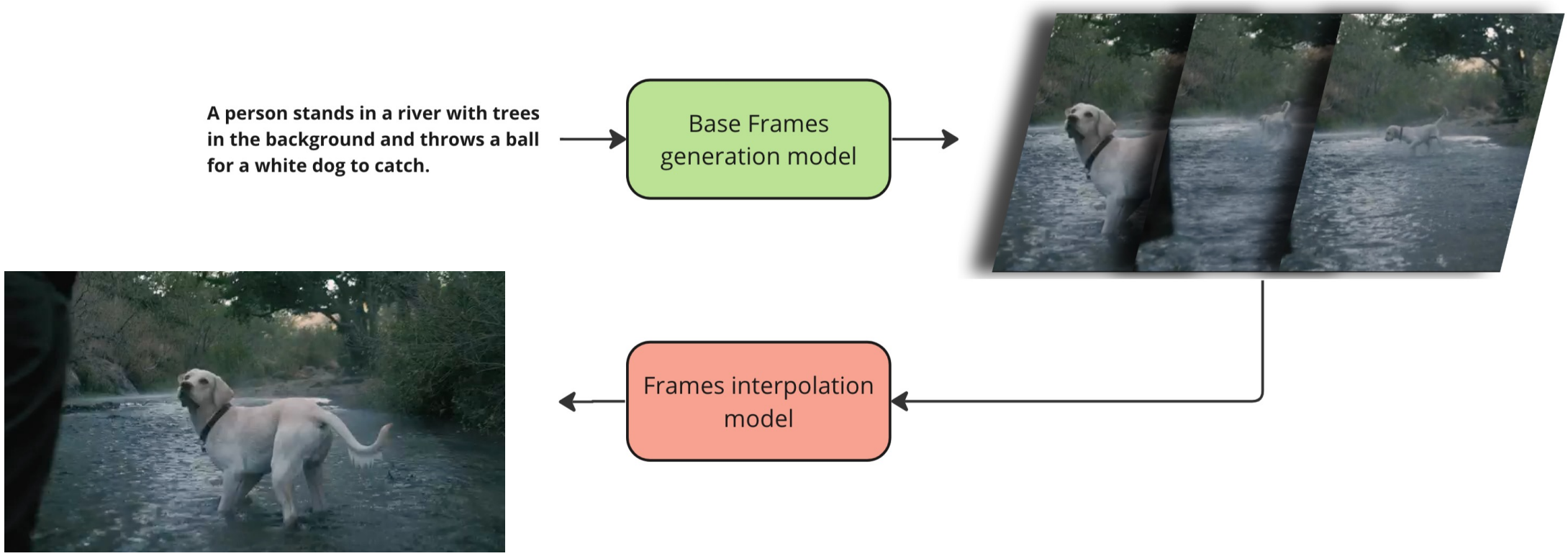
Same seed(!)
Same data for training (!!)

Kandinsky 4.0 XL

If we don't have enough resources to train one big model, we can complicate the pipeline a little bit, but simplify each part of this pipeline

Kandinsky Video 1.0/1.1 pipeline

Base Frames generation model
+ Frames interpolation model

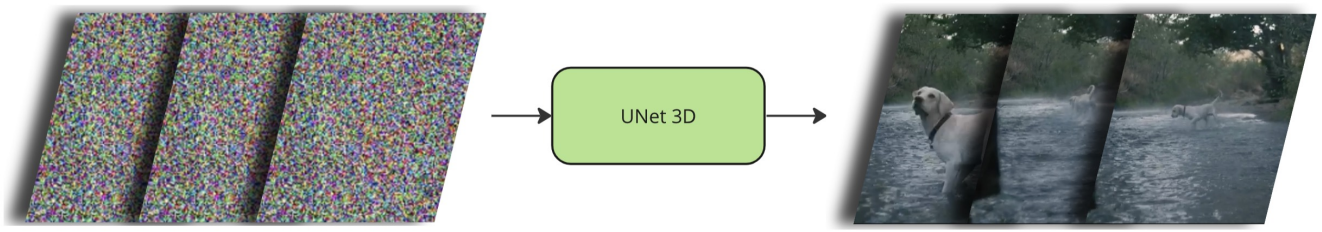


KANDINSKY

Kandinsky Video 1.0

Base Frames:

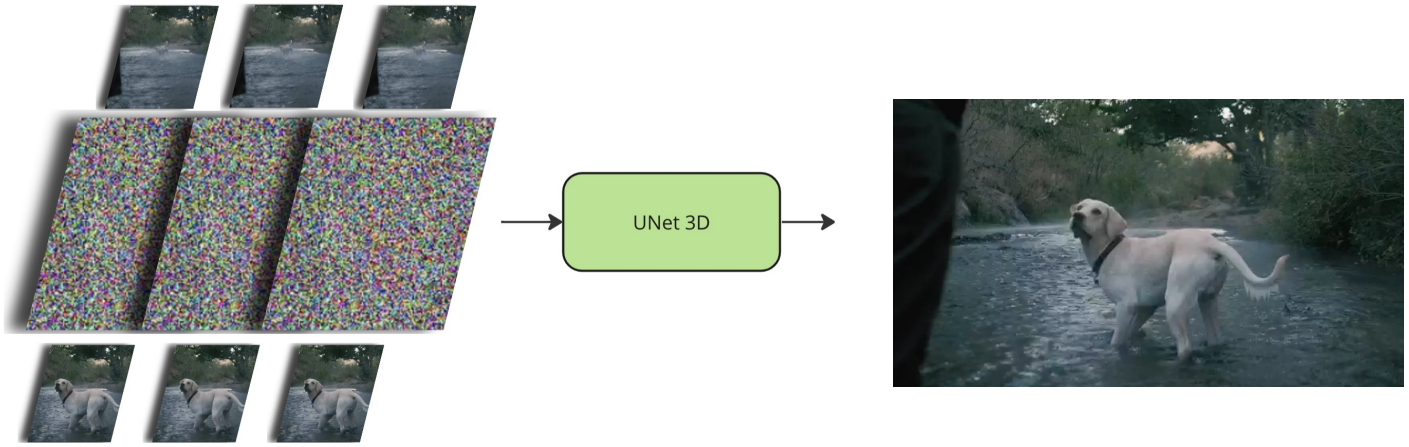
Shape: (bs, t, c, w, h)



Interpolation:

(base frames are concatenated with noise)

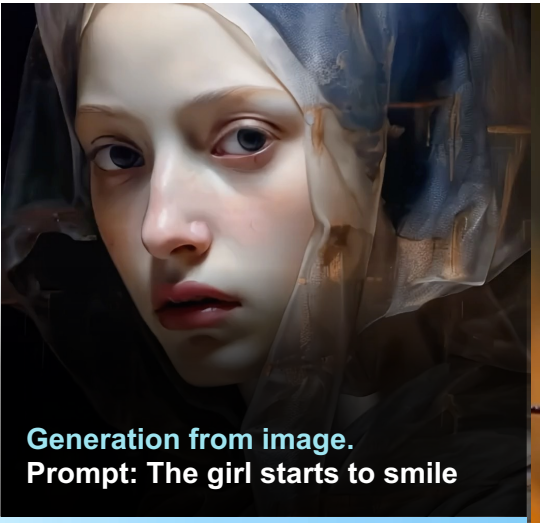
- Shape: (bs, t+2, c, w, h)
- Shape: (bs, t, c*3, w, h)



