



Эволюция методов диффузионной генерации видео

Константин Соболев

AIRI, Научный сотрудник

Руководитель команды Video Generative AI в составе
лаборатории Fusion Brain

Содержание

- 01 Мотивация
- 02 Введение в диффузионные модели
- 03 Первые работы по видео диффузии
- 04 Адаптация T2I моделей на T2V
- 05 Диффузионные трансформеры для видео
- 06 Sora и аналоги
- 07 Выводы и перспективы

01



Мотивация

Мотивация

→ За последние 2 года диффузионные модели модели генерации видео сделали огромный скачок в качестве

2022



HD разрешение, до 128 кадров

2023



Full HD разрешение, до 113 кадров

2024



Full HD разрешение, до 1800 кадров

Мотивация

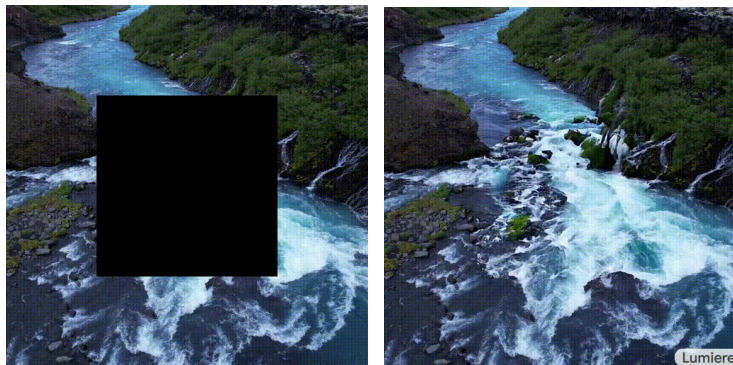
→ Появилось множество новых направлений для применения моделей диффузионной генерации видео:

- Text-to-Video
- Image-to-Video
- Video Inpainting
- Video Editing
- Video Enhancement

→ Области применения:

- Развлечения
- Искусство
- Реклама

Video Inpainting



Video Enhancement



Video Editing



02

Введение в диффузионные модели

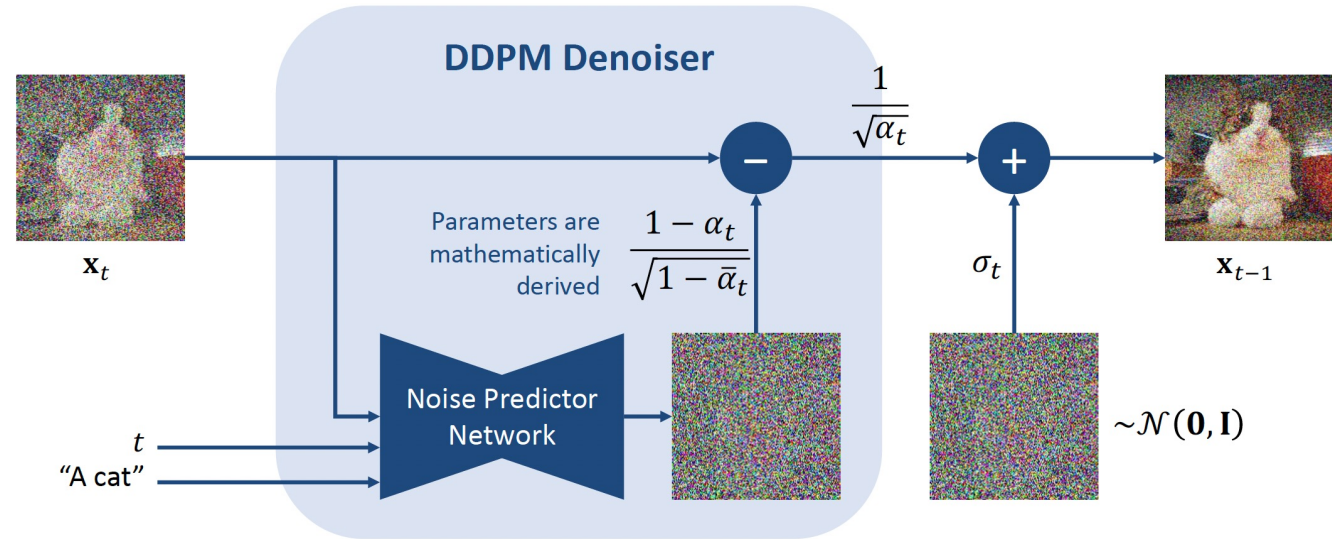
Введение в диффузионные модели

Прямой процесс диффузии – последовательное добавление шума

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

Обратный процесс диффузии – последовательное удаление шума с помощью аппроксимации нейросети

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t))$$



Данные

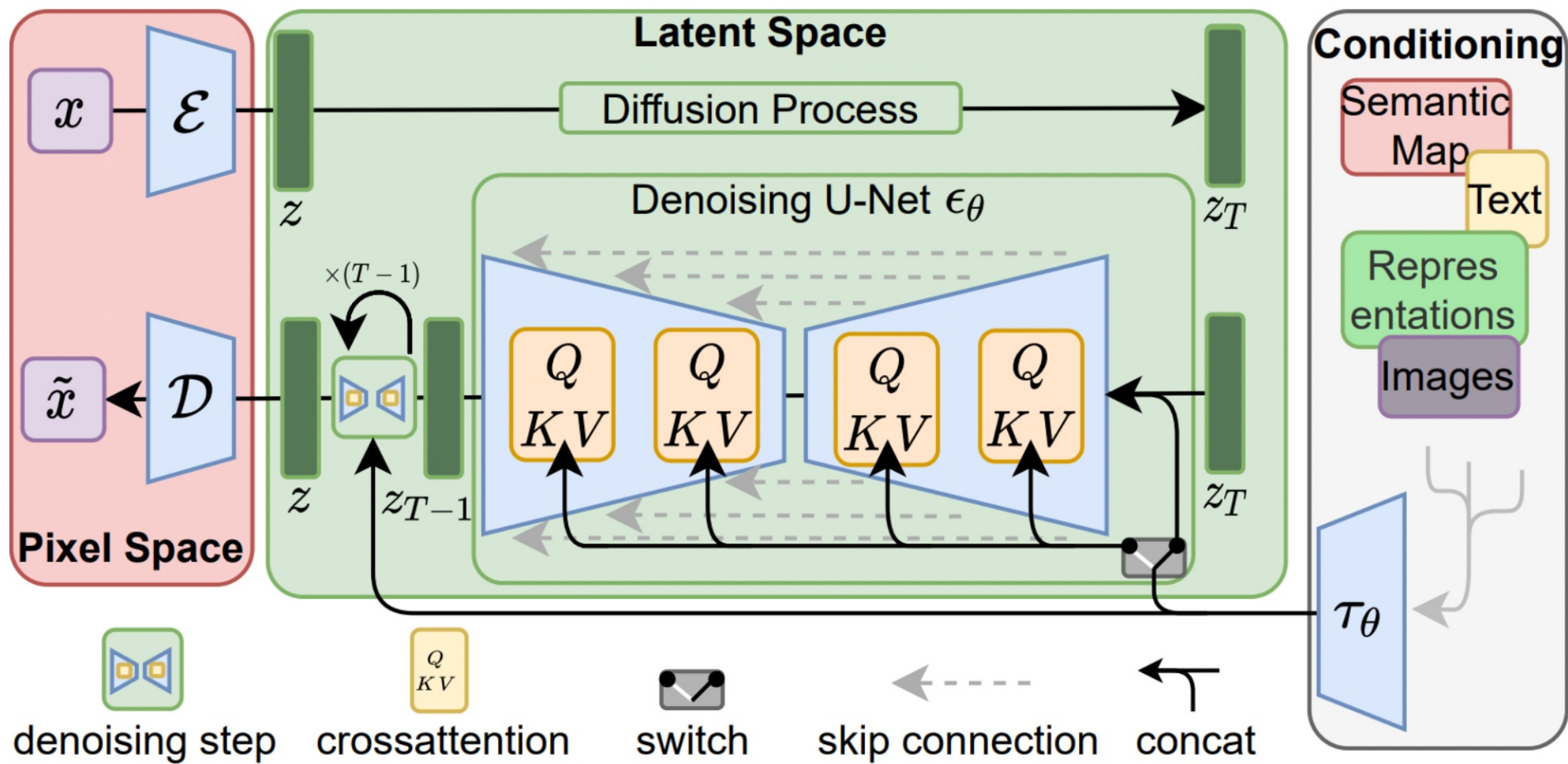


Forward diffusion process (fixed)

