NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

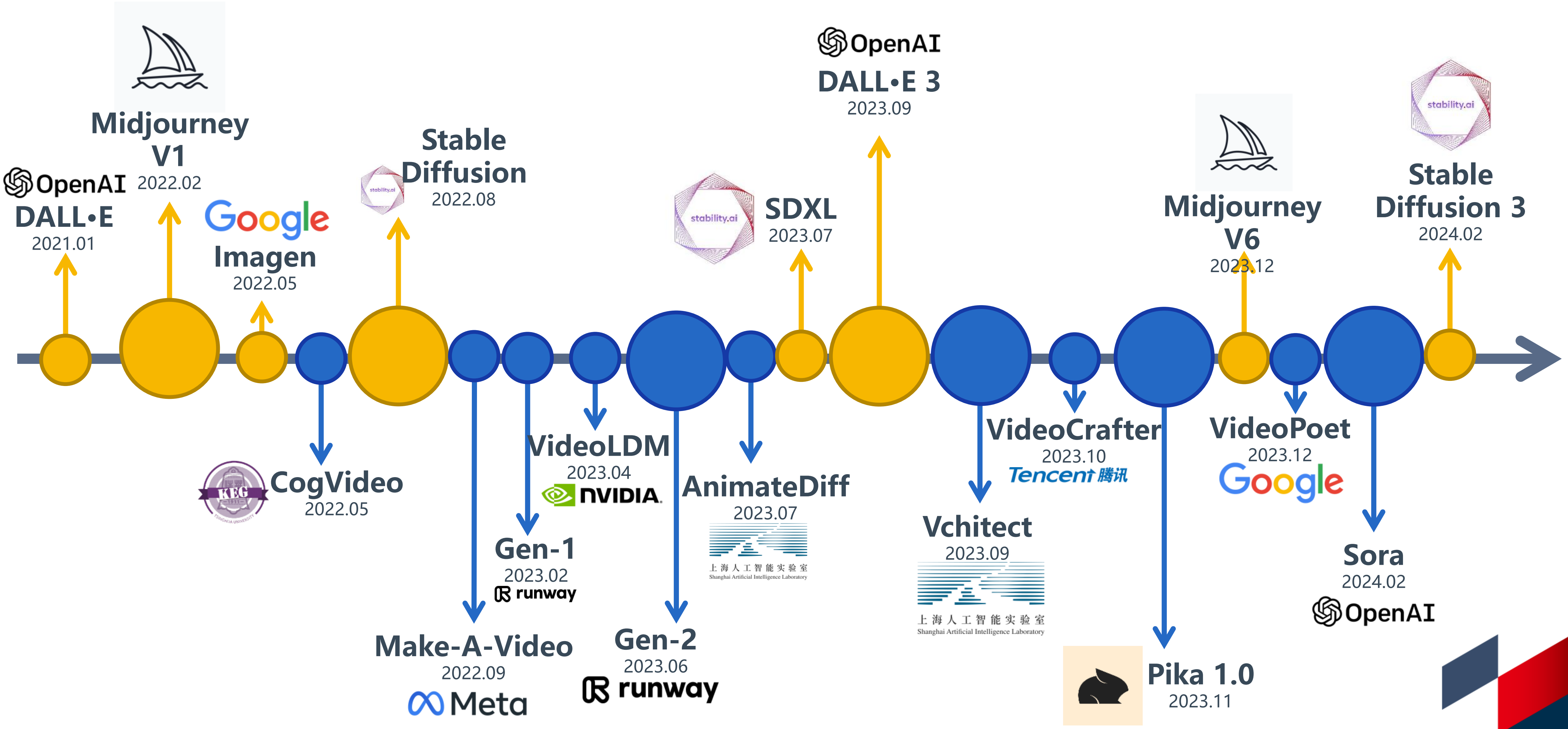
Vchitect: Building Open-Source Foundation System for Video Generation

Ziwei Liu（刘子纬）

https://liuziwei7.github.io/

Nanyang Technological University

The Timeline from T2I to T2V

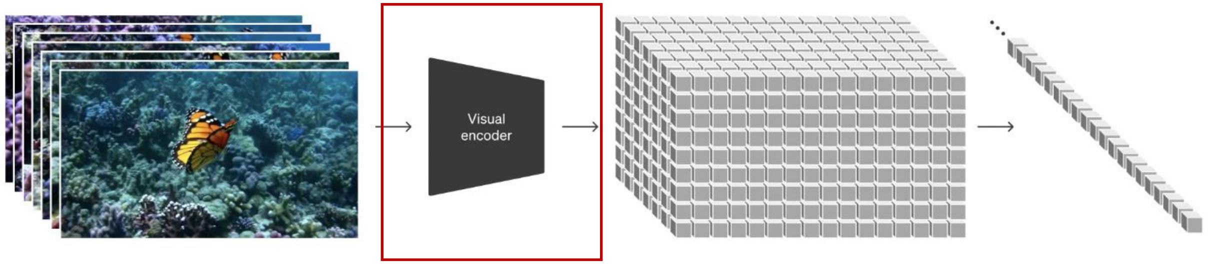


Sora - Capabilities

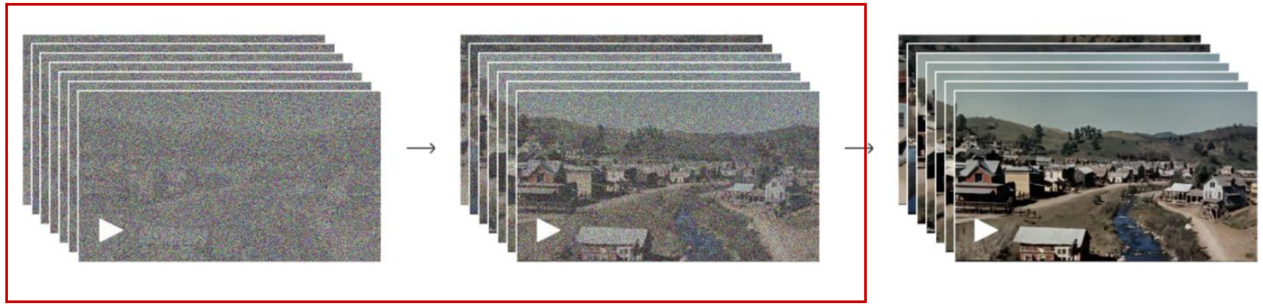
- **High Quality**
  - The generated video with excellent realism and details
- **Long Video Generation**
  - Generate videos up to a minute long
- **Diverse Resolution Generation**
  - Generate videos with variable durations, resolutions, aspect ratios
- **Physics-Aware Rendering**
  - Understand and simulate the physical world in motion

Sora - Method

- **Network Architecture**
  - Video Compression Network
  - LLM– GPT4
  - Diffusion Transformer Model
- **Training Data**
  - Possibly Physical Engine Synthetic Data