KANDINSKY

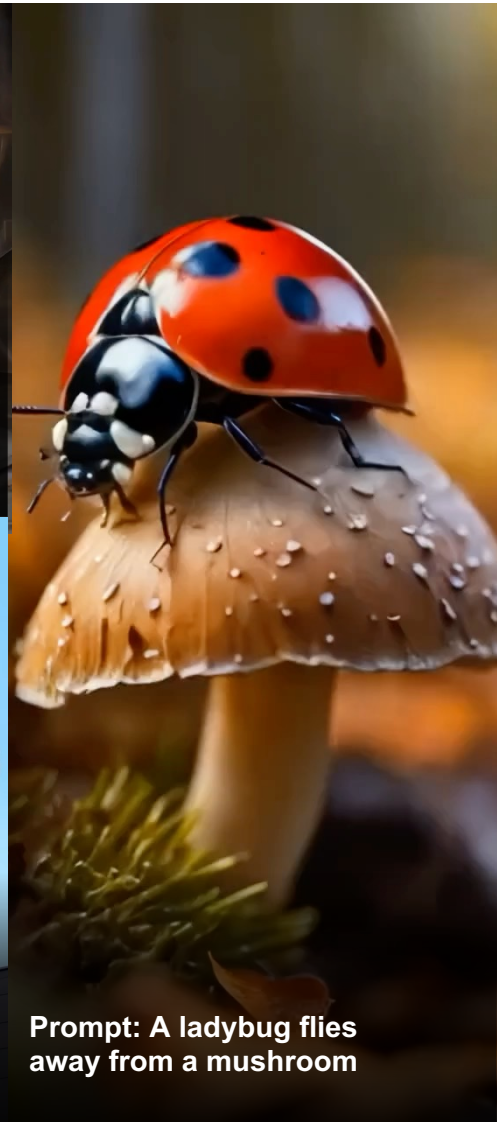
Kandinsky Video 1.1

17.5B parameters

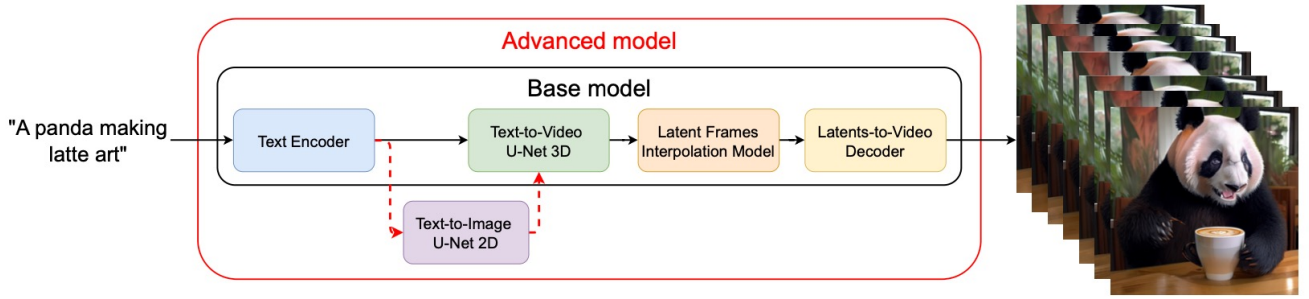


Generation from image.
Prompt: The girl starts to smile

Prompt: Large white octopus sitting on top of a building



Prompt: A ladybug flies away from a mushroom



https://evalcrafter.github.io

KANDINSKY

Prompt: Red Car Drift

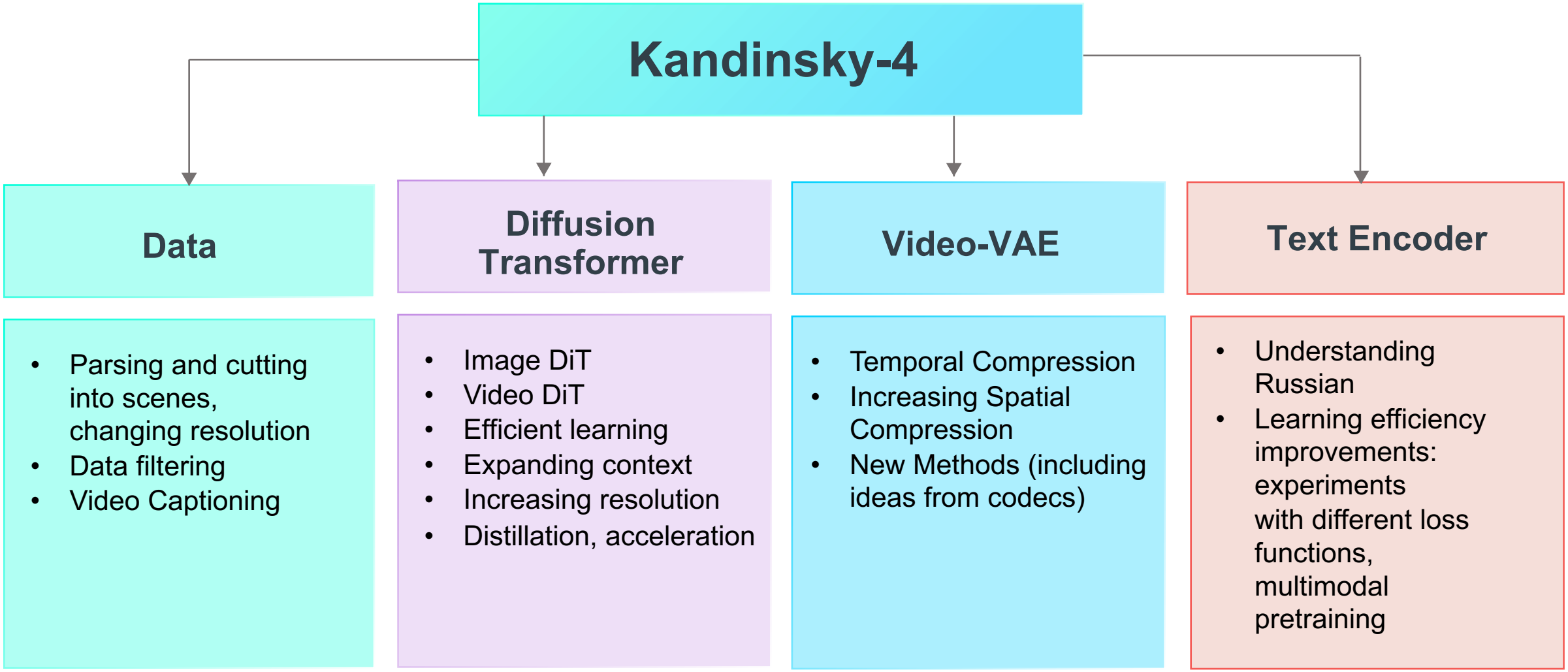
Prompt: The Russian flag is waving in the wind

Generation from image.
Prompt: Prompta: Cartoon

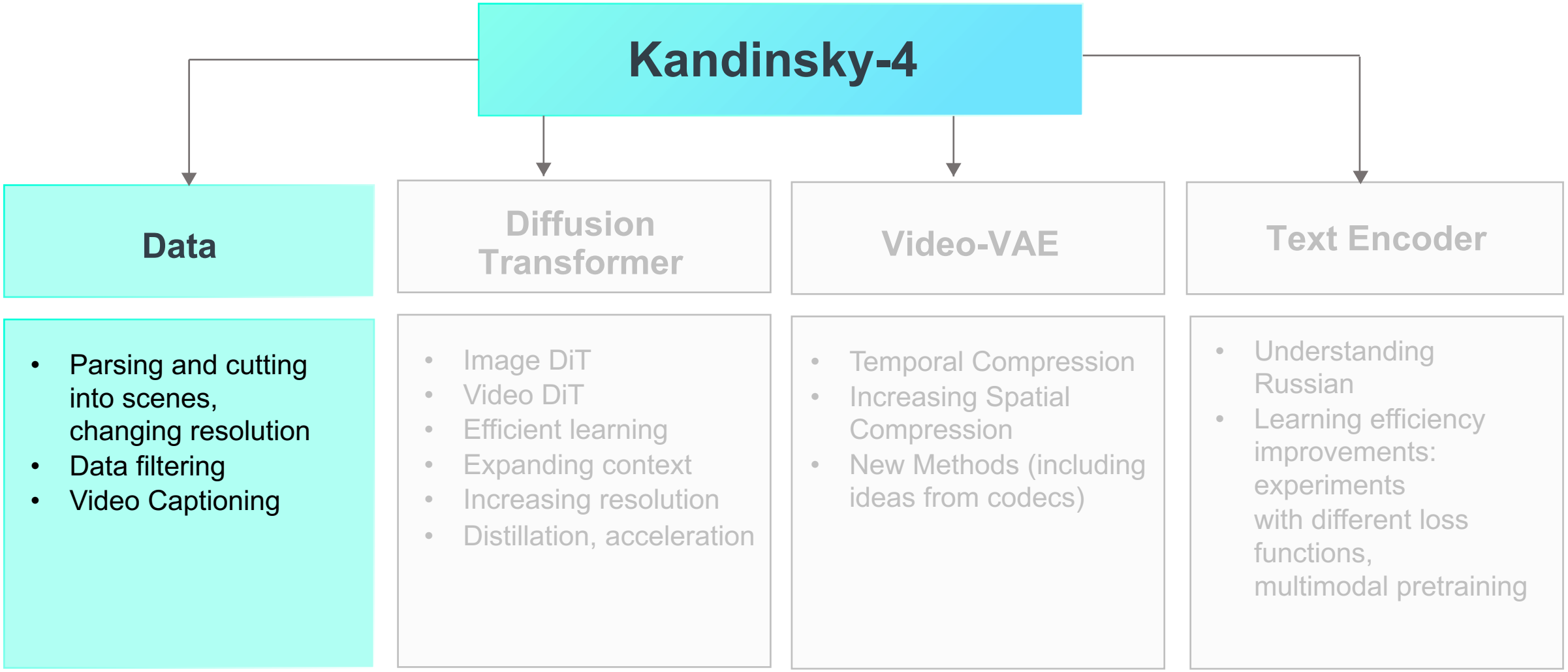
Prompt: Cheburashka makes coffee in a cafe

But if you have the enough GPU (computing resources), a single model shows the best quality!