Reverse denoising process (generative)

Шум

Stable Diffusion



Stable Diffusion



child's crayon drawing of minions



Minion



Biblically accurate Minion



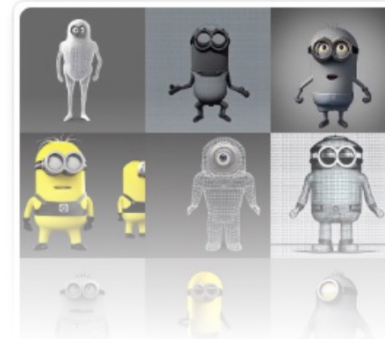
Historical footage of a riot caused by Minions, Nuremberg 1930s, grainy, detailed



Totoro from my Neighbour



vintage photograph of minions operating a tank during ww2



03

Первые работы по видео диффузии

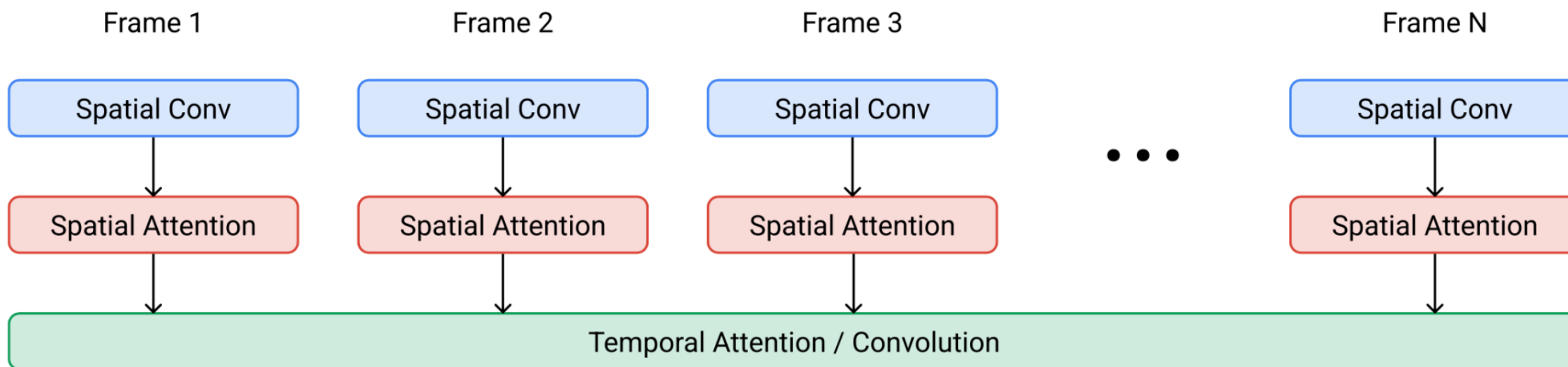
Постановка задачи

2D output \rightarrow 3D output



Video Diffusion Models

- 3D U-Net работающий в размерности пространства и времени
- Картиночные 2D свертки \rightarrow пространственные 3D свертки, то есть от 2D к (2+1)D свертке
 - Kernel size: $(3 \times 3) \rightarrow (1 \times 3 \times 3)$
 - Feature vectors: $(\text{height} \times \text{width} \times \text{channel}) \rightarrow (\text{frame} \times \text{height} \times \text{width} \times \text{channel})$
- Пространственное внимание: не меняем
- Добавляем слой временного внимания: attend across the temporal dimension (spatial axes as batch)



Video Diffusion Models



Make-A-Video – каскадная генерация

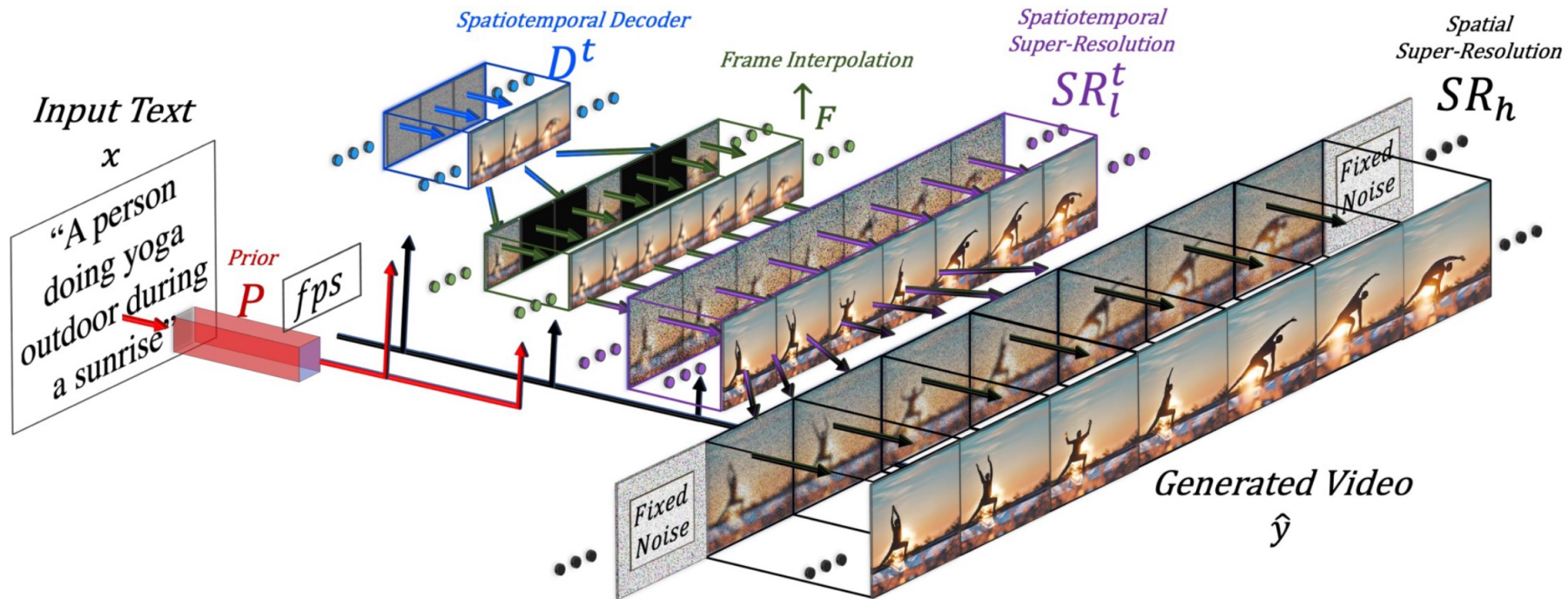
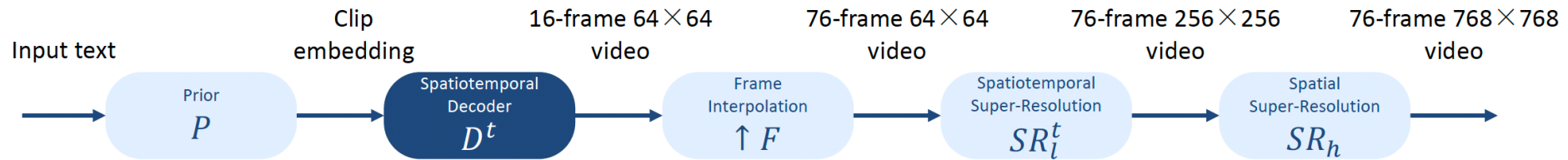
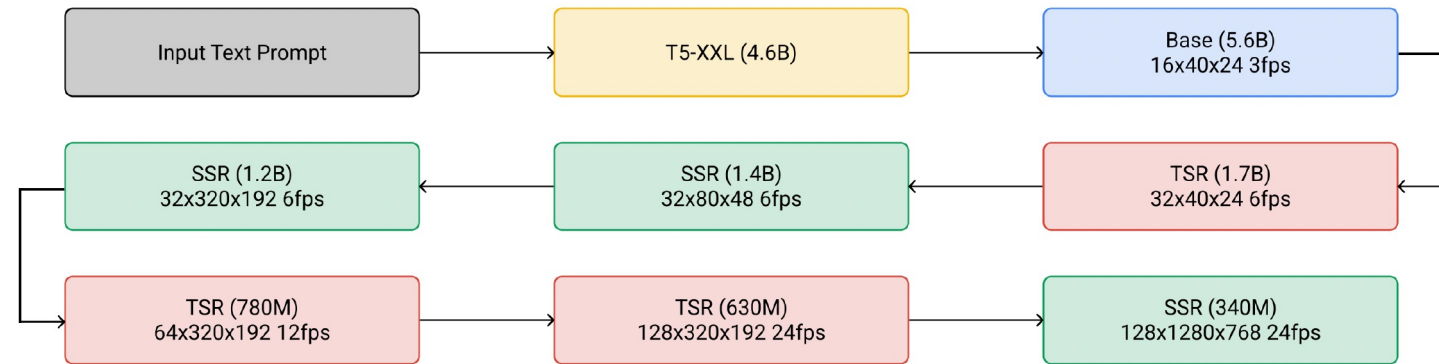
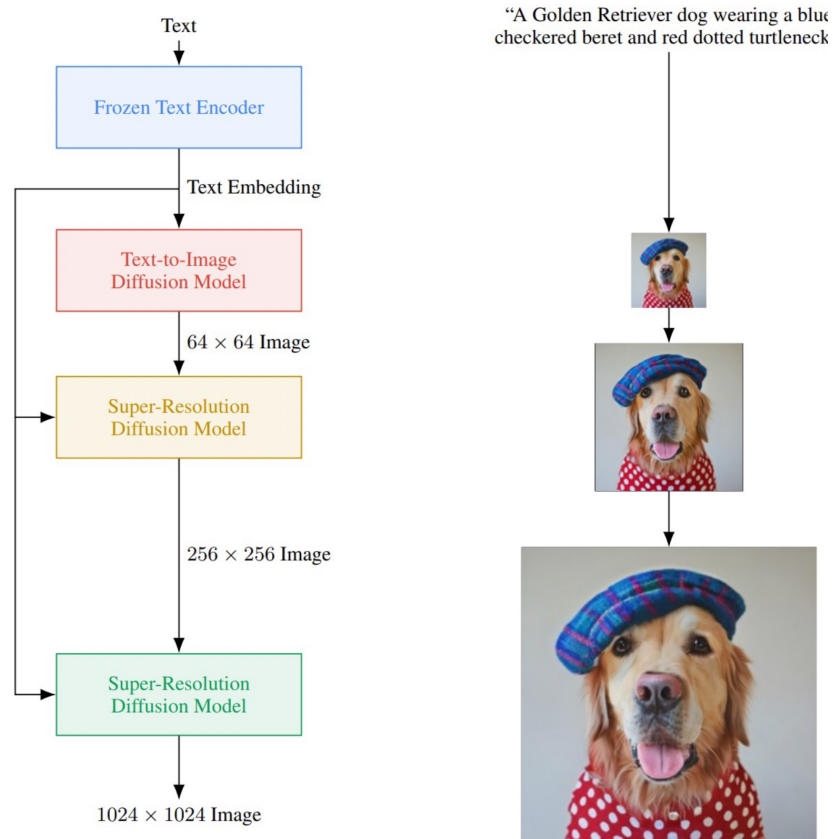