Video Compression Network

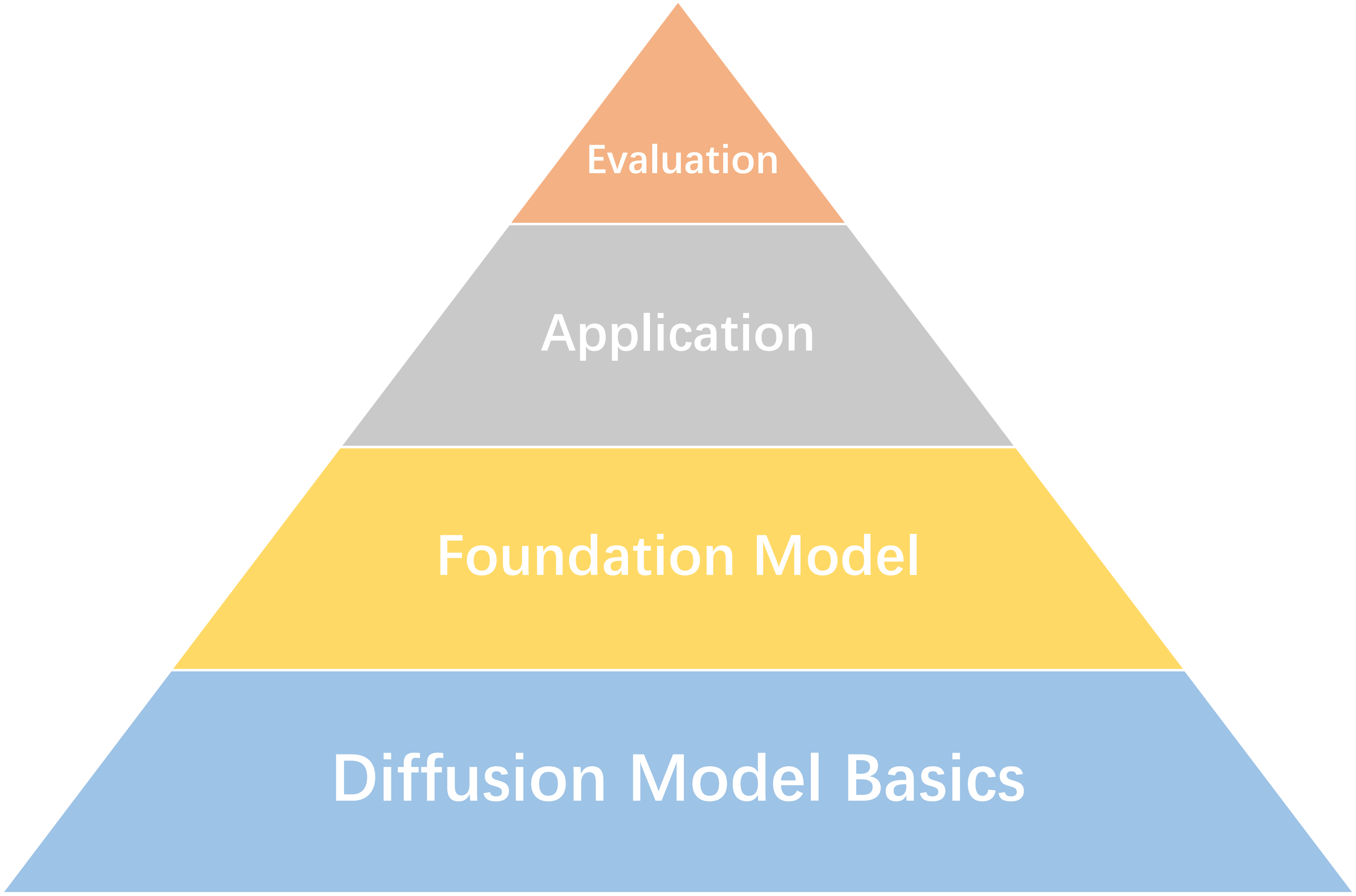


Diffusion Transformer + GPT4

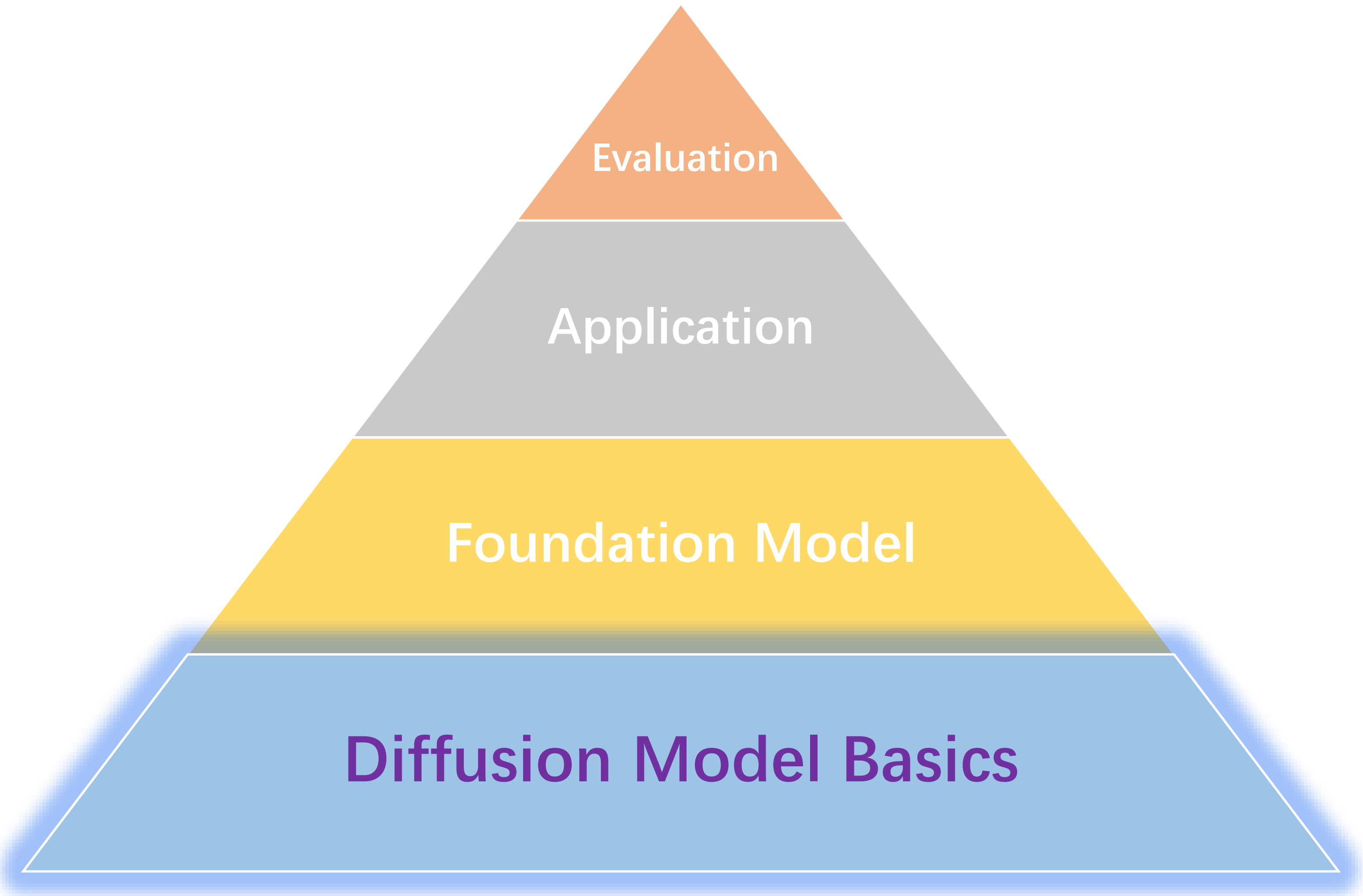
Sora - Weakness

- **Incorrect simulation of the physics in a complex scene**
  - Unable to accurately simulate complex physical processes (such as glass breaking)
- **Inaccurate physical geometry relationships**
  - The relative height of people and houses is unrealistic
  - Inaccurate projective geometry

Video Generation



Video Generation



FreeU: Free Lunch in Diffusion U-Net

reverse process / denoising process

gradually denoise to image

Image

Noise

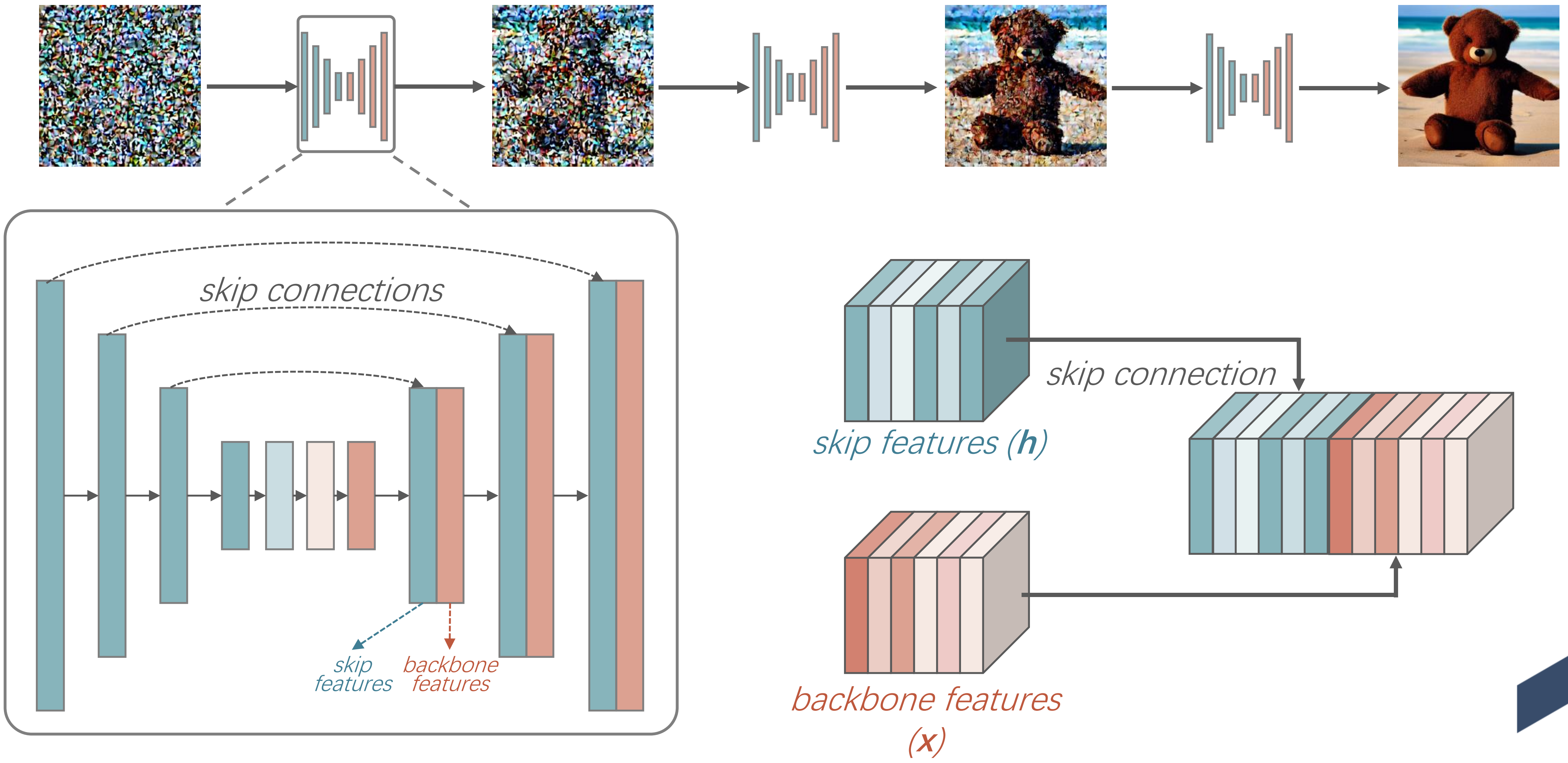
$\mathbf{x}_0$ $\mathbf{x}_{t-1}$ $\mathbf{x}_t$ $\mathbf{x}_T$