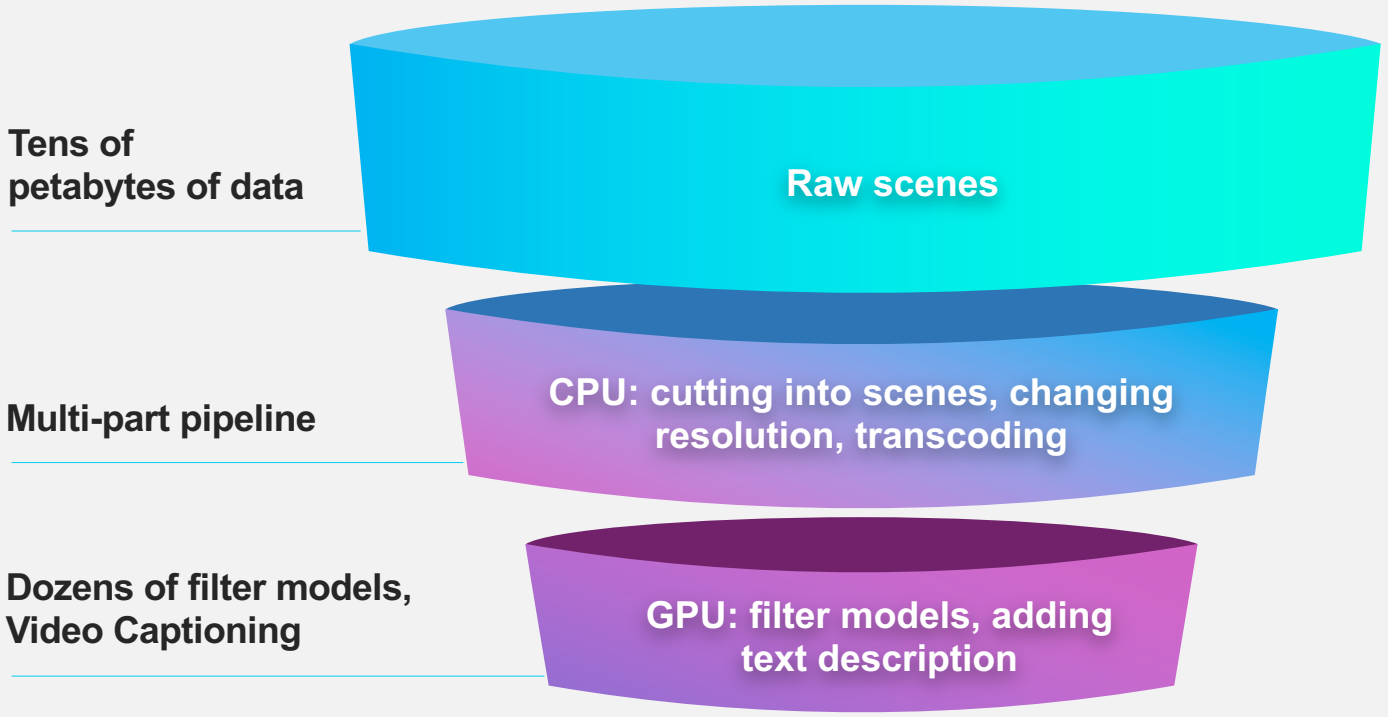
Main directions of development



Data


Kandinsky-4
Data
Diffusion Transformer
Video-VAE
Text Encoder
• Parsing and cutting into scenes, changing resolution
• Data filtering
• Video Captioning
• Image DiT
• Video DiT
• Efficient learning
• Expanding context
• Increasing resolution
• Distillation, acceleration
• Temporal Compression
• Increasing Spatial Compression
• New Methods (including ideas from codecs)
• Understanding Russian
• Learning efficiency improvements: experiments with different loss functions, multimodal pretraining

Data



«Raw»:
~400,000 hours
~150 million «text-video» pairs

After processing:
~170,000 hours
~60 million «text-video» pairs

Approximately 40% of the data is included in the training

KANDINSKY

Pipeline

«Raw» videos table

dataset_name	obs_endpoint	obs_path	status	process_start	width	height	duration	fps	id	process_node
youtube	obs.ru-moscow-1.hc.sbercloud.ru	gigaeye-data/youtube/2024-02-1/1.mp4	process	2024-07-12 11:15:32.847311 +03	1920	1080	00:01:11	24	1	job-3cc09c65-a72d-5b6e05155

10^9 rows

«Processed» videos table

id	raw_id	obs_endpoint	text prompt	width	height	duration	dynamicity	aesthetic_score
10	1	gigaeye-kandinsky-spark/db_pipeline/scenes/1/0:01:16.mp4	The scene gives off a casual and relaxed atmosphere, with the woman seemingly enjoying….	1920	1080	12	0.976	0.812
11	1	gigaeye-kandinsky-spark/db_pipeline/scenes/1/0:01:42.mp4	The scene is in black and white, giving it a classic feel…	1920	1080	13	0.312	0.115

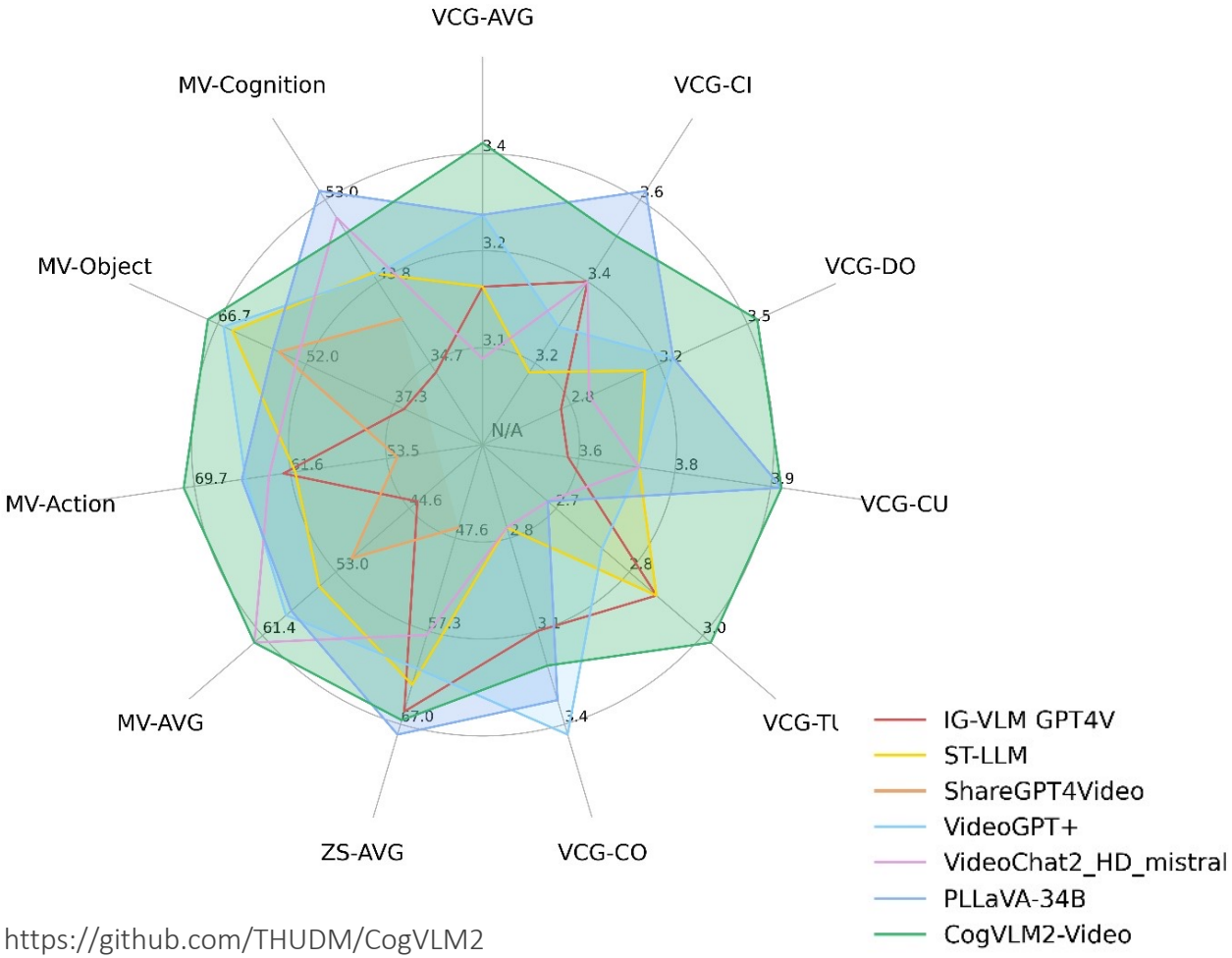
10^8 rows

Optical flow

KANDINSKY

Video captioning

CogVLM2-Video



https://github.com/THUDM/CogVLM2

Есть проблемы

20% descriptions are inaccurate

0,5% descriptions are incorrect

Main pipeline bottleneck (2.2 svsc)

A group of individuals engaging in various outdoor activities on a paved pathway surrounded by lush greenery. The scenes depict people walking, riding bicycles, and skateboarding, with some individuals in red and white attire. The pathway is well-maintained, with long shadows cast by the trees, suggesting it's either early morning or late afternoon.