Imagen & Imagen Video

Imagen (for images)



Make-A-Video – примеры



A dog wearing a Superhero outfit with red cape flying through the sky



A golden retriever eating ice cream on a beautiful tropical beach at sunset, high resolution

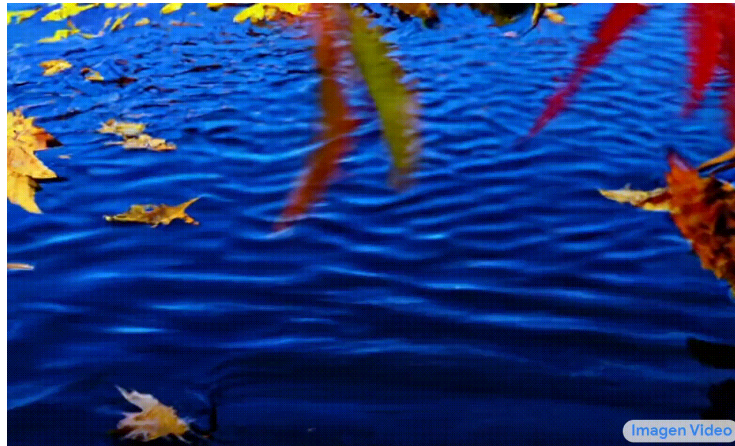


A teddy bear painting a portrait

Imagen Video



A happy elephant wearing a birthday hat walking under the sea.



A bunch of autumn leaves falling on a calm lake to form text 'Imagen Video'. Smooth.



Flying through an intense battle of ships in the ocean.

04

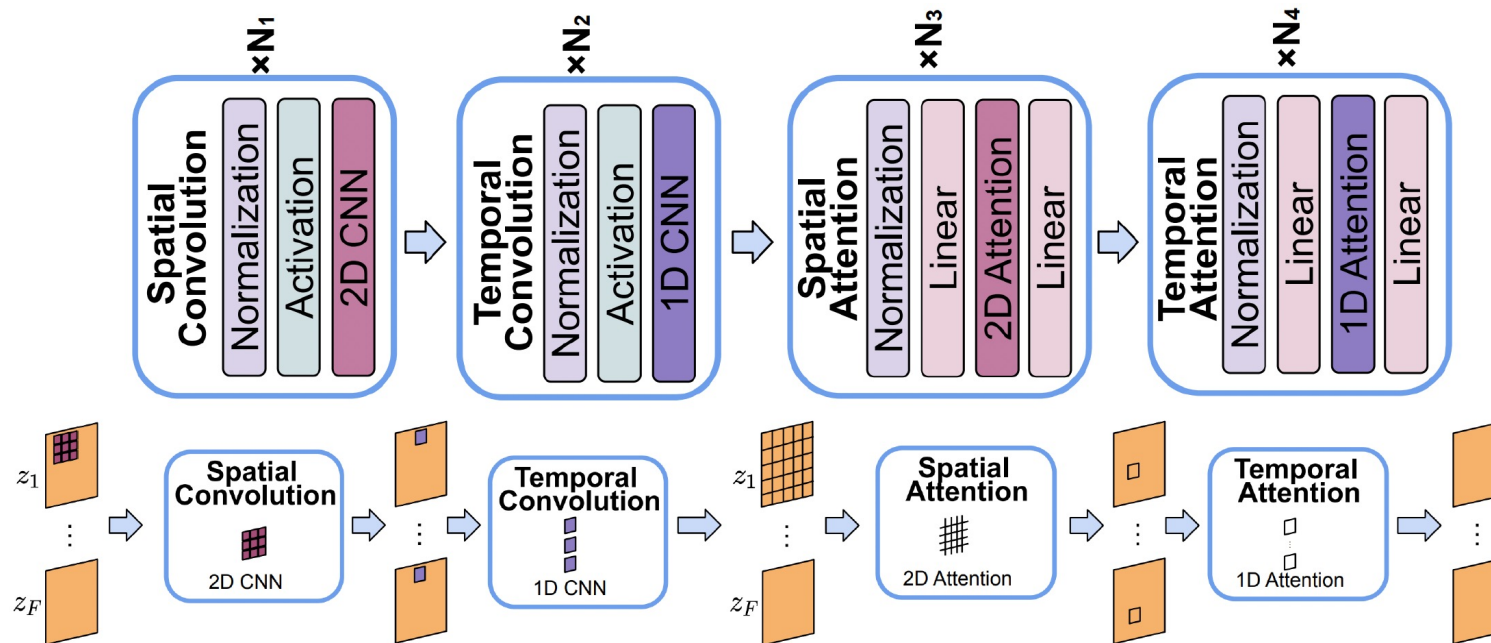
Адаптация T2I моделей на T2V

ModelScopeT2V

- T2V сложнее учить, чем T2I, а датасеты изображений гораздо более многообразны, чем видео (5B vs 10M);
- Видео данные менее многочисленны и разнообразны, чем картиночные (миллиарды vs десятки миллионов)ж
- В Open Source есть качественные T2I модели (Stable Diffusion);

Решение:

- Переиспользовать веса T2I моделей, расширяя архитектуру из 2D в 3D: добавить временную свертку и временное внимание.



ModelScopeT2V

Models	FID-vid (\downarrow)	FVD (\downarrow)	CLIPSIM (\uparrow)
NÜWA [62]	47.68	-	0.2439
CogVideo (Chinese) [20]	24.78	-	0.2614
CogVideo (English) [20]	23.59	1294	0.2631
MagicVideo [71]	-	1290	-
Video LDM [3]	-	-	0.2929
Make-A-Video [51]	13.17	-	0.3049
ModelScopeT2V (ours)	11.09	550	0.2930

Table 1: **Quantitative comparison with state-of-the-art models on MSR-VTT.** We evaluate the models with three metrics (*i.e.*, FID-vid [15], FVD [57], and CLIPSIM [61]).