gradually adds Gaussian noise to the data

forward process / diffusion process

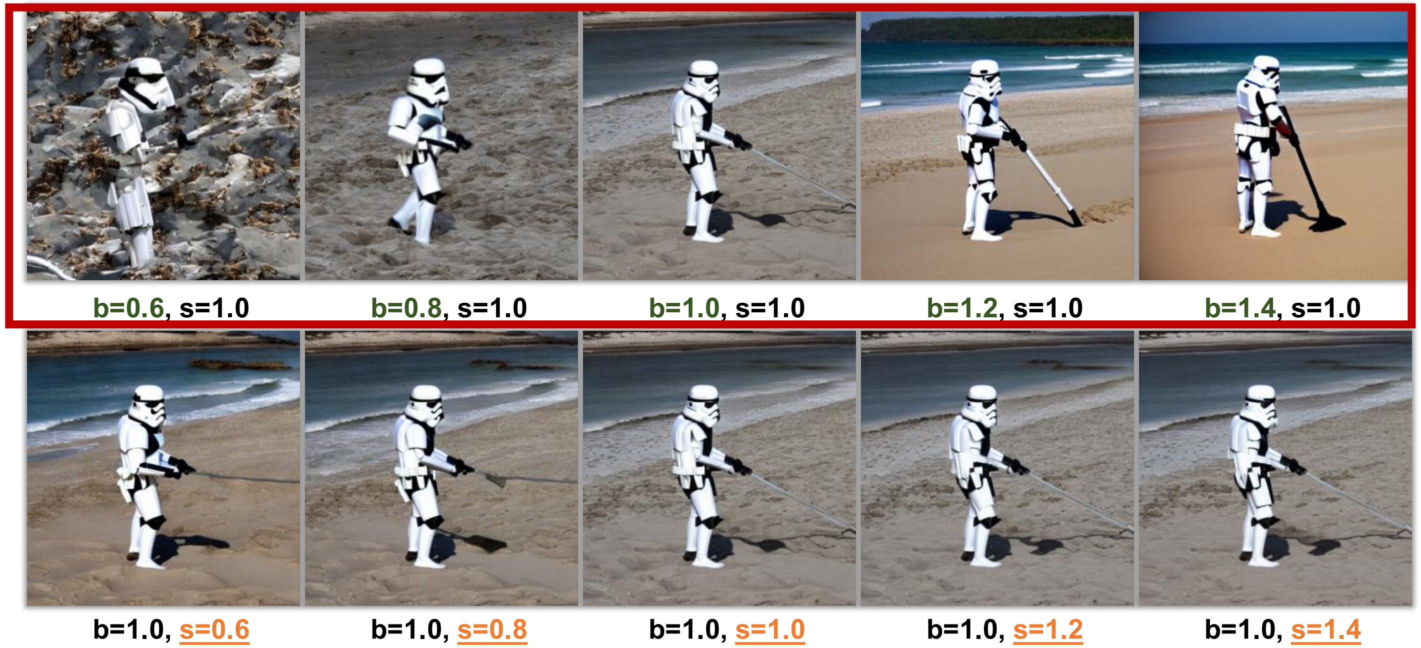
Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