KANDINSKY

Comparison with captions from LAION

Revolights-eclipse

(600, 329)

Grosjean's car was ripped onto two pieces after careering through the guardrails

(768, 512)

(768, 512)

Shooting stars elec ireland

(640, 360)

(640, 320)

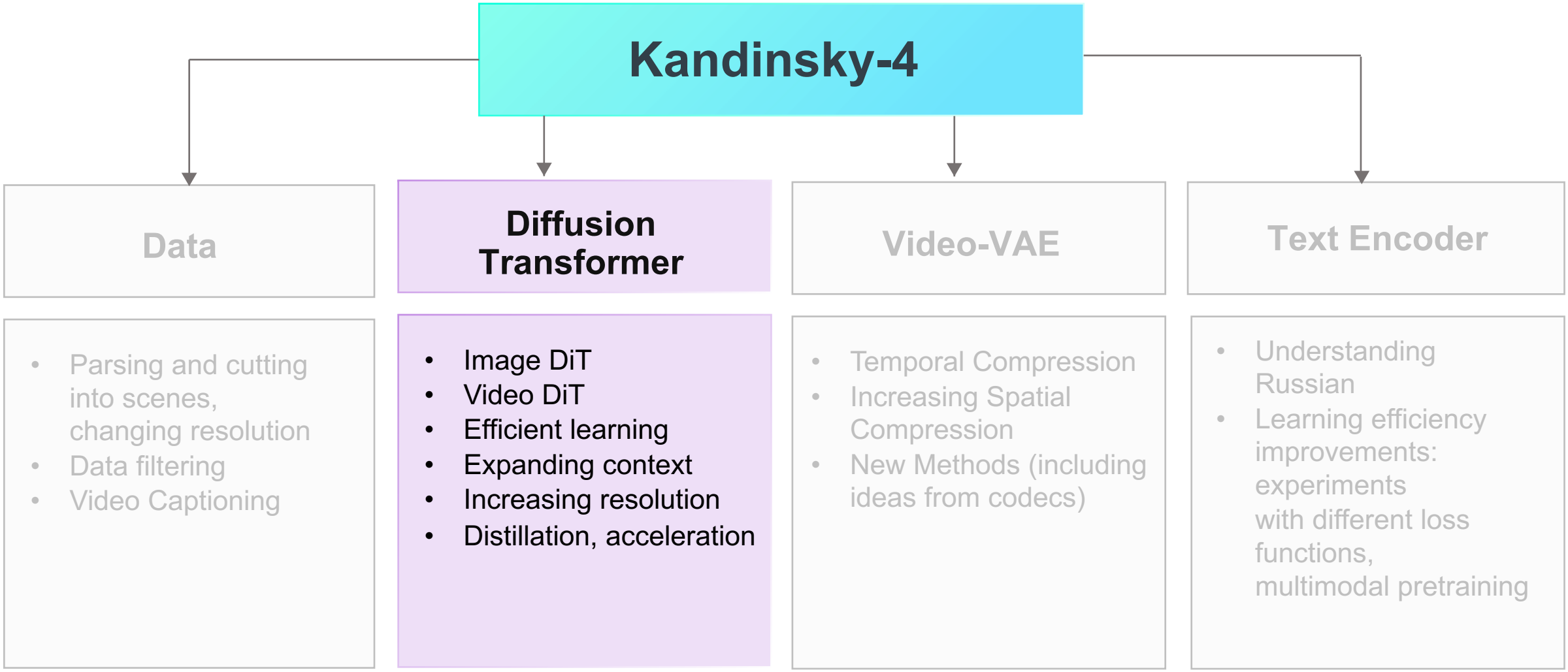
2013 DACIA DUSTER 1.5DCI AMBIANCE 4X2 ****OWN THIS CAR FROM £33 PER WEEK

(682, 512)

(640, 512)