ModelScopeT2V



Prompt Hyper-realistic photo of an abandoned industrial site during a storm.



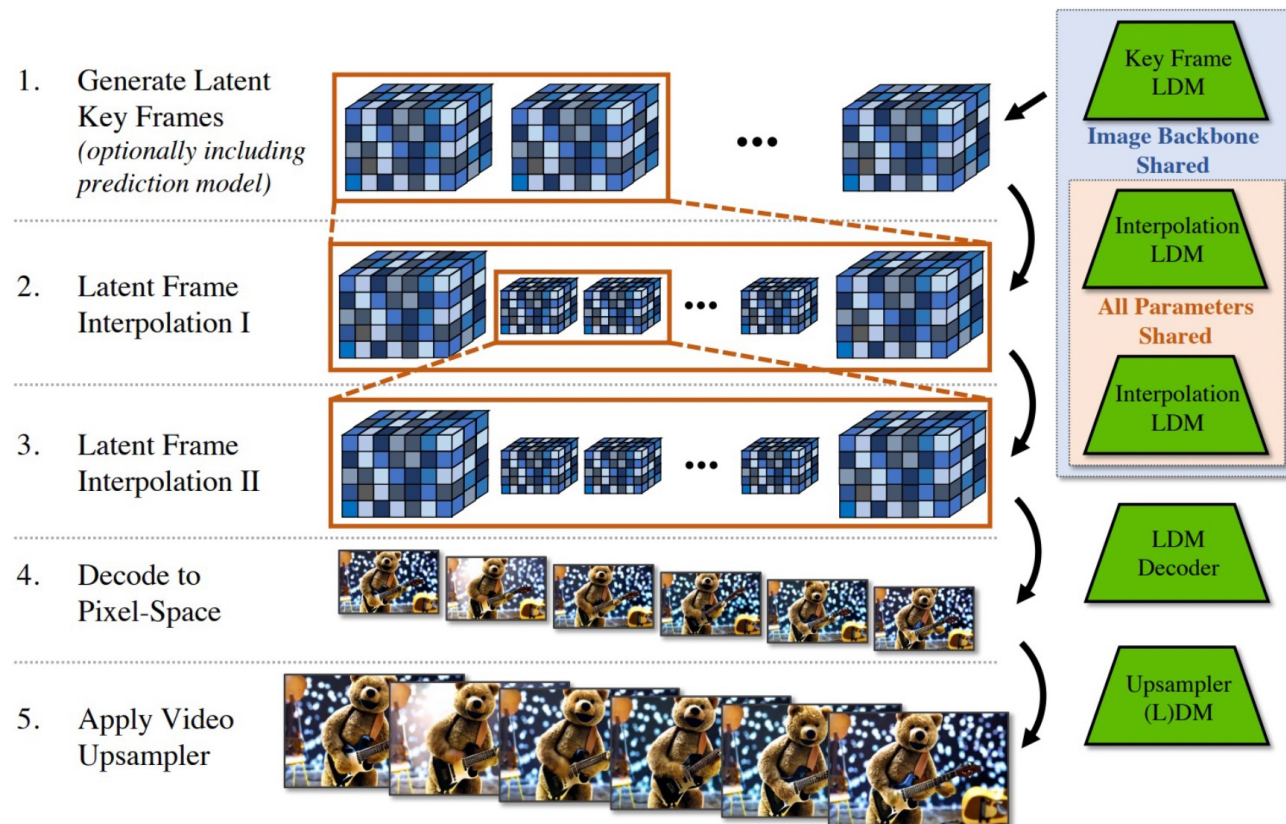
Align your Latents & Stable Video Diffusion

→ Align Your Latents:

- Расширение Stable Diffusion;
- Каскад из 5 моделей позволяет генерировать Full HD видео до 113 кадров;

→ Stable Video Diffusion:

- Та же архитектура, что и в Align Your Latents;
- Показали важность фильтрации данных.

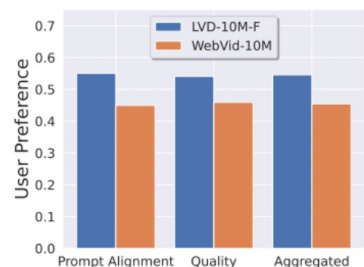


Stable Video Diffusion

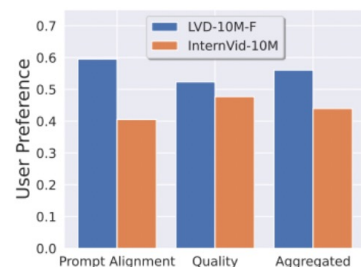
Стадии обучения

- Стадия I: Предобучение на изображениях (инициализация весов Stable Diffusion 2.1)
- Стадия II: Создание предварительной обучающей базы видеоданных:

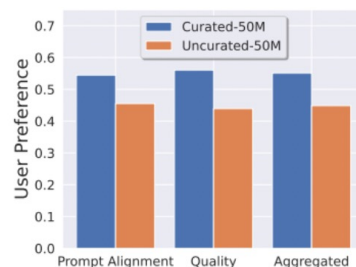
→ Фильтрация данных для предобучения повышает качество модели



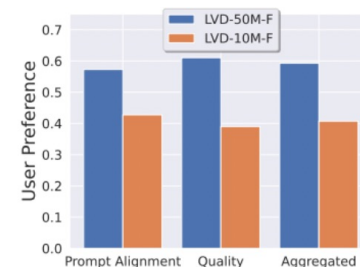
(a) User preference for *LVD-10M-F* and WebVid [6].



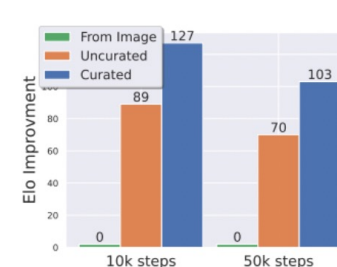
(b) User preference for *LVD-10M-F* and InternVid [96].



(c) User preference at 50M samples scales.

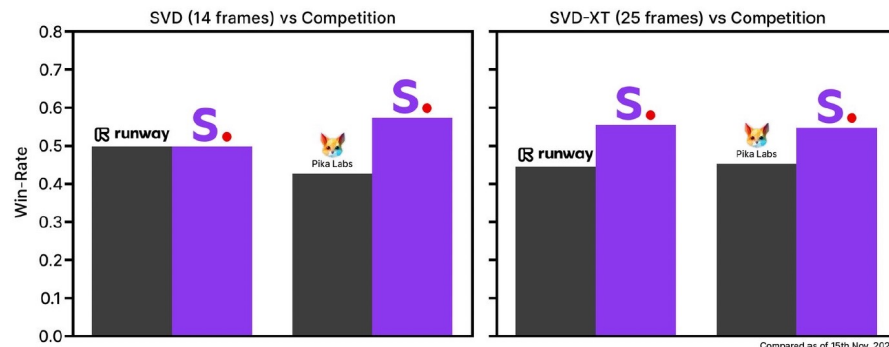


(d) User preference on scaling datasets.



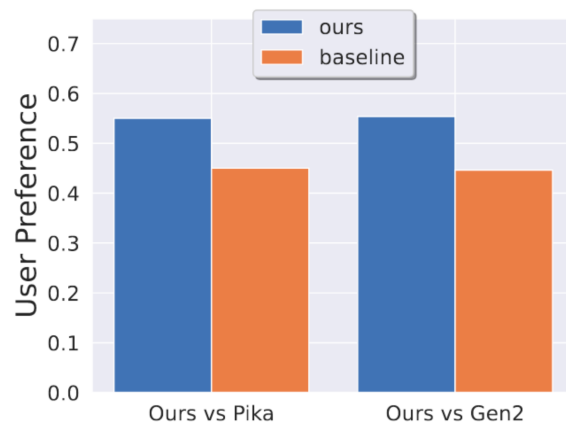
(e) Relative ELO progression over time during Stage III.