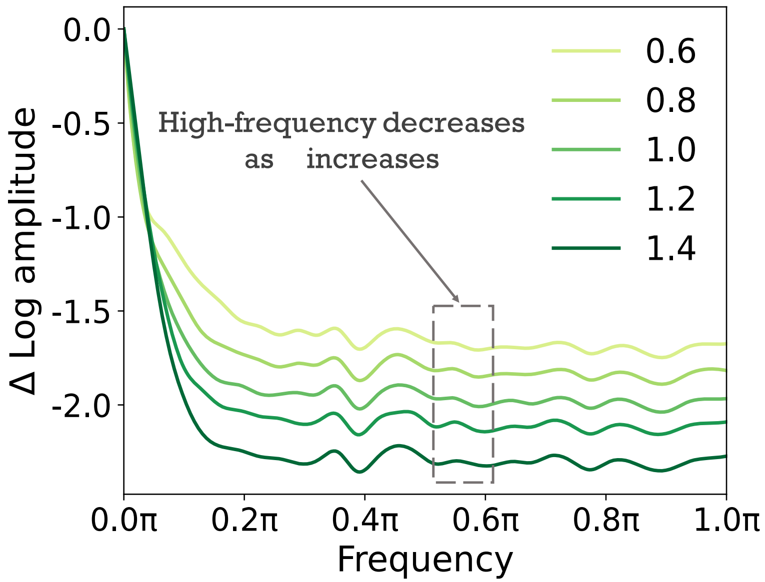
FreeU: Free Lunch in Diffusion U-Net



FreeU: Free Lunch in Diffusion U-Net

- **<u>Backbone</u>**: primarily contributes to denoising

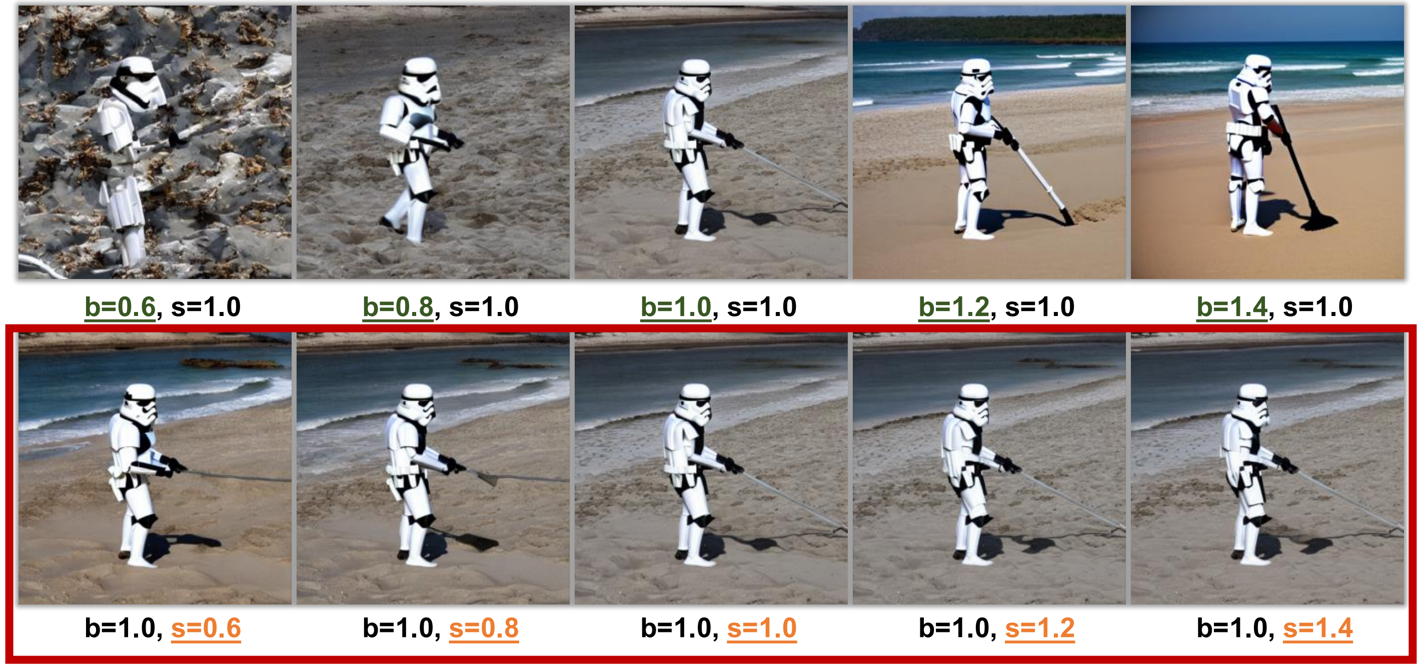


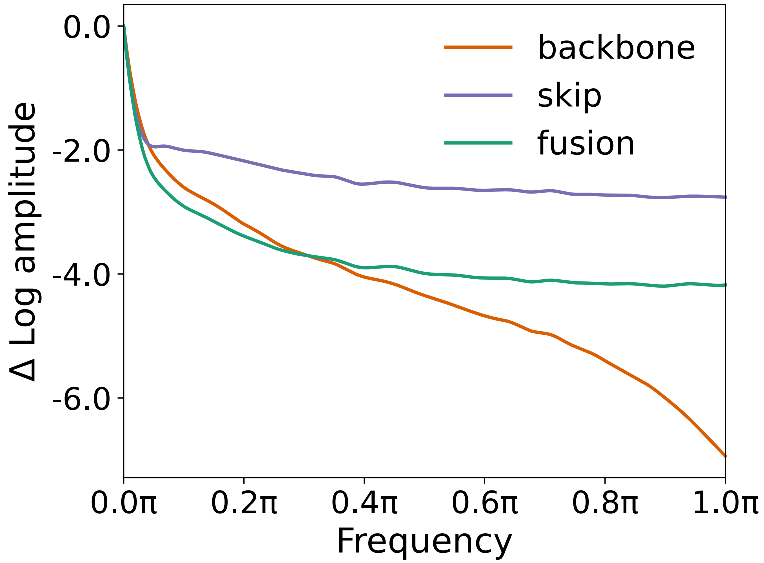


Fourier relative log amplitudes of variations of b

FreeU: Free Lunch in Diffusion U-Net

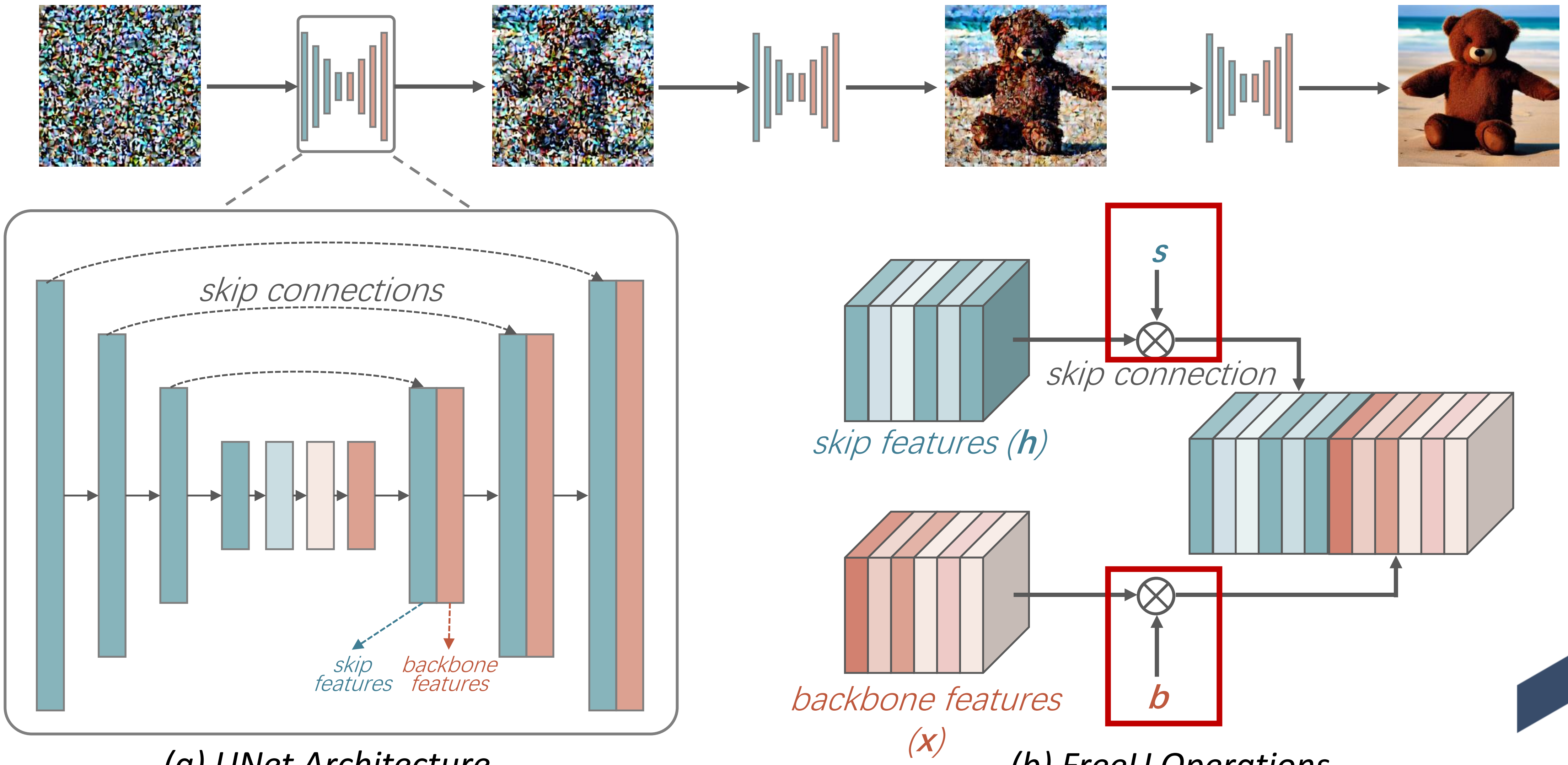
- **Backbone**: primarily contributes to denoising
- **Skip**: introduce high-frequency features into the decoder module





Fourier relative log amplitudes of backbone, skip, and their fused feature maps

FreeU: Free Lunch in Diffusion U-Net



(a) UNet Architecture

(b) FreeU Operations

Visual Results: Text-to-Image

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

SD | SD + FreeU | SD | SD + FreeU | SD | SD + FreeU

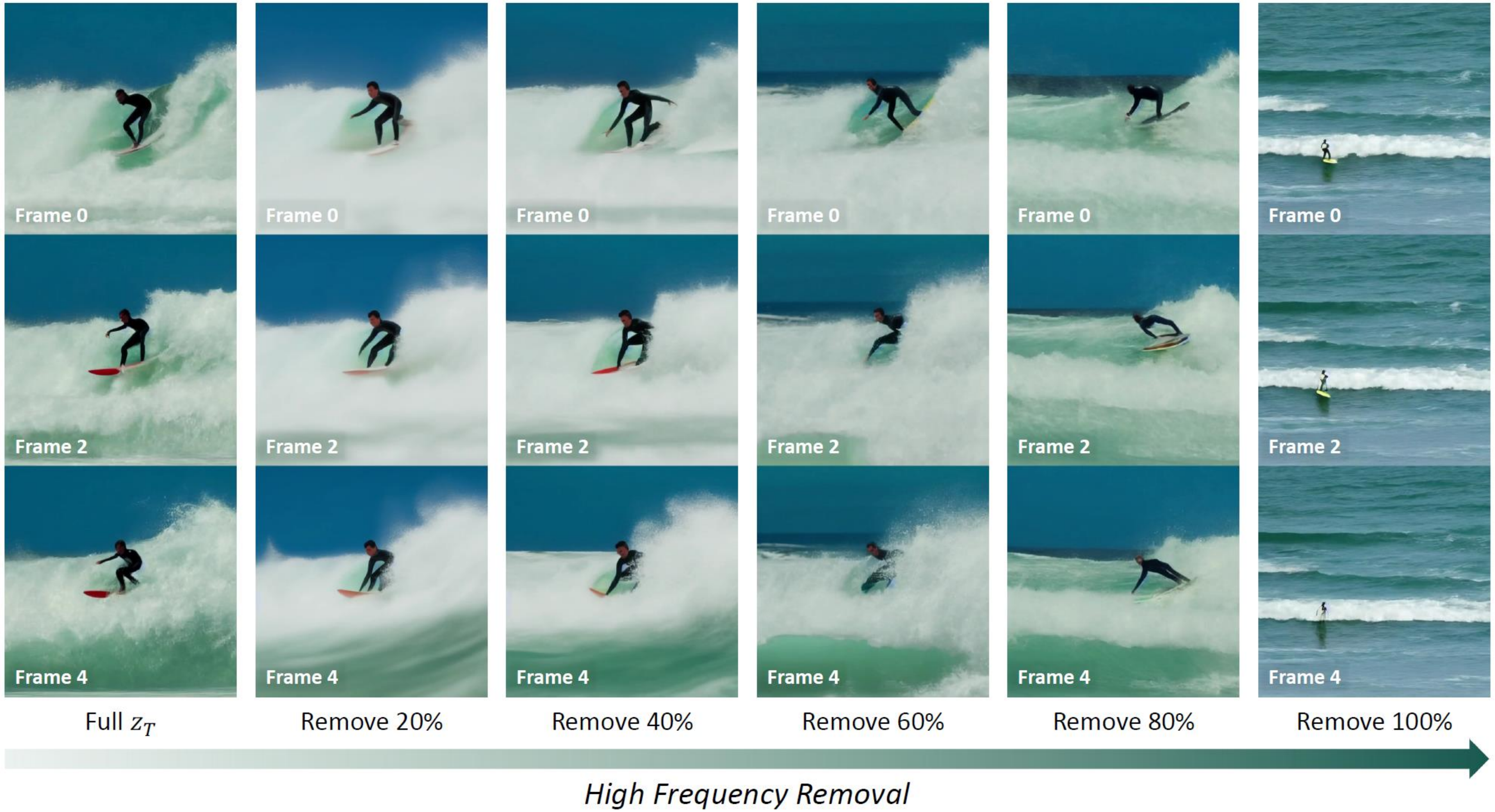
a blue car is being filmed | *Mother rabbit is raising baby rabbits* | *A bridge is depicted in the water*

a baby in a red shirt | *a attacks an upset cat and is then chased off* | *A teddy bear walking in the snowstorm*

A cat riding a motorcycle. | *A panda standing on a surfboard in the ocean* | *A boy is playing pokemon*