KANDINSKY

Main part: Diffusion Transformer



Working with a diffusion transformer

01

Pretrain type and training stability

02

Joint learning on images and videos

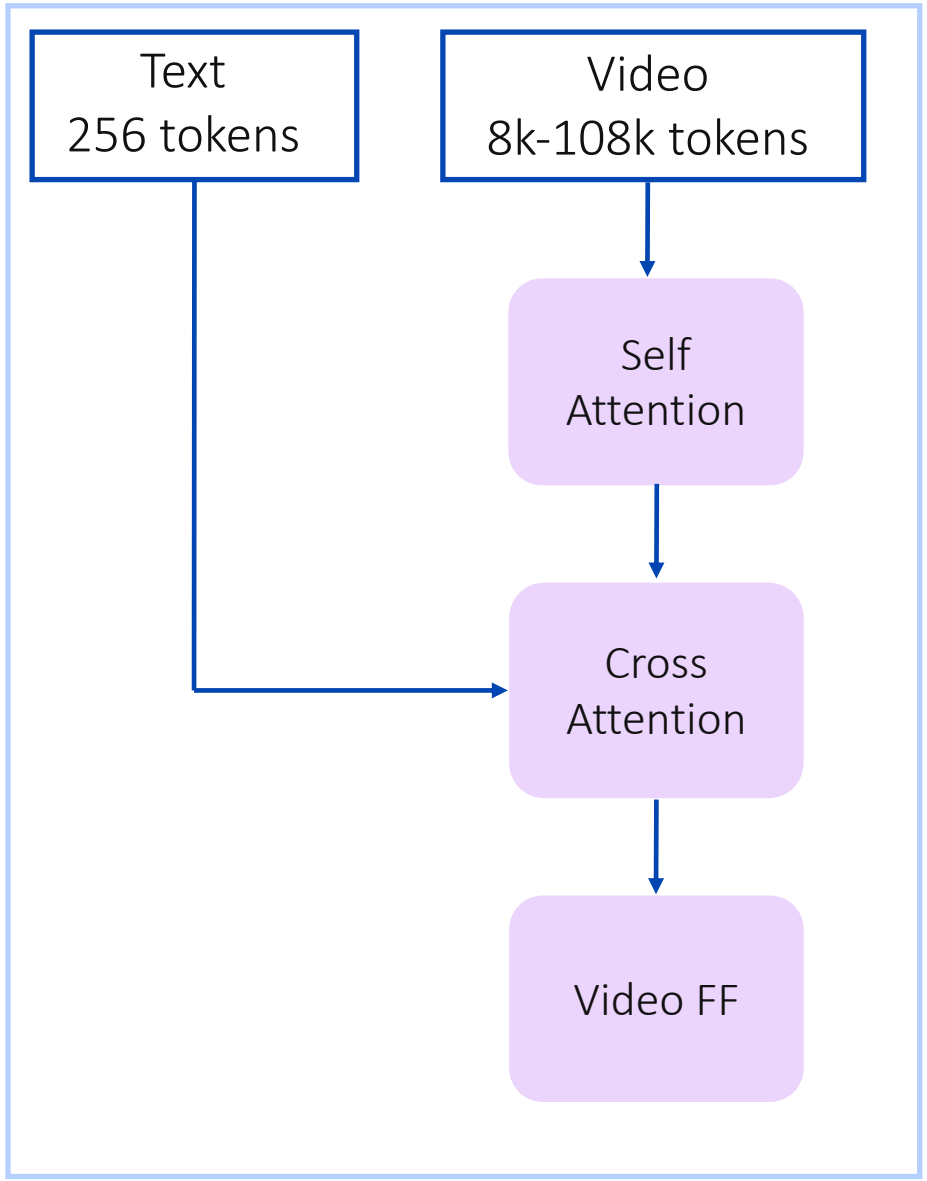
03

Long videos

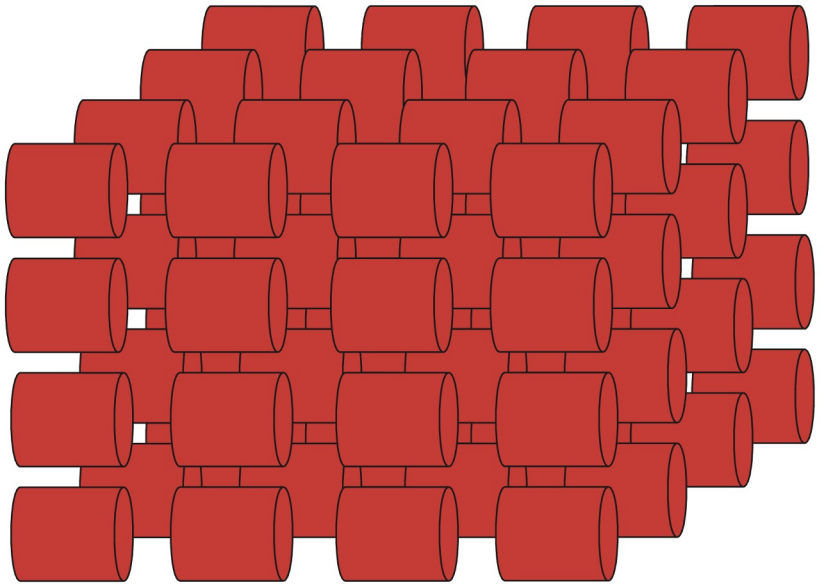
04

High resolution

DiT: Baseline

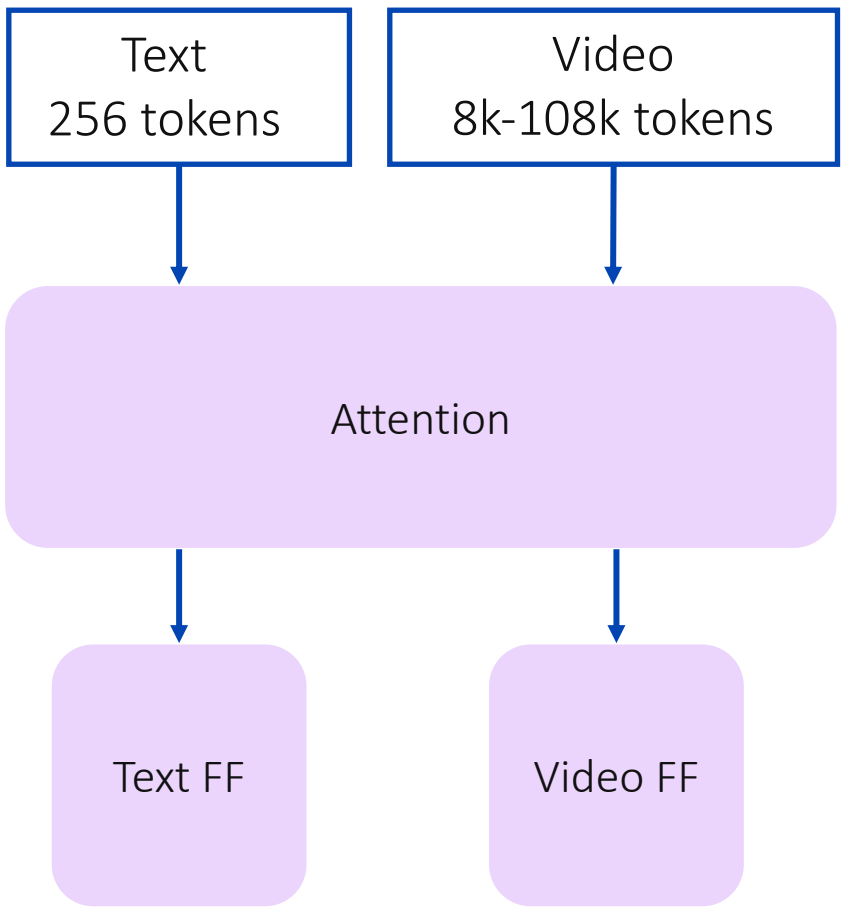


Full Attention:

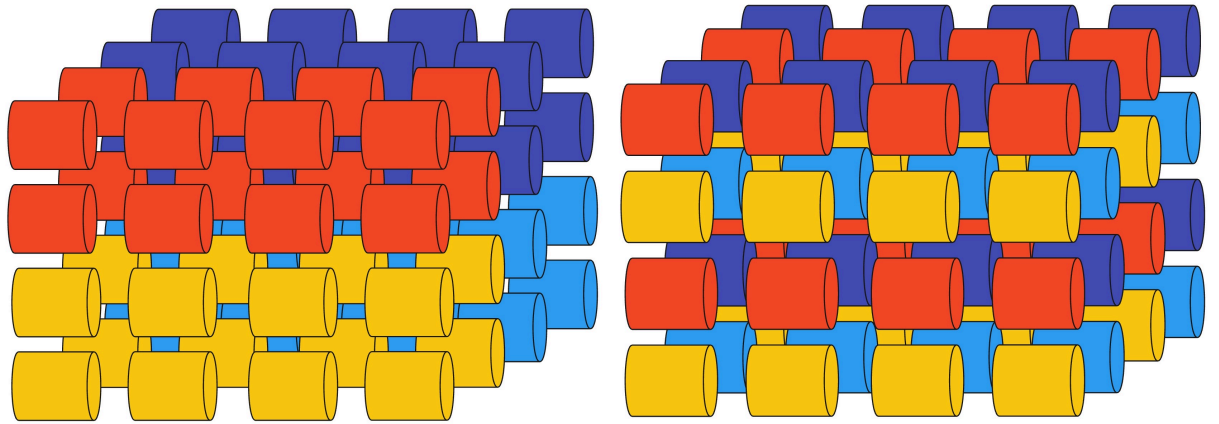


KANDINSKY

MMDiT, Sparse Attention



Sparse Attention for resolutions > 256:

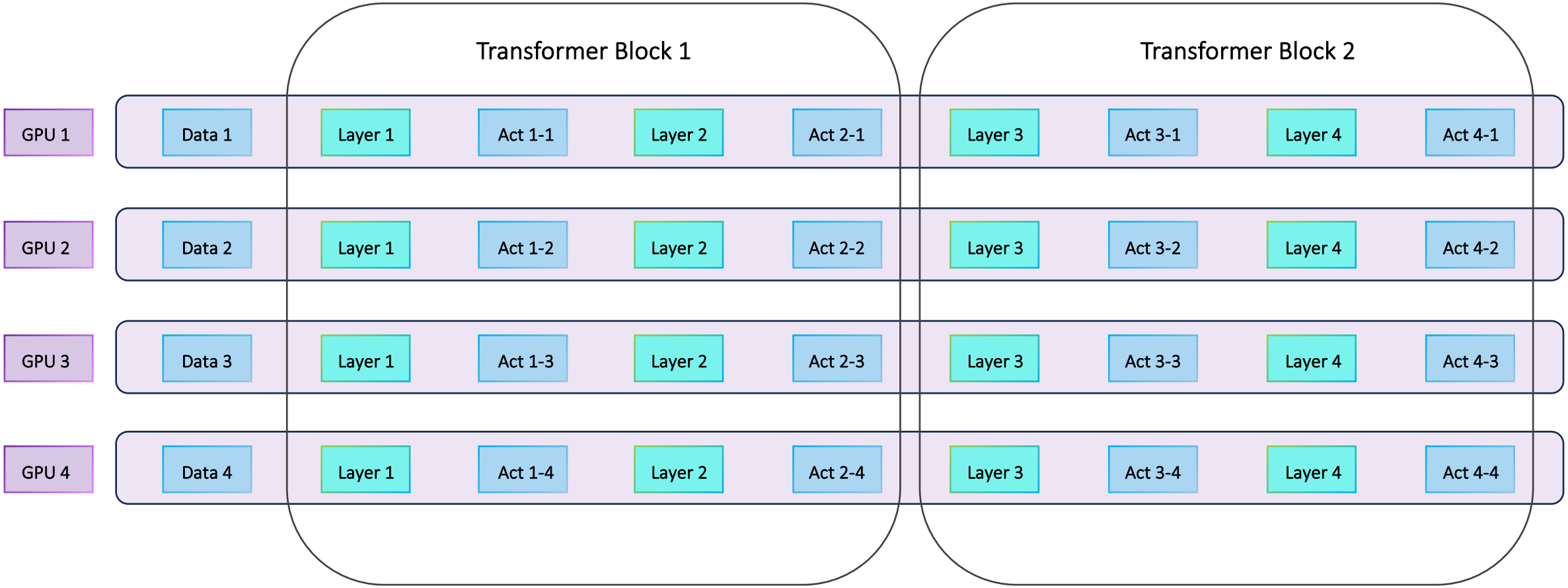


- Learning speedup **x4 times**
- Inference speedup **x6 times**
- Memory consumption decrease **x1.3 times**
- **x1.5** more parameters

KANDINSKY

Training Pipeline: Baseline

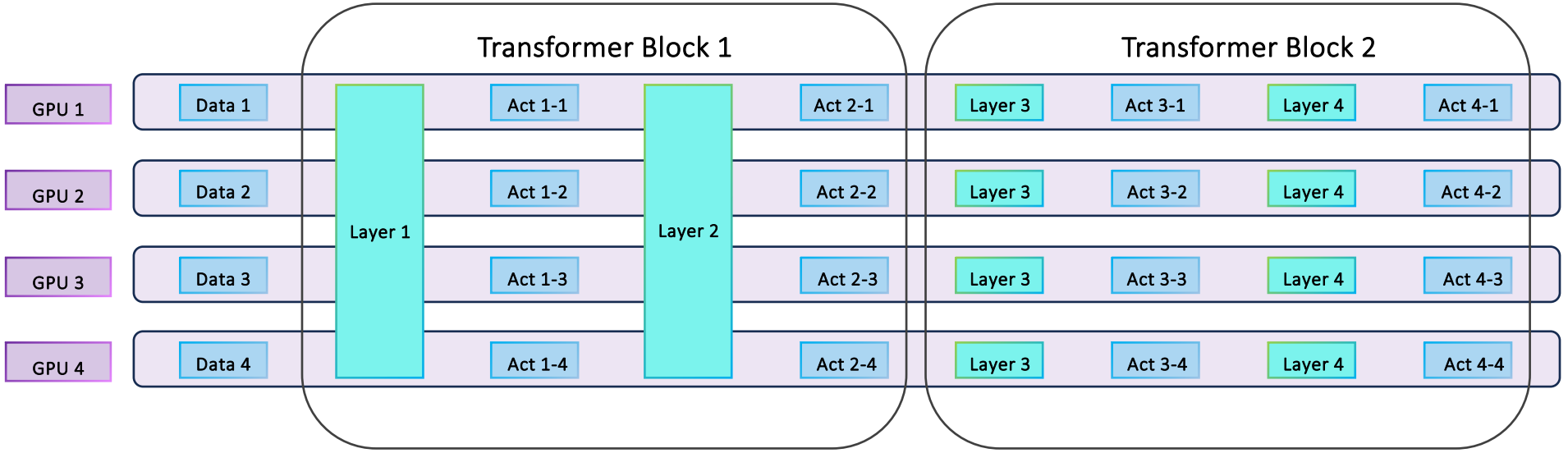
The main factors of memory consumption: model weights, optimizer state, activations, attention