→ Стадия III: Дообучение модели на высококачественных видеоданных



Stable Video Diffusion

Method	FVD (↓)
CogVideo (ZH) [41]	751.34
CogVideo (EN) [41]	701.59
Make-A-Video [78]	367.23
Video LDM [8]	550.61
MagicVideo [110]	655.00
PYOCO [27]	355.20
SVD (ours)	242.02

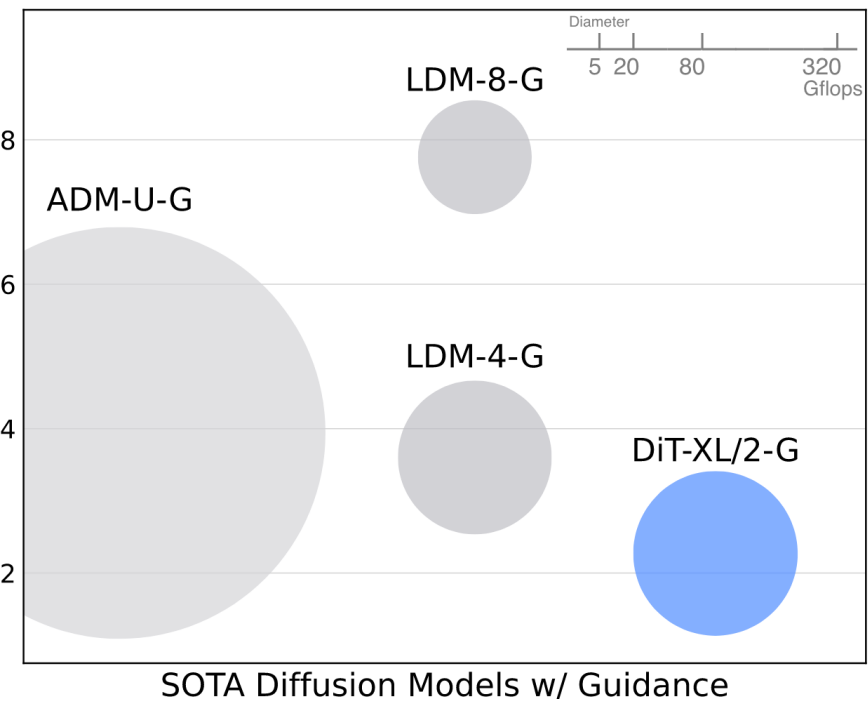
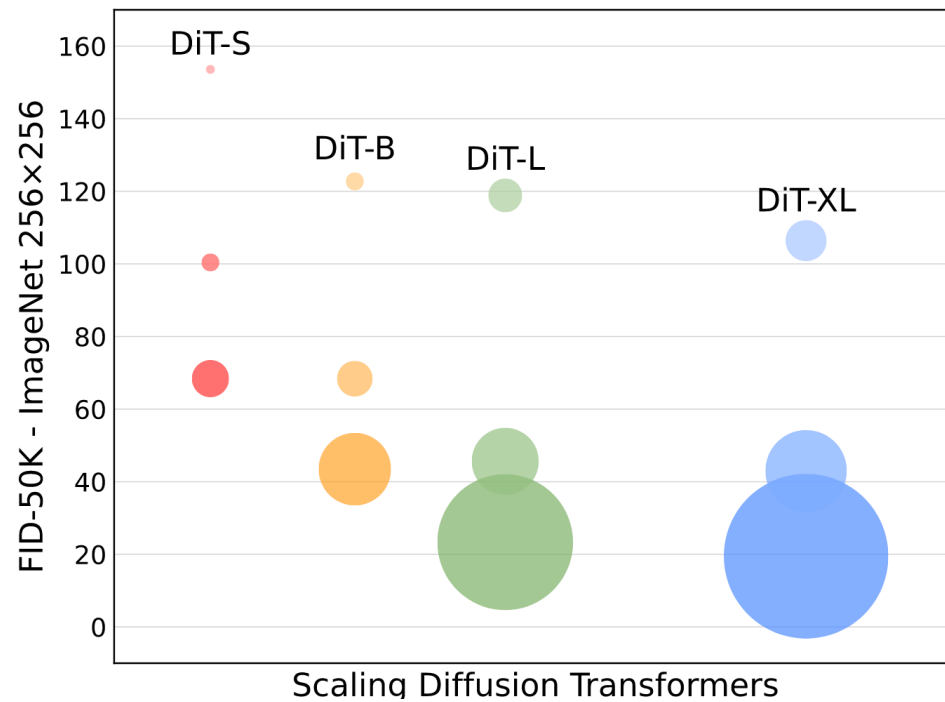


05

Диффузионные трансформеры для видео

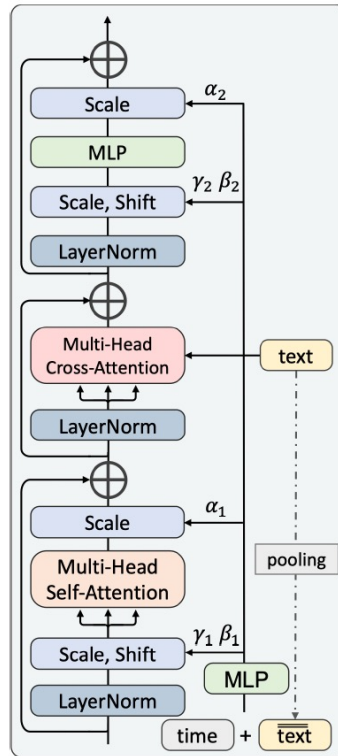
Мотивация

- U-Net плохо растет в качестве с увеличением параметров
- Нужна новая архитектура, которая хорошо масштабируется
- Диффузионный трансформер (DiT) – отличный вариант

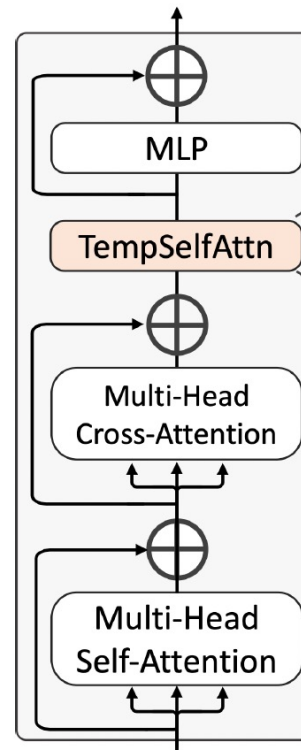


GenTron

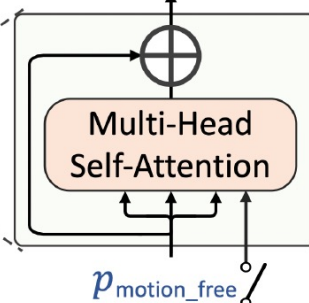
- Transformer-based architecture extended from DiT (class-conditioned transformer-based LDM)
- Train T2I -> insert temporal self-attn -> joint image-video finetuning (motion-free guidance)



Text-to-image architecture



Text-to-video architecture



Insert temporal self-attn for video generation

Motion-Free Mask

	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7
T_0	1	0	0	0	0	0	0	0
T_1	0	1	0	0	0	0	0	0
T_2	0	0	1	0	0	0	0	0
T_3	0	0	0	1	0	0	0	0
T_4	0	0	0	0	1	0	0	0
T_5	0	0	0	0	0	1	0	0
T_6	0	0	0	0	0	0	1	0
T_7	0	0	0	0	0	0	0	1

Motion-free guidance to allow joint image-video finetuning

GenTron



A fantasy landscape trending on Artstation, 4k



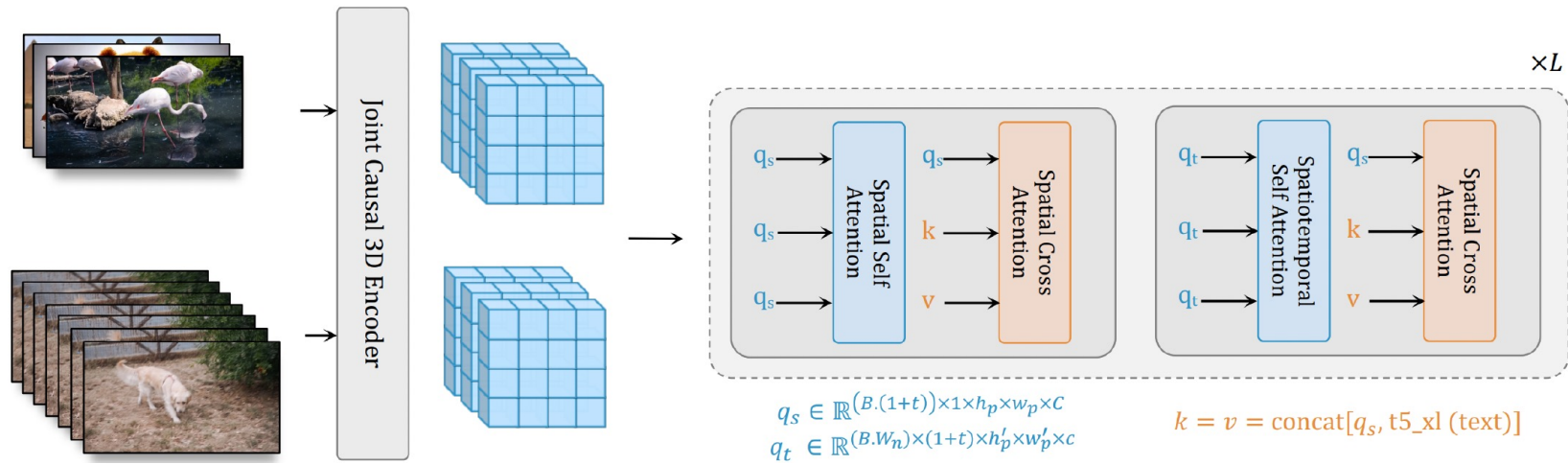
An astronaut flying in space, 4k high resolution