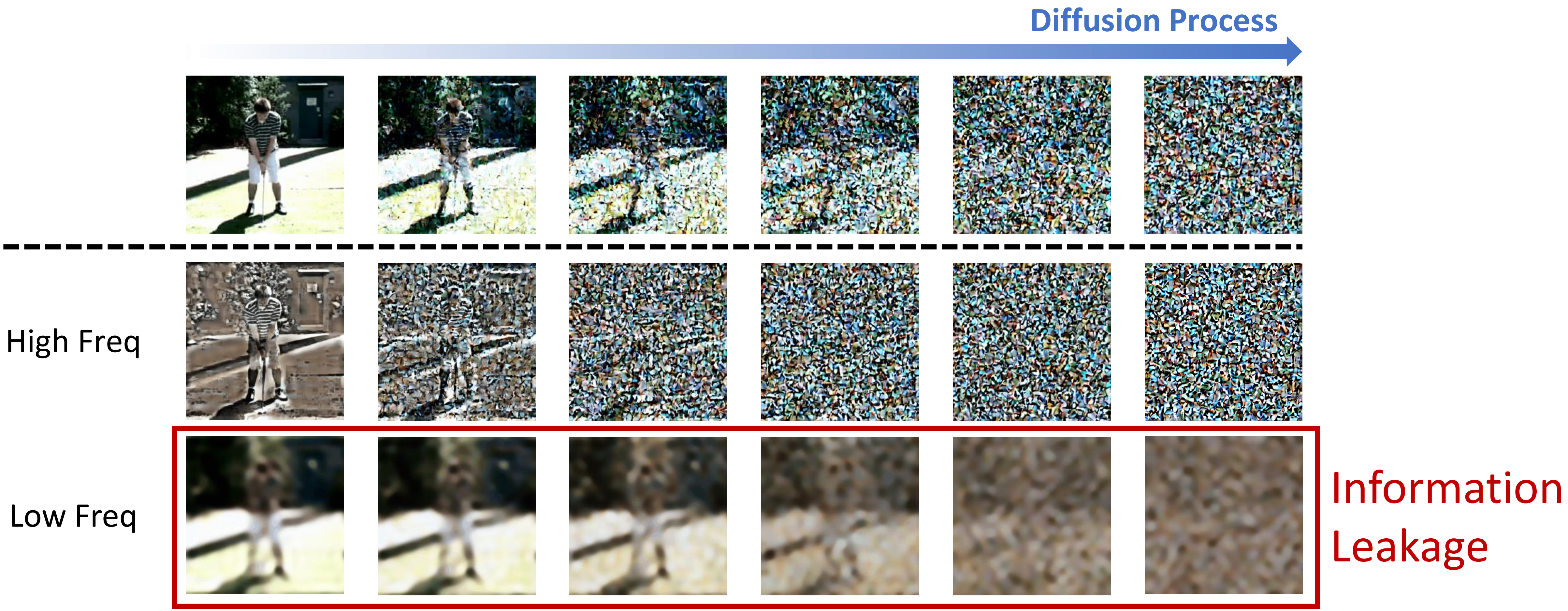
FreeInit

Bridging initialization gap in video diffusion models



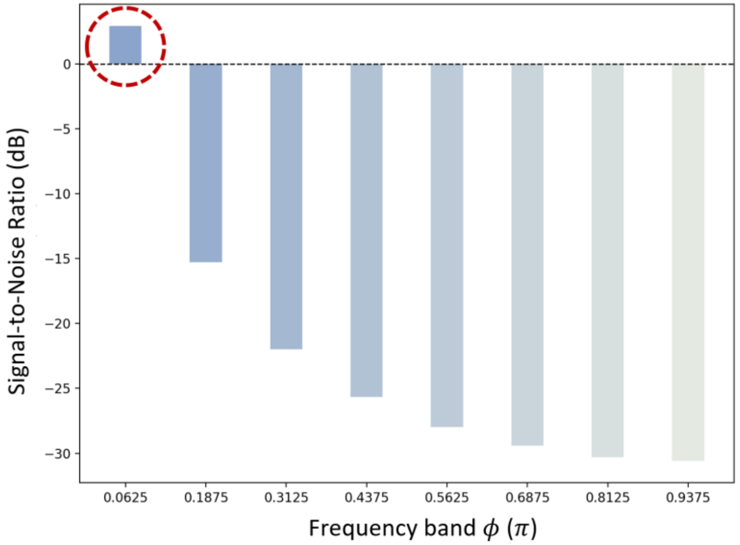
We observe that the spatio-temporal low-frequency components of the initial noise dominate the generation results.

Observation: Initialization Gap



However, the diffusion process at training cannot fully corrupt the low-frequency information, leaking correlations to initial noise

Observation: Initialization Gap



Initial noise at Training: High SNR at low-frequency band, information leaked

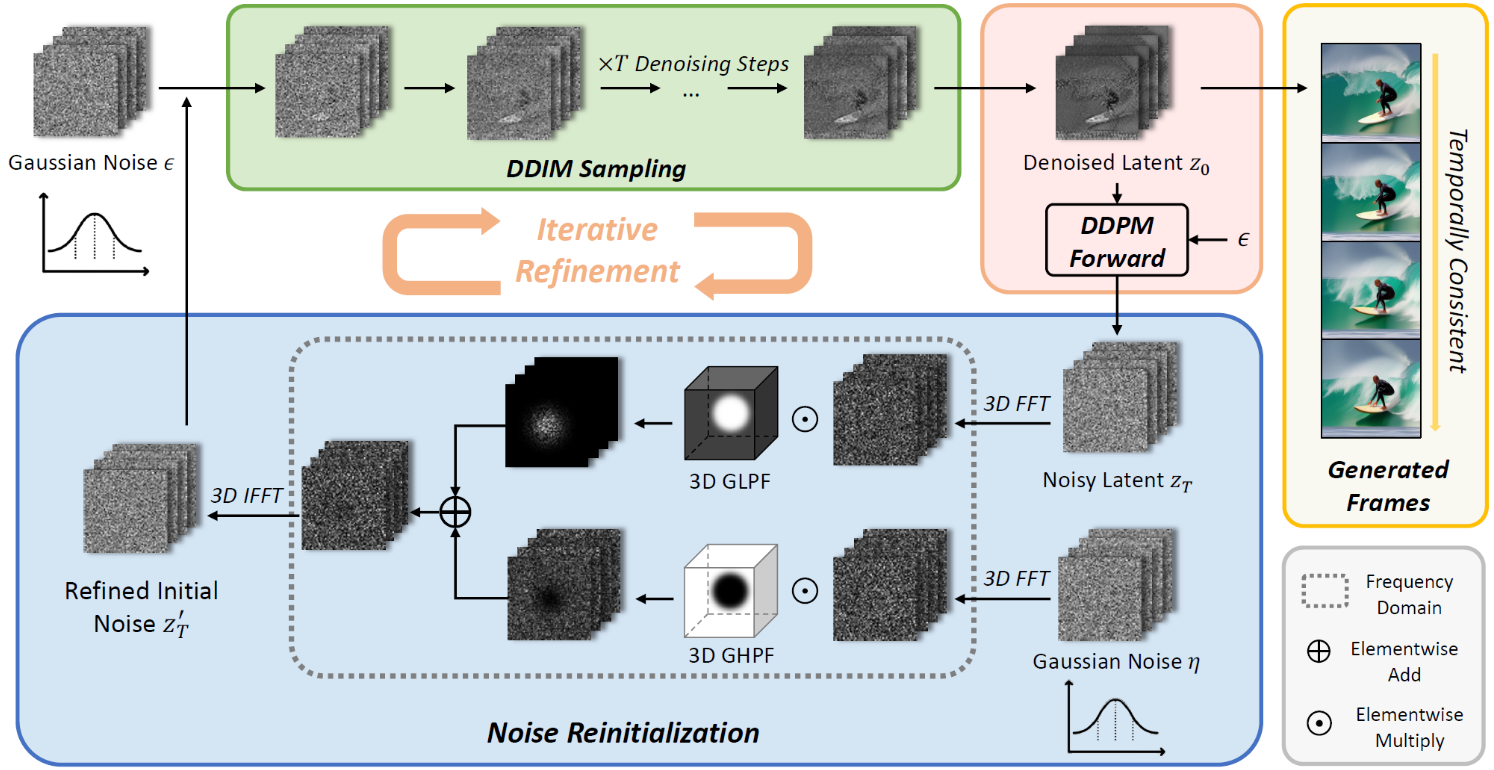
Initial noise at Inference: i.i.d Gaussian Noise, no temporal correlations

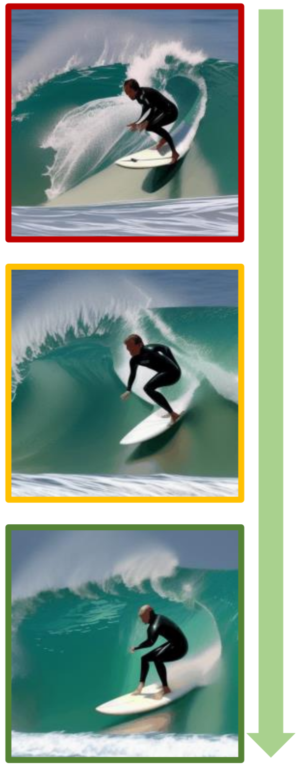
- **This causes an implicit *training-inference gap*:**
  - At training, the initial noise contain temporal correlations at low-frequency band
  - While at inference, the initial noise is pure Gaussian White Noise, lacking temporal correlations

FreeInit

We propose a training-free approach – FreeInit, to bridge this gap:

- the initial noise at inference is iteratively refined towards the training distribution, gradually enhancing the generation quality



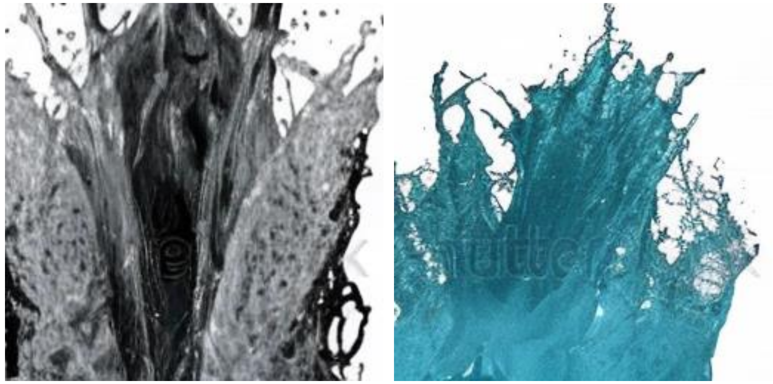


Gradually Enhanced Generation Quality

Visual Results

AnimateDiff | AnimateDiff + FreeInit | ModelScope | ModelScope + FreeInit | VideoCrafter | VideoCrafter + FreeInit

A panda standing on a surfboard in the ocean in sunset.