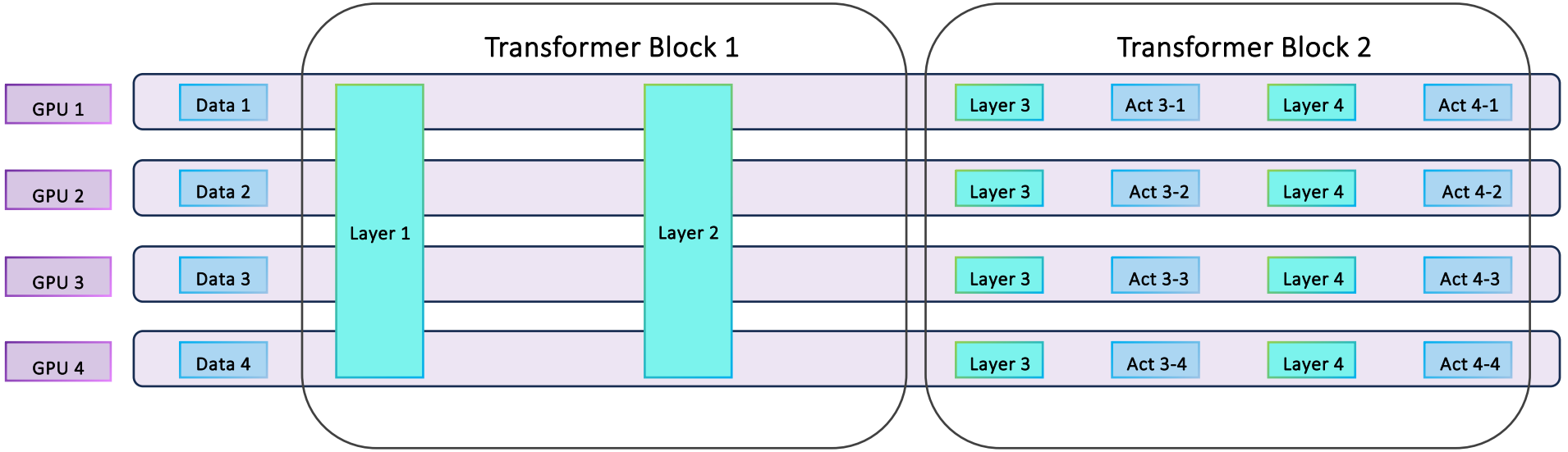
KANDINSKY

Training Pipeline: Engineering optimizations (FSDP)

FSDP:



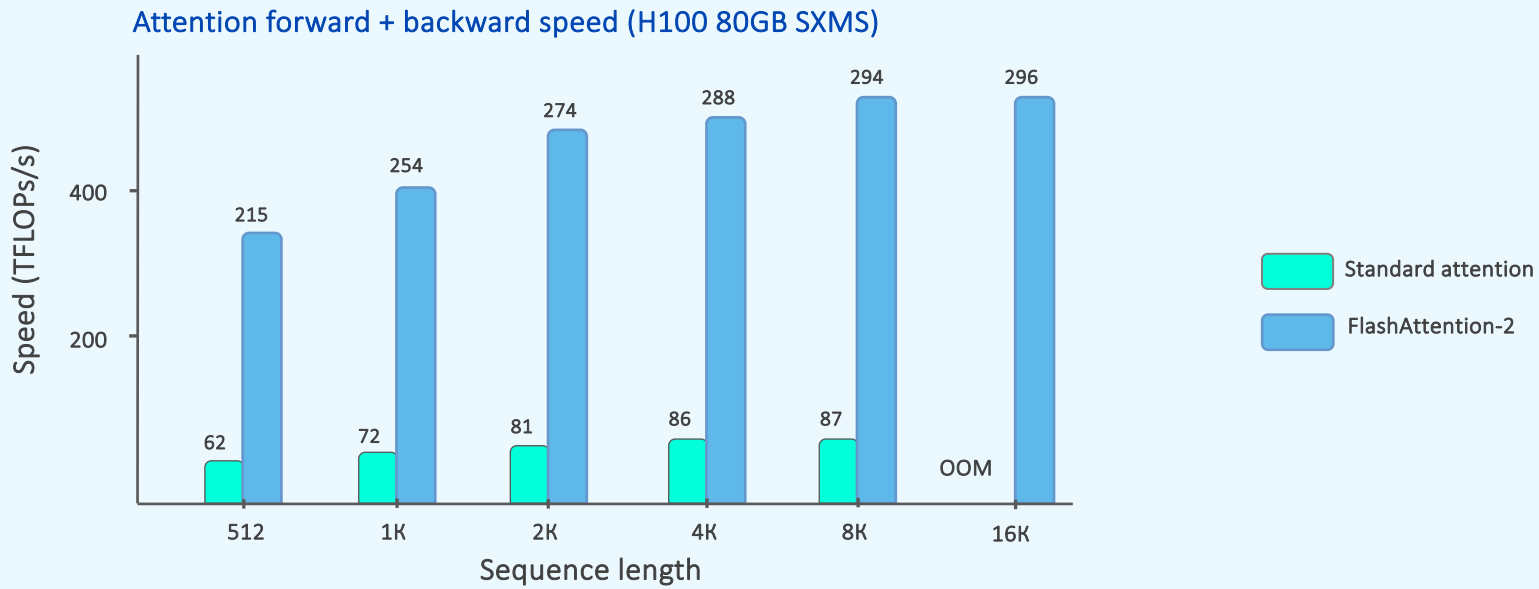
FSDP + Activation Checkpointing:



KANDINSKY

Training Pipeline: Engineering optimizations (Flash Attention)

Flash Attention:

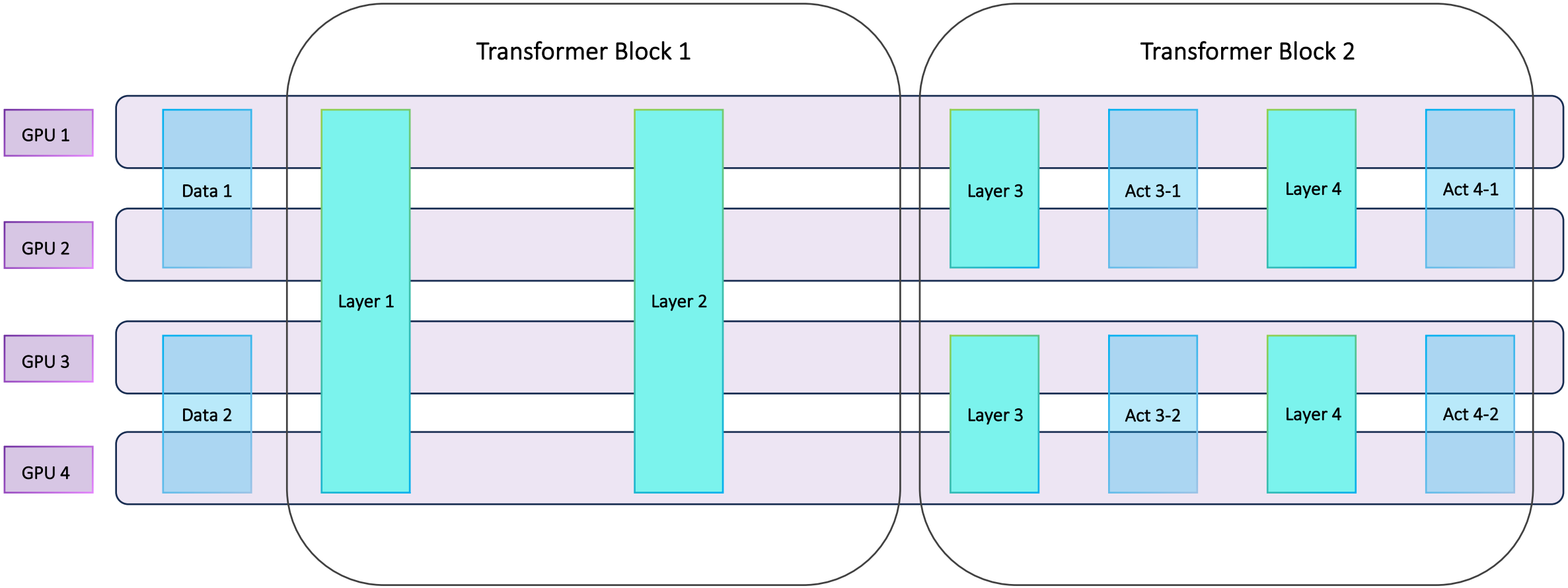


Memory consumption decrease more than 15 times:

Strategy 96 GPU, 3B DiT	DDP	DDP + Flash Attn	FSDP+ Flash Attn	FSDP + Checkpointing + Flash Attn 2
Memory 256x256, 2.5s, 24fps	OOM	76 Gb	20 Gb	5 Gb
Step Speed 256x256, 2.5s, 24fps	inf	0.6 s	0.4 s	0.5 s

KANDINSKY

Training Pipeline: Engineering optimizations (Ring Flash Attention)



KANDINSKY

MMDiT: training on videos

Number of parameters	Resolution	fps	Duration	Video Dataset	GPU
2.7B	256x256	24	<10 sec	31M 120B Tokens	64