Traveler walking alone in the misty forest at sunset, 4k

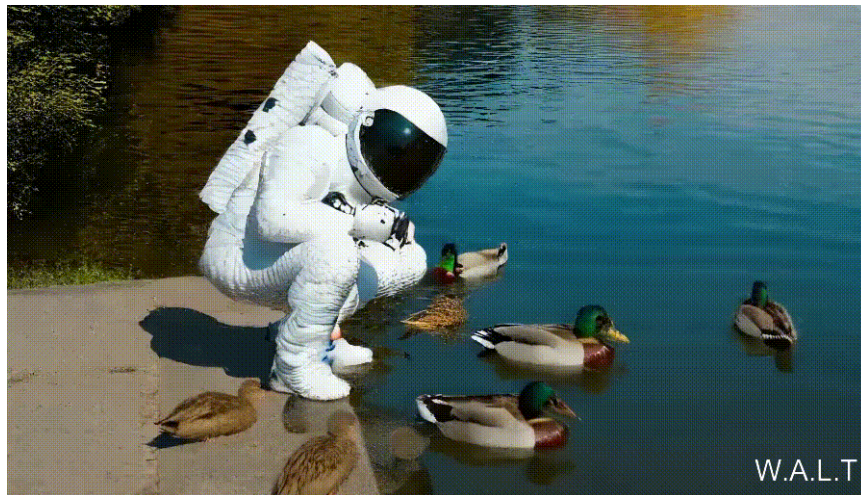
W.A.L.T.

- Transformer-based denoising diffusion backbone
- Joint image-video training via unified image/video latent space (created by a joint 3D encoder with causal 3D conv layers, allowing the first frame of a video to be tokenized independently)
- Window attention to reduce computing/memory costs
- Cascaded pipeline for high-quality generation



W.A.L.T.

Method	IS (↑)	FVD (↓)
CogVideo (Chinese) [37]	23.6	751.3
CogVideo (English) [37]	25.3	701.6
MagicVideo [88]	-	699.0
Make-A-Video [66]	33.0	367.2
Video LDM [4]	33.5	550.6
PYoCo [24]	47.8	355.2
W.A.L.T (Ours) 419M (video only)	26.8	598.8
W.A.L.T (Ours) 419M (video + image)	31.7	344.5
W.A.L.T (Ours) 3B (video + image)	35.1	258.1



06

Sora и аналоги

Sora-like модели

15 февраля 2024 года Open AI анонсировали свою модель генерации видео под названием Sora, которая существенно превосходит все предыдущие модели.

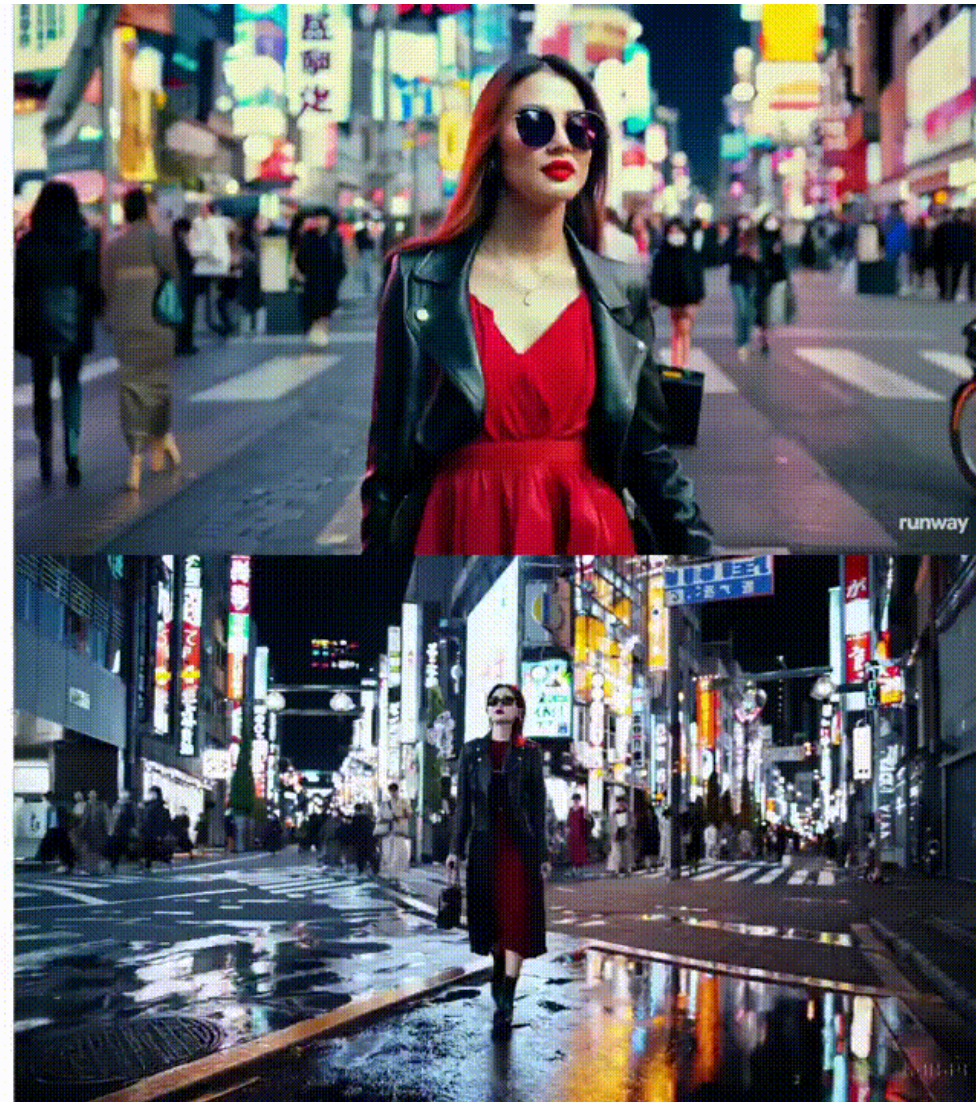
- Мало технических деталей
- Большое разрешение – 2048 x 1080
- Длинные видео ~ 2k кадров
- Диффузионный трансформер

 runway

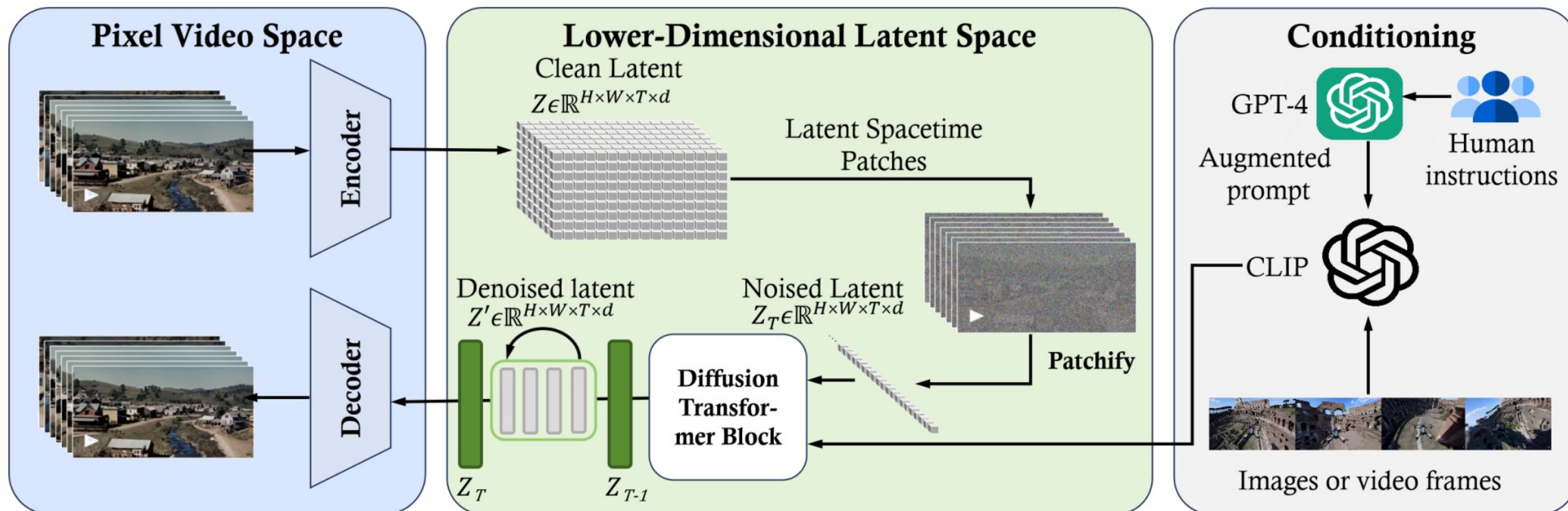
Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and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