Splash of turquoise water in extreme slow motion, alpha channel included.

A cute raccoon playing guitar in a boat on the ocean

Vampire makeup face of beautiful girl, red contact lenses.

An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with umbrellas

Snow rocky mountains peaks canyon. snow blanketed rocky mountains surround and shadow deep canyons. The canyons twist and bend through the high elevated mountain peaks.

FreeInit can be readily applied to various text-to-video models, effectively improving temporal consistency and visual appearance

Visual Results

FreeNoise

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Tuning-Free Longer Video Diffusion via Noise Rescheduling

✅ totally no tuning ✅ less than 20% extra time ✅ support 512 frames

Motivation

- **Directly generating longer videos leads to poor quality**

Training-inference: The model is trained on 16 frames, but is required to generate 64 frames.

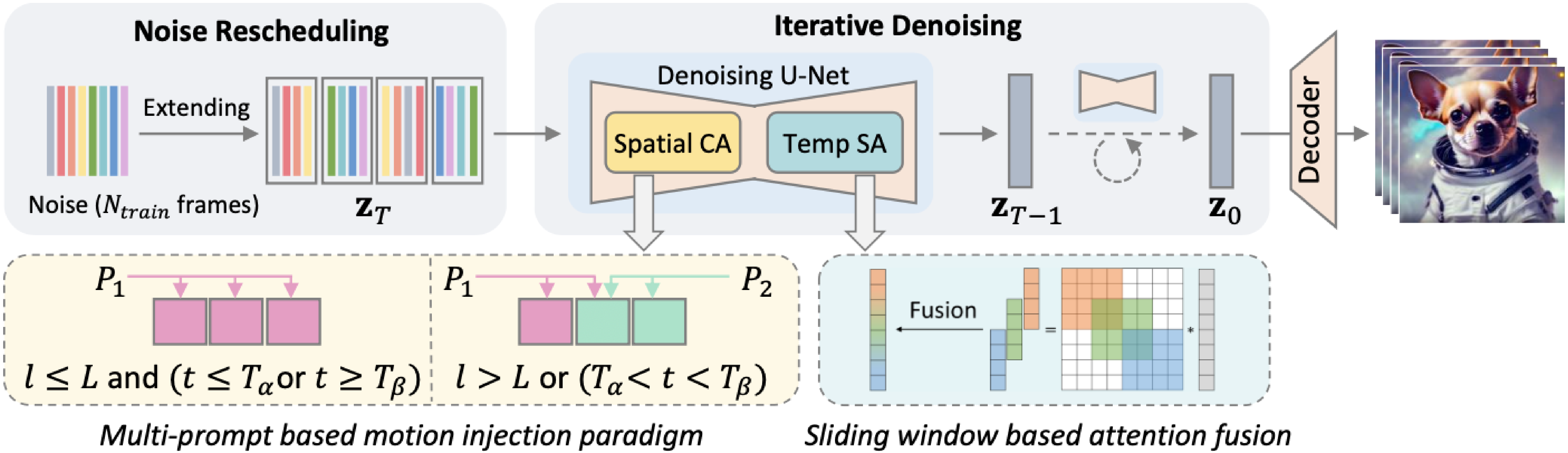
Direct 16 Frames | Direct 64 Frames


"A chihuahua in astronaut suit floating in space, cinematic lighting, glow effect"

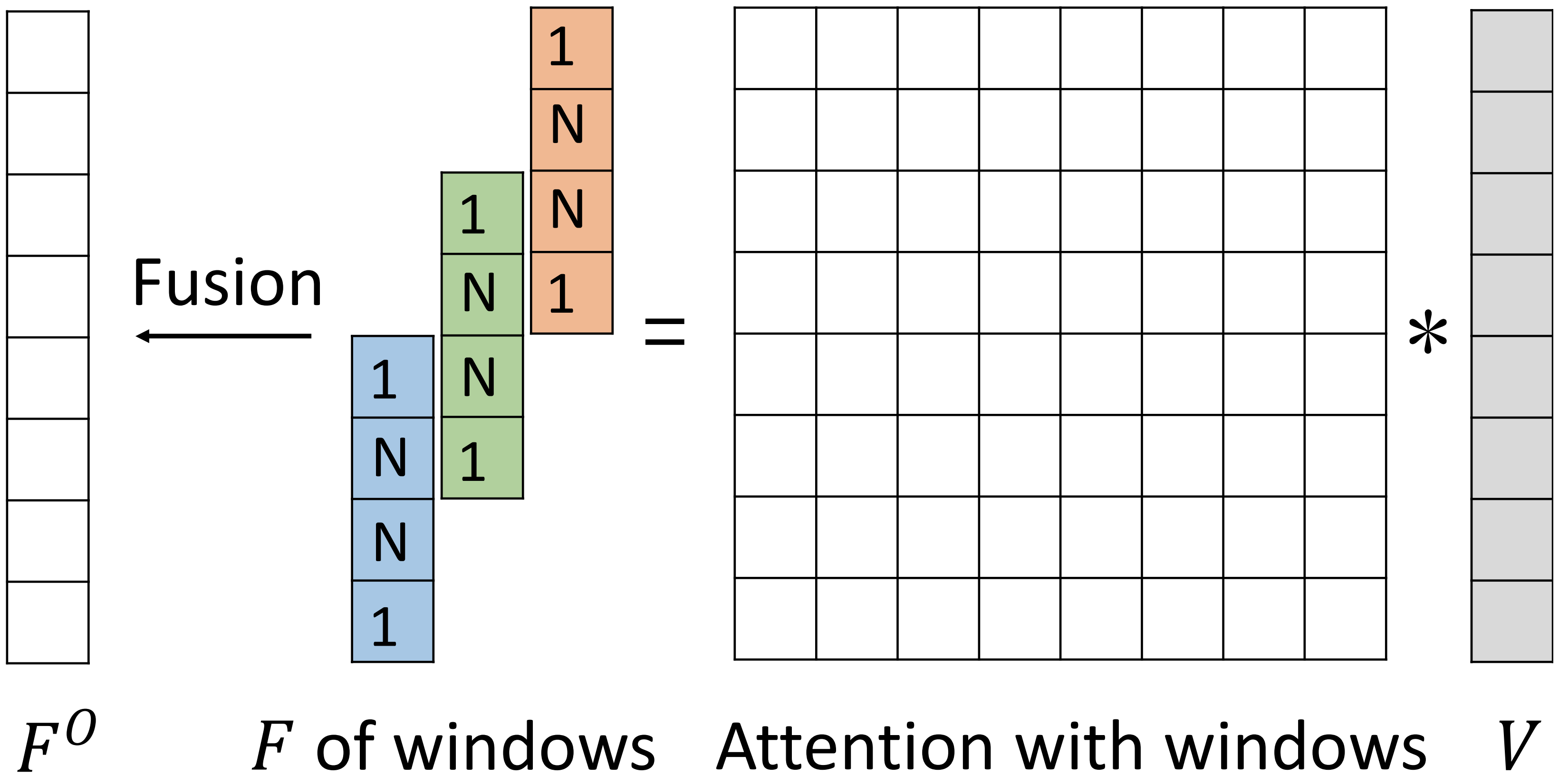

"A video of milk pouring over strawberries, blueberries, and blackberries. "

Method Overview

- **Core Designs:**
  - Local Window Fusion (for quality)
  - Noise Rescheduling (for consistency)
  - Motion Injection (for multi-prompt)