SORA: base compute

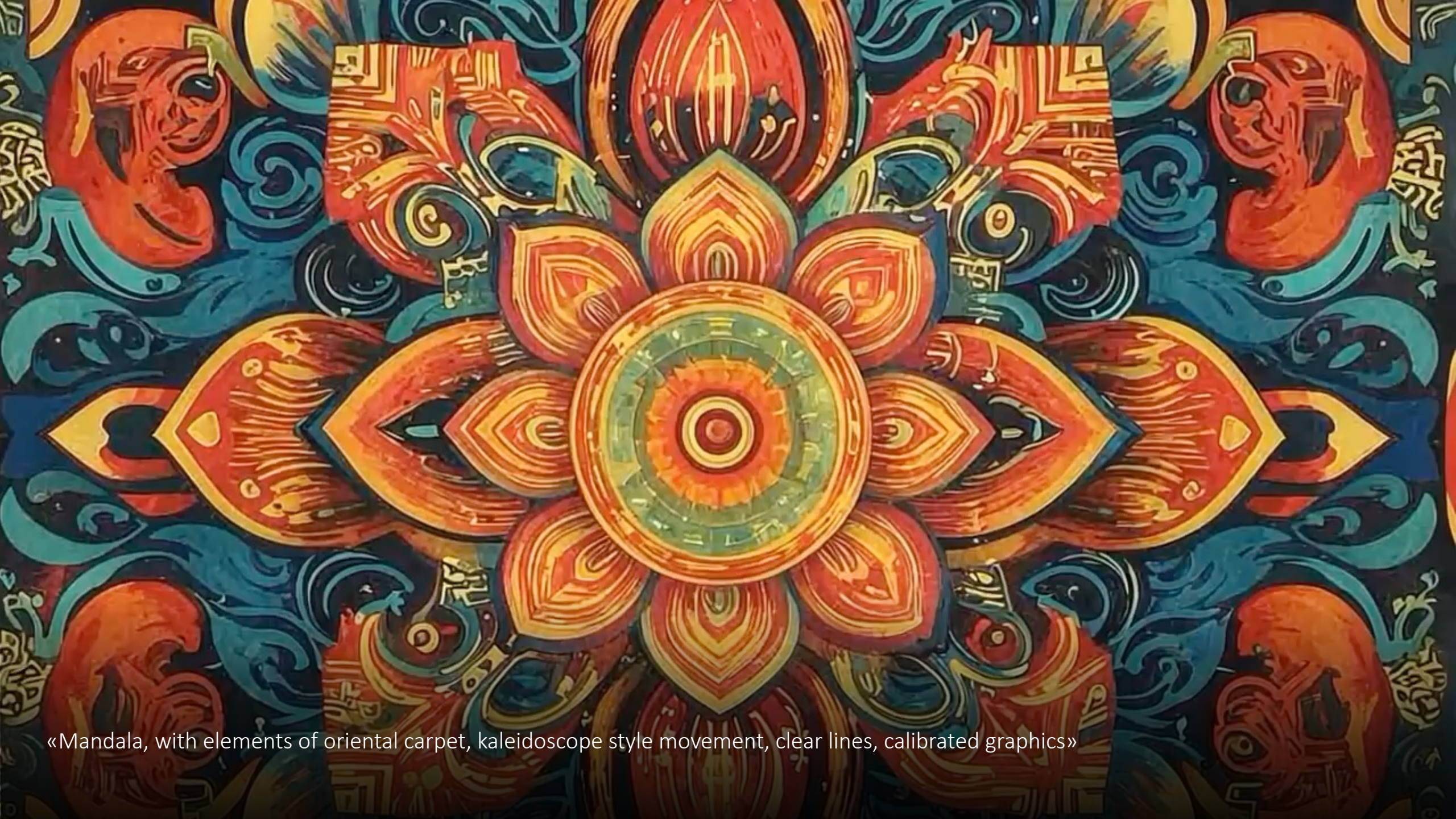
A panda, dressed in a small, red jacket and a tiny hat, sits on a wooden stool in a serene bamboo forest…

A knight riding on a horse through the countryside

Confident teddy bear surfer rides the wave in the tropics

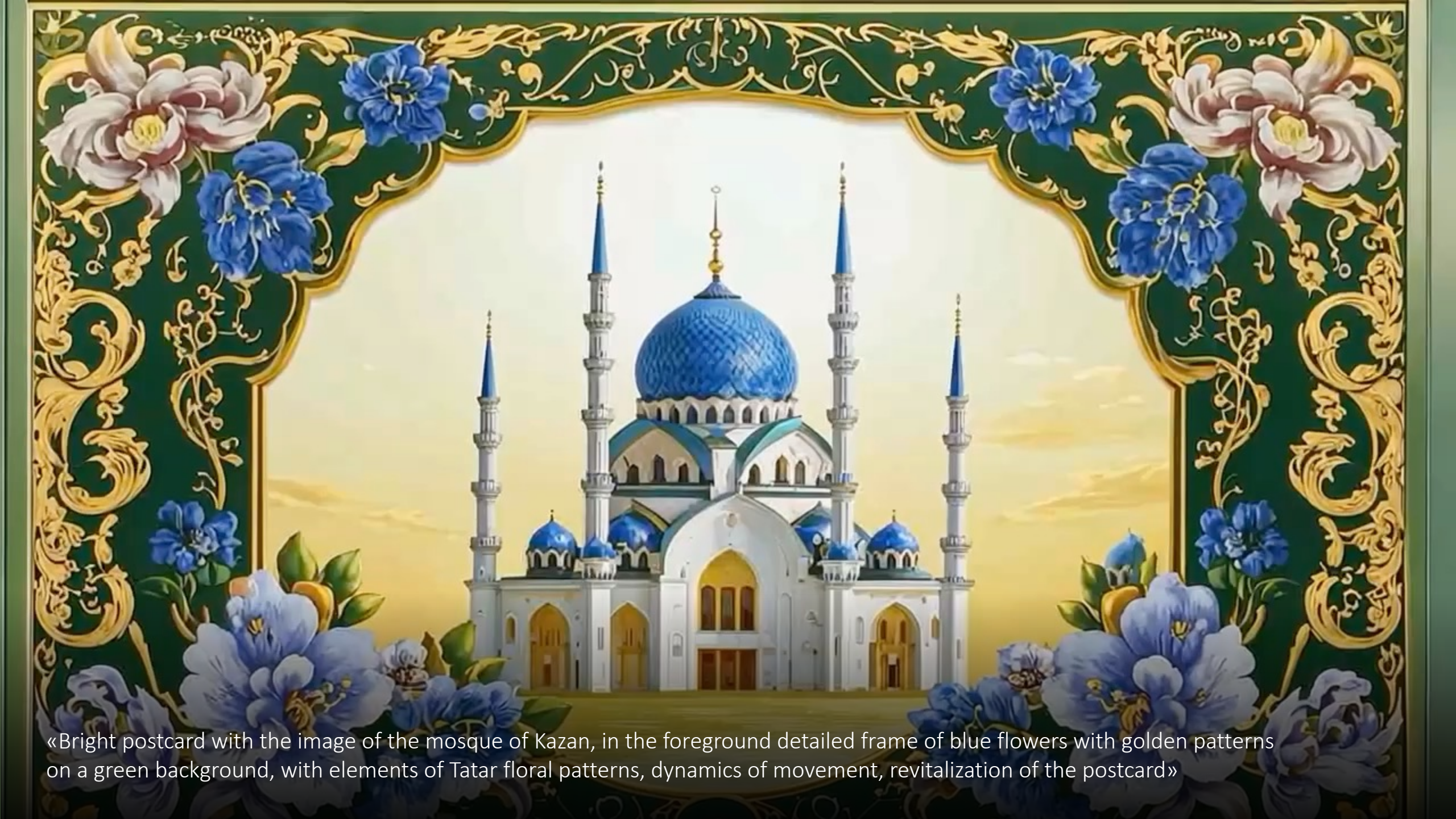
KANDINSKY

54



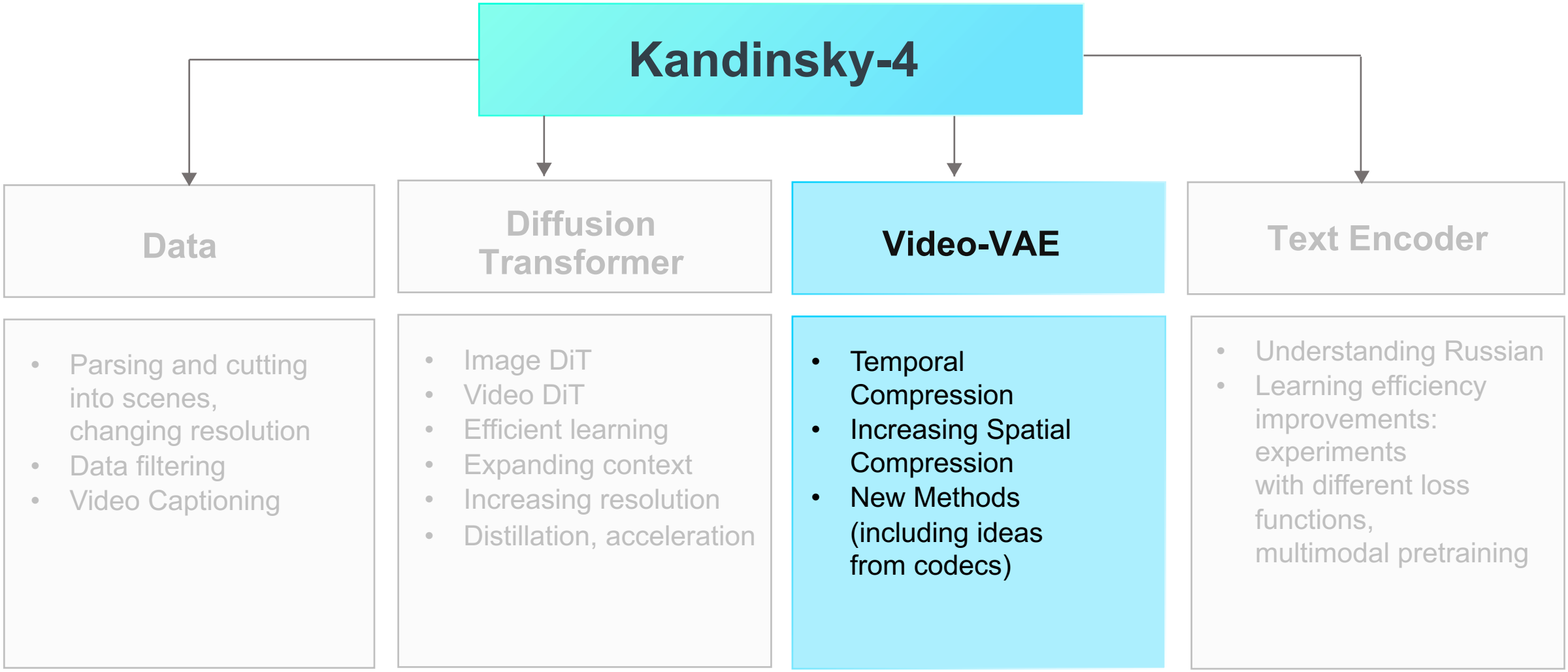
«Mandala, with elements of oriental carpet, kaleidoscope style movement, clear lines, calibrated graphics»