@ProperPrompter

 Sora



Архитектура

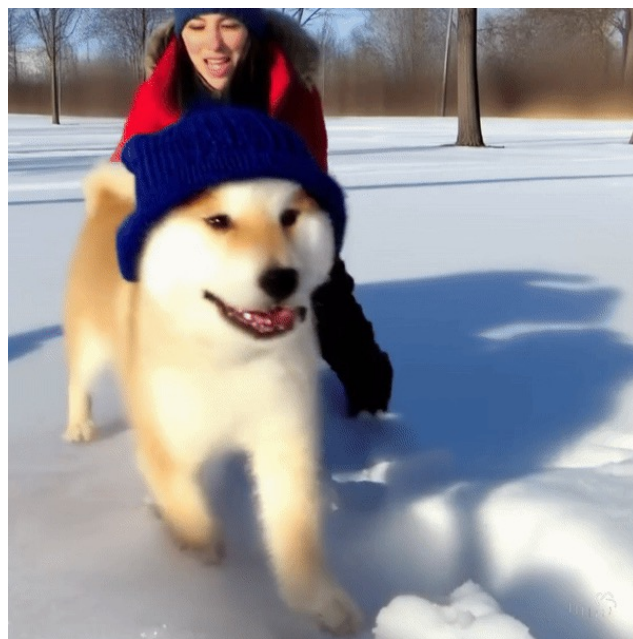


Скейлинг

Авторы показали, что диффузионные трансформаторы эффективно масштабируются. Ниже мы приводим сравнение примеров видео с фиксированными семплами и входными данными по мере обучения. Качество примеров заметно улучшается по мере увеличения количества обучающих вычислений.



Базовая модель



4x модель



32x модель

Список моделей

С момента анонса Sora ряд компаний представили свои модели генерации видео сопоставимого качества.

Модель	Компания	Страна	Статус	Детали
Sora	Open AI	США	Анонс	1080р видео до 60 секунд
Vidu	ShengShu	Китай	Анонс	1080р видео до 16 секунд
Veo	Google	США	Анонс	1080р видео до 60 секунд
Kling	Kuaishou	Китай	Открытый доступ	1080р видео до 5 секунд
Dream Machine	Luma Labs	США	Открытый доступ	~720р видео до 5 секунд
Gen-3	Runway AI	США	Открытый доступ (платный)	~720р видео до 10 секунд
CogVideoX	Tsinghua University	Китай	Открытый код и веса	720×480 видео до 6 секунд
OpenSora & OpenSoraPlan	AI Community	-	Открытый код и веса	~720р видео до 4 секунд

Открытый доступ: Kling & Dream machine

Два проекта от энтузиастов из интернета

- Оба генерируют 720p видео 4 секунды;
- У обоих проектов уже третья версия модели;
- Модели ощутимо меняются от версии к версии.

Ссылки на проекты:

- Open Sora: <https://github.com/hpcaitech/Open-Sora>
- Open Sora Plan: <https://github.com/PKU-YuanGroup/Open-Sora-Plan>

Генерации Open Sora Plan



Генерации Open Sora



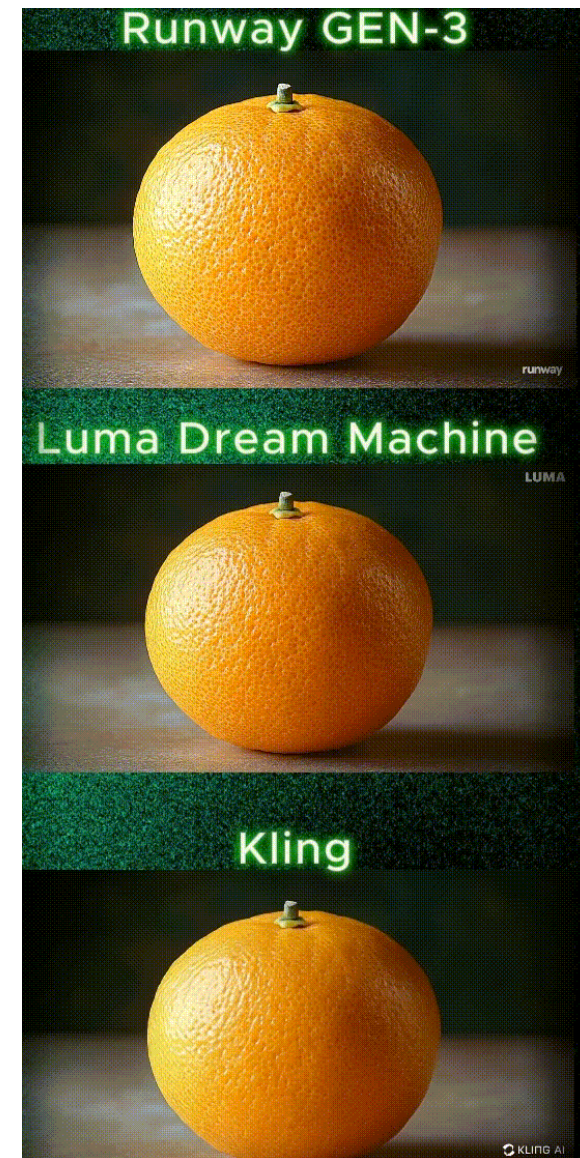
Open Access: Kling vs Dream Machine vs Gen3

Модель	Детали	Режимы	Сайт
Kling	1080p видео до 5 секунд	Text-to-Video Image-to-Video	https://kling.kuaishou.com
Dream Machine	~720p видео до 5 секунд		https://lumalabs.ai/dream-machine
Gen-3	~720p видео до 10 секунд		https://runwayml.com