Local Window Fusion

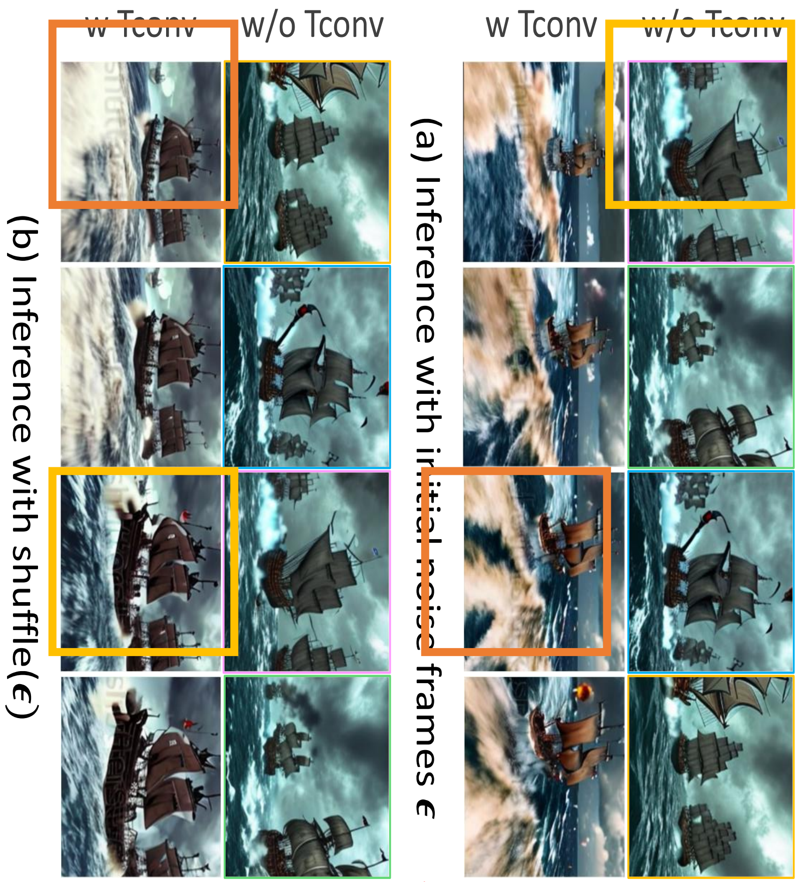


Only apply to temporal attention, negligible additional costs

Noise Rescheduling

(a) Inference with ϵ_1 (b) Inference with $[\epsilon_1, \epsilon_2]$

(c) Inference with ϵ_2 (d) Sliding window inference with $[\epsilon_1, \epsilon_2]$



Observations:

- New random noises bring a significantly different video.
- Temporal attention module is order-independent.
- Temporal convolution module is order-dependent.

Solution:

- Rescheduling Noise bans the influence of temporal attention but preserves the influence of temporal convolution, introducing new content while maintaining the main subjects and scenes.

Motion Injection

$$\textbf{Motion Injection} := \begin{cases} \mathrm{Attn}_{\mathrm{cross}}\left(\widetilde{Q}, l_{\widetilde{K}}(\widetilde{P}), l_{\widetilde{V}}(\widetilde{P})\right), & \text{if } T_\alpha < t < T_\beta \text{ or } l > L, \\ \mathrm{Attn}_{\mathrm{cross}}(\widetilde{Q}, l_{\widetilde{K}}(P_1), l_{\widetilde{V}}(P_1)), & \text{otherwise} \end{cases}$$

GenL

Ours w/o Motion Injection

Ours

"An astronaut resting on a horse" → "… riding …"

Results

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

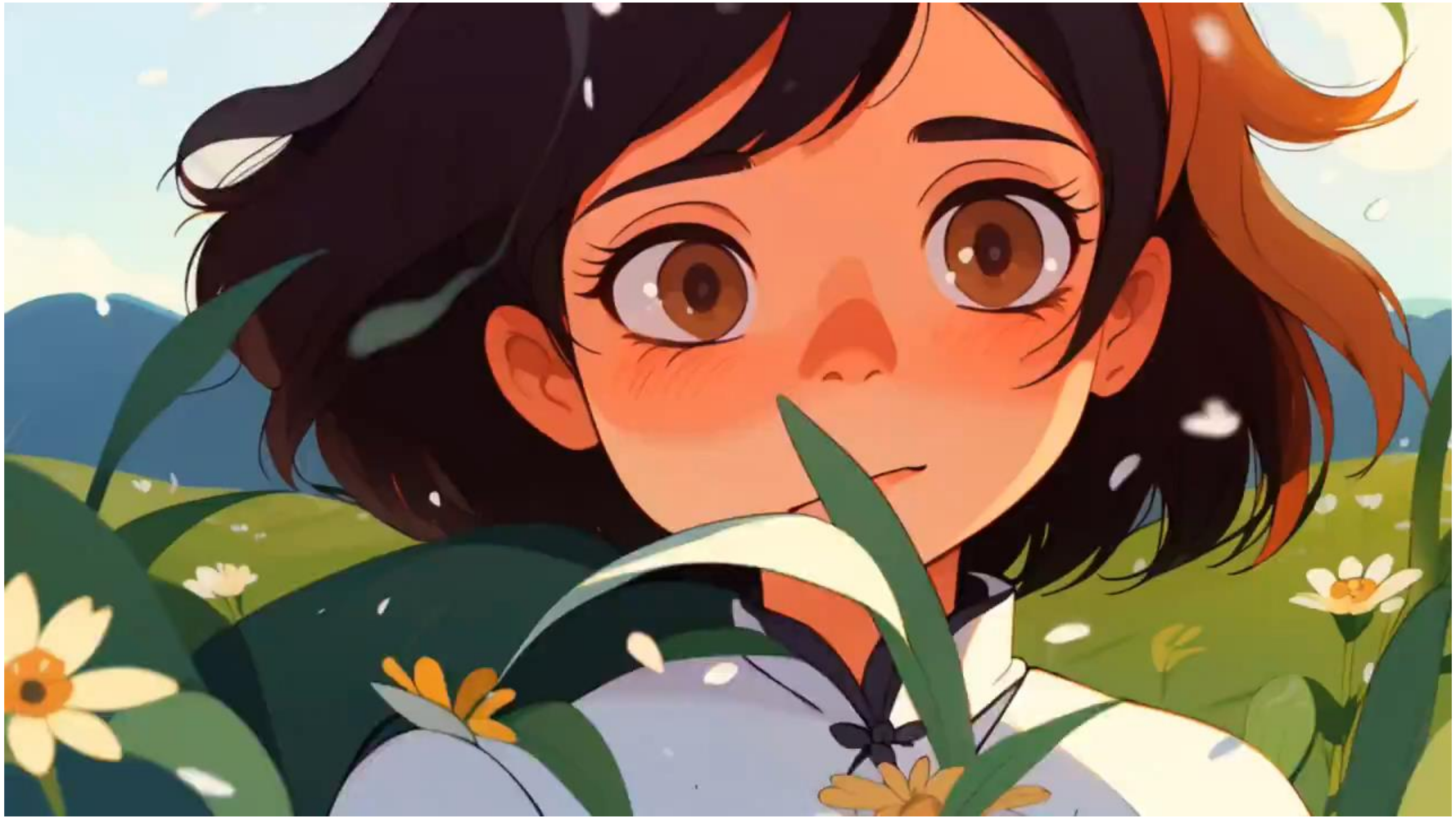
Direct	Sliding	GenL	Ours
			

"A chihuahua in astronaut suit floating in space, cinematic lighting, glow effect"

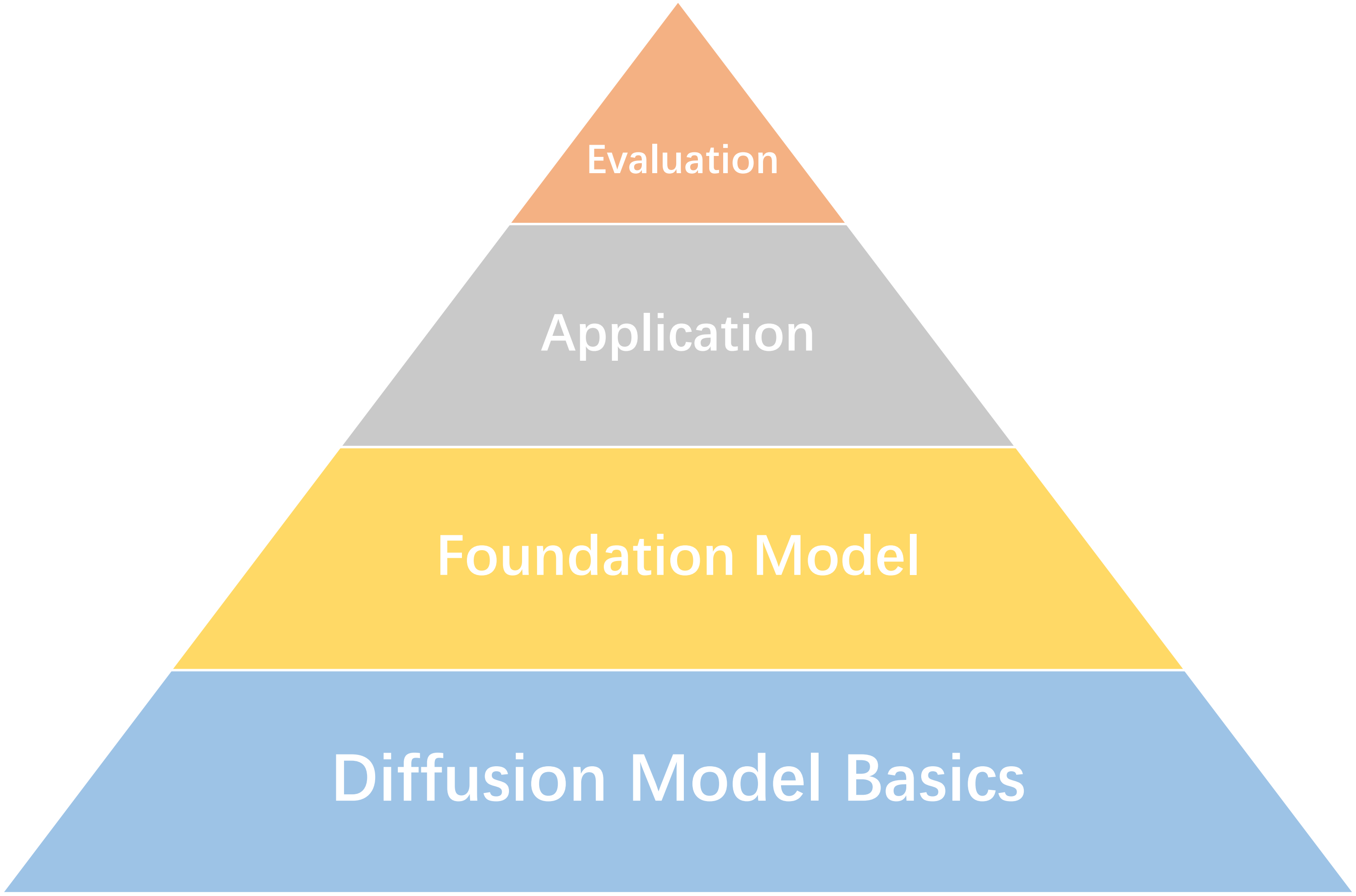
			

"A very happy fuzzy panda dressed as a chef eating pizza in the New York street food truck"