«Beautiful Tatar girl on the background of the Kazan Kremlin, elements of Tatar culture, Karina in art style, maximum camera zoom, facial details»



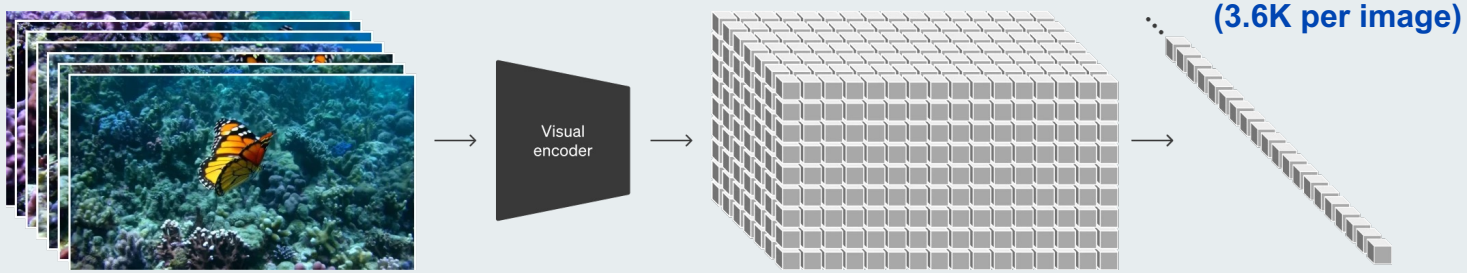
«Bright postcard with the image of the mosque of Kazan, in the foreground detailed frame of blue flowers with golden patterns on a green background, with elements of Tatar floral patterns, dynamics of movement, revitalization of the postcard»

Main directions of development



Video-VAE

High computational complexity



32K tokens

128K tokens

Open Source solutions do not have a high enough compression ratio

VAE model	Compression (TxHxW)	Number of tokens
—	Without compression	55 296k
Image VAE	1x8x8	864k
Open-Sora v1.2 Open-Sora Plan v1.2 CV-VAE CogVideoX	4x8x8	216k
Cascade Video-VAE (Ours)	**8x8x8**	**108k**
Goal	8x16x16	27k

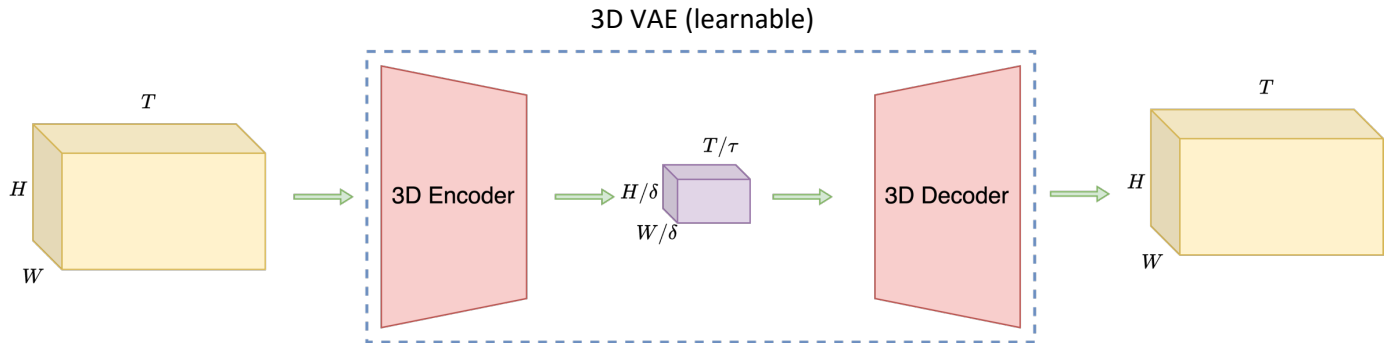
OpenSora: https://github.com/hpcaitech/Open-Sora
OpenSoraPlan: https://github.com/PKU-YuanGroup/Open-Sora-Plan
CV-VAE: https://github.com/AILab-CVC/CV-VAE

KANDINSKY

Video-VAE

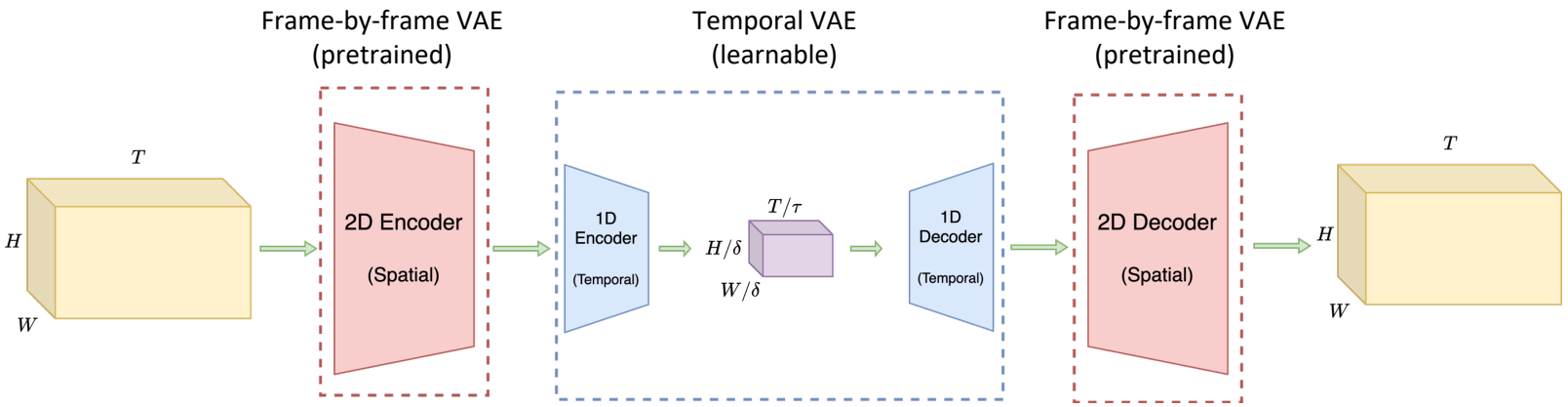
Default algorithm – 3D VAE:

- Small context - 17 frames
- Low compression ratio - 4
- Requires lots of resources - 32 GPU



Cascade VAE: 2D + 1D:

- Large context - 128 frames
- Higher compression ratio - 8
- Requires less resources - 32 GPU



KANDINSKY

Video-VAE

Original video

Reconstruction - x512 times smaller dimension!

04

Models Application & Conclusions

General World Models

runway

A world model is an AI system that builds an internal representation of an environment, and uses it to simulate future events within that environment.

We believe the next major advancement in AI will come from systems that understand the visual world and its dynamics, which is why we're starting a new long-term research effort around what we call general world models.

Sora

Sora serves as a foundation for models that can understand and simulate the real world, a capability we believe will be an important milestone for achieving AGI.
We believe the capabilities Sora has today demonstrate that continued scaling of video models is a promising path towards the development of capable simulators of the physical and digital world, and the objects, animals and people that live within them.

World Models – unlimited source of data for training Foundation Models

General World Models: симуляторы

Using video generation models to simulate our reality (or any other virtual one) and to adjust the robot's actions will be significant improvement in this area

Cons - long and computationally expensive

KANDINSKY

General World Models: downstream-tasks

A fundamental model of computer vision that can enhance other computer vision tasks

- Synthetic data generation
- Pretrain for perception tasks: depth map, detection, segmentation (example - Marigold)
- Pretrain for generative tasks: 3D object/scene generation

General World Models: visual content

- Entertainment, advertising, cinema, art
- Video generation by text query
- Image animation

KANDINSKY

CONTACTS

Denis Dimitrov

Scientific advisor, AIRI
Managing director in data science,
Sber AI

@DENDI_MATH_AI

@DENDIMITROV