Midjourney + Dream Machine

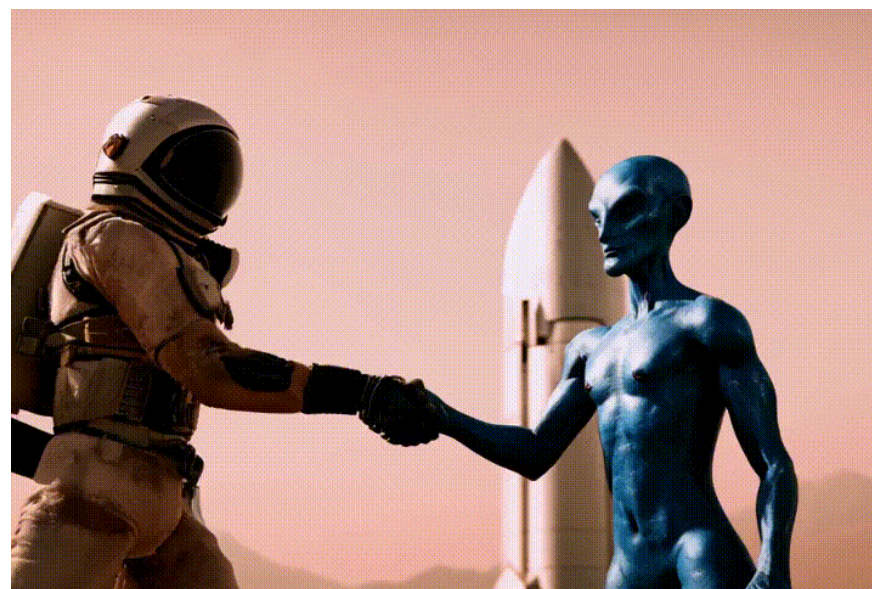
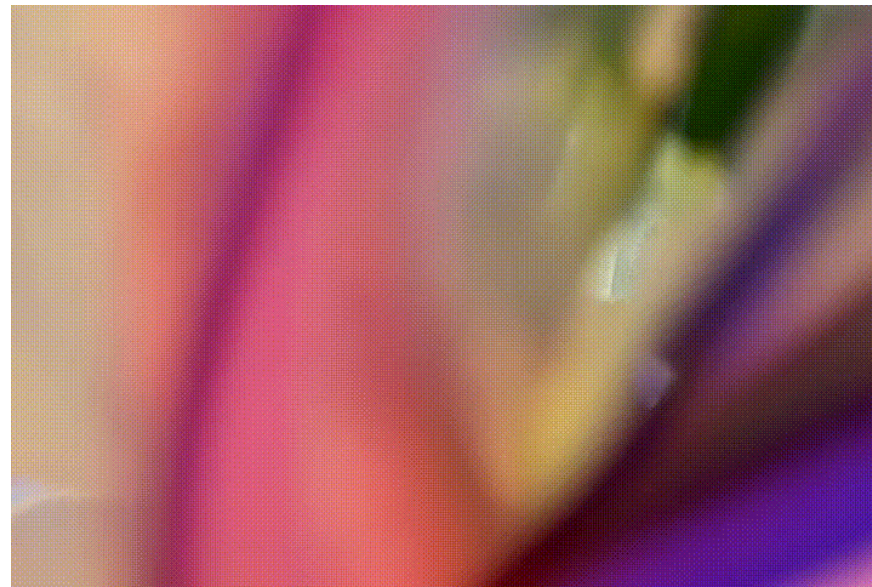


Midjourney + Gen 3



Open Source: OpenSora & OpenSoraPlan

- Генерирует 720 × 480 видео до 6 секунд
- Доступны 2B и 5B модели
- Разработала команда Tsinghua University
- Код и веса: <https://github.com/THUDM/CogVideo>



Open Source: Open Sora & Open Sora Plan

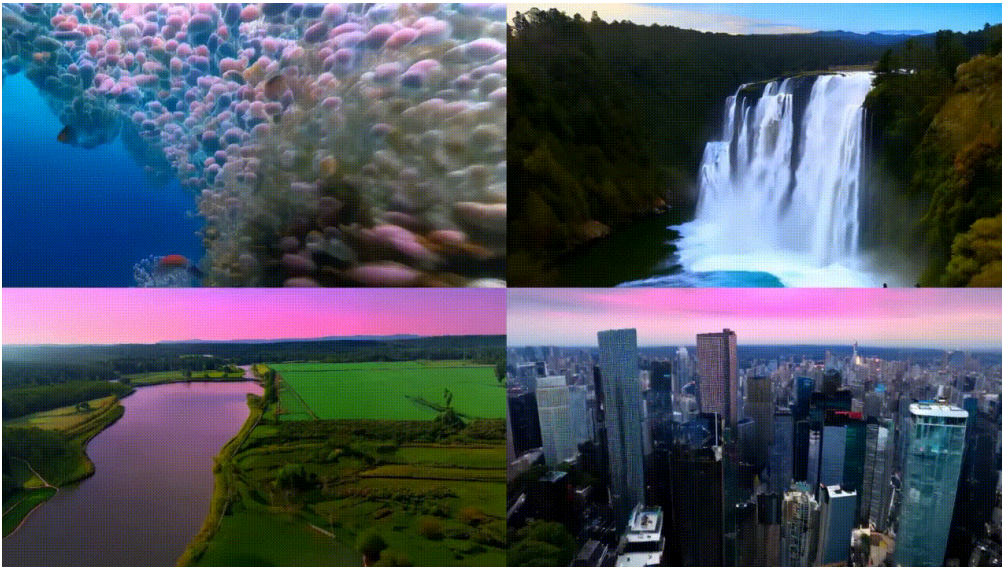
Два проекта от энтузиастов из интернета

- Оба генерируют 720p видео 4 секунды;
- У обоих проектов уже третья версия модели;
- Модели ощутимо меняются от версии к версии.

Ссылки на проекты:

- Open Sora: <https://github.com/hpcaitech/Open-Sora>
- Open Sora Plan: <https://github.com/PKU-YuanGroup/Open-Sora-Plan>

Генерации Open Sora Plan



Генерации Open Sora



Недостатки

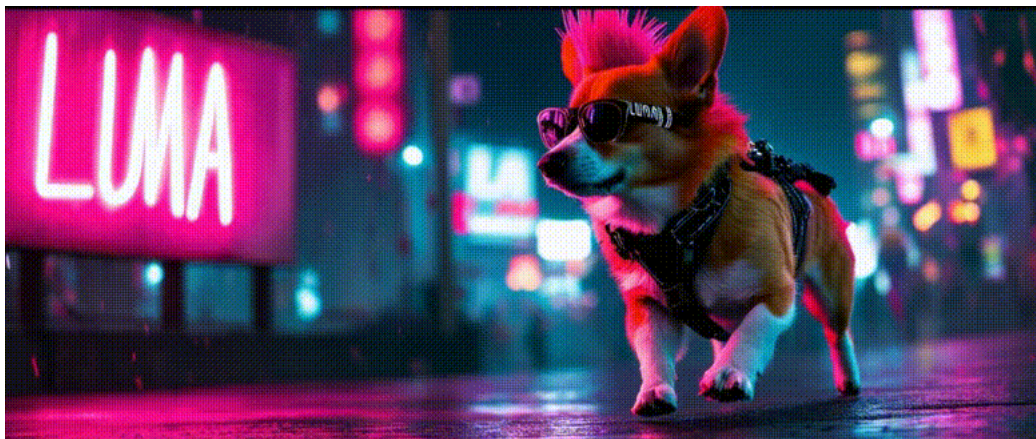
Морфинг объектов



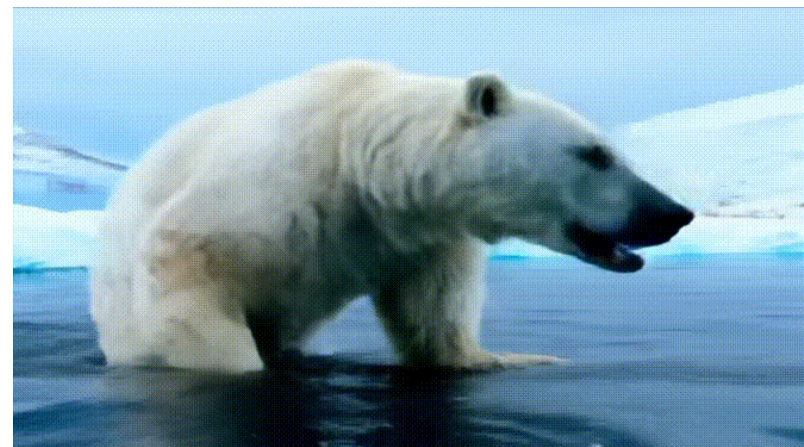
Ошибки в отрисовке текста



Нереалистичные движения



Физиологические ошибки



Недостатки

- Современные модели недостаточно хорошо понимают текст, а так же наш мир;
- Генерация длинных видео (больше 1 минуты) требует огромной вычислительной сложности;
- Отсутствуют крупные публичные наборы данных обучения (авторские права);
- Слабый контроль видео (текстовое описание, изображение).

Prompt: Step-printing scene of a person running, cinematic film shot in 35mm.



Prompt: Archeologists discover a generic plastic chair in the desert, excavating and dusting it with great care.



07

Выводы и перспективы

Выводы


- За последние 3 года в области диффузионные модели совершили прорыв в области генерации видео;
- Основные ингредиенты хорошей модели – это много качественных данных и много GPU;
- Уже существует множество сервисов и открытых моделей, с которыми можно бесплатно ознакомиться;
- Помимо генерации развиваются другие направления: Editing, Inpainting, Enhancement;
- Несмотря на неожиданный прирост в качестве с выходом Sora, данные модели в задаче генерации видео осталось много нерешенных проблем;

Контакты



Соболев Константин

AIRI, Научный сотрудник
Руководитель команды Video Generative AI

 sobolev@airi.net



airi.net



[airi_research_institute](https://www.telegram.com/@airi_research_institute)



[AIRI Institute](https://vk.com/AIRI_Institute)



[AIRI Institute](https://www.youtube.com/AIRI_Institute)



Telegram

Artificial Intelligence
Research Institute