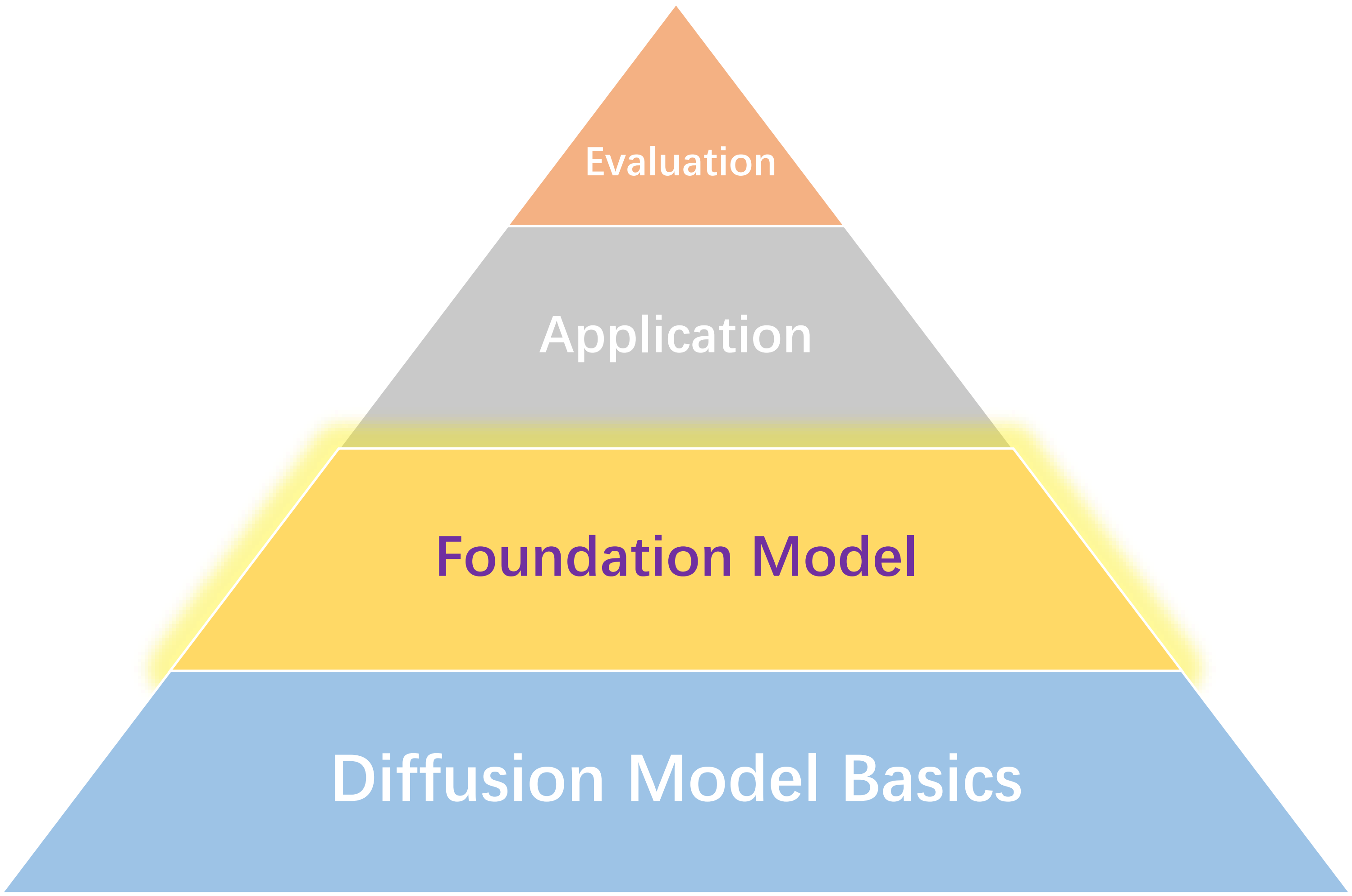
Results



Video Generation

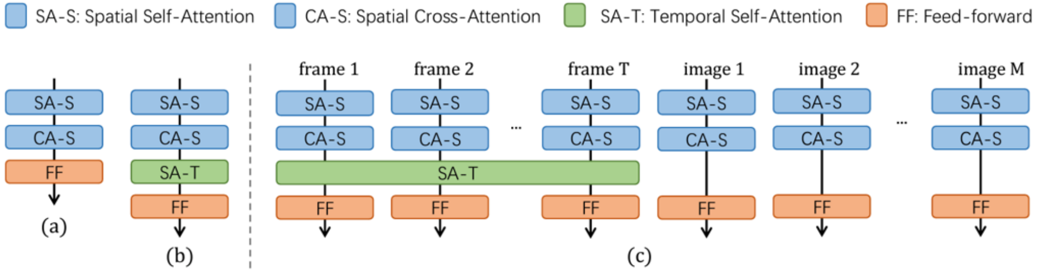


Video Generation



书生·筑梦 视频生成大模型

- 支持**故事性、多镜头**的视频生成大模型，具备**分钟级4K视频生成**能力
- 实现转场流畅、故事连贯、画质高清，在多维度评测指标中综合领先



模型架构

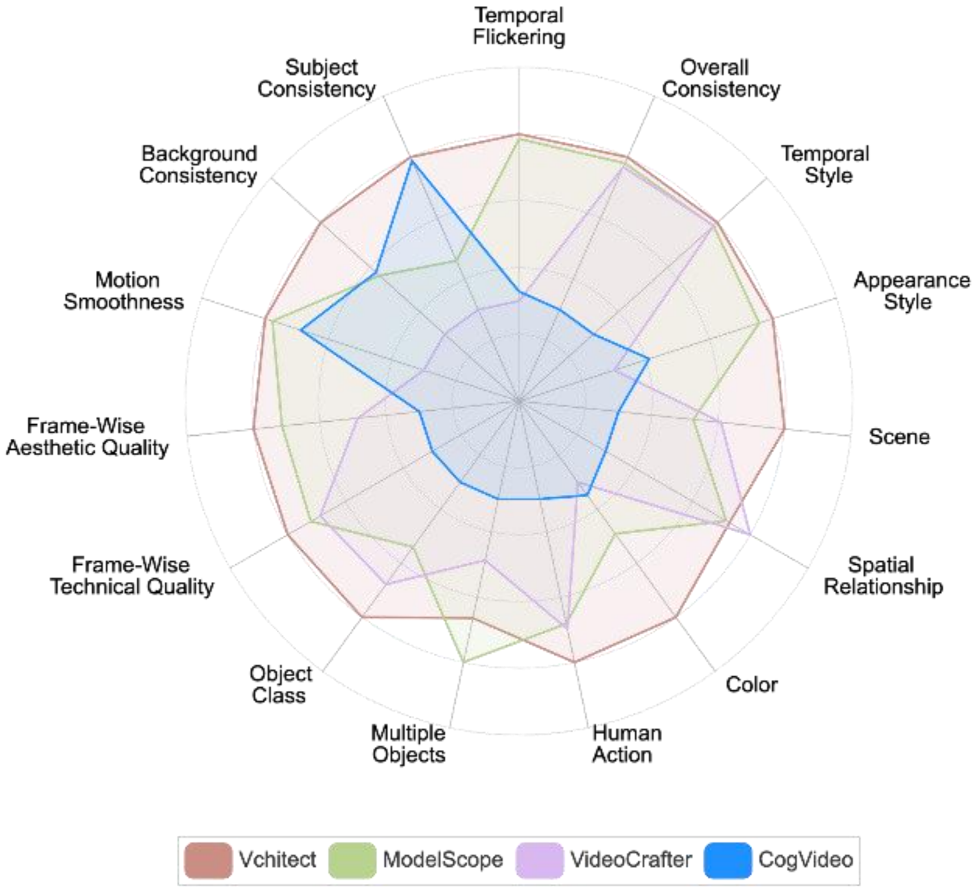
长视频生成

文生视频

图生视频（转场）

图生视频（驱动）

书生·筑梦 视频生成大模型



与开源模型能力对比

筑梦

Meta EmuVideo

筑梦

谷歌 Lumiere

与闭源模型能力对比

书生·筑梦 视频生成大模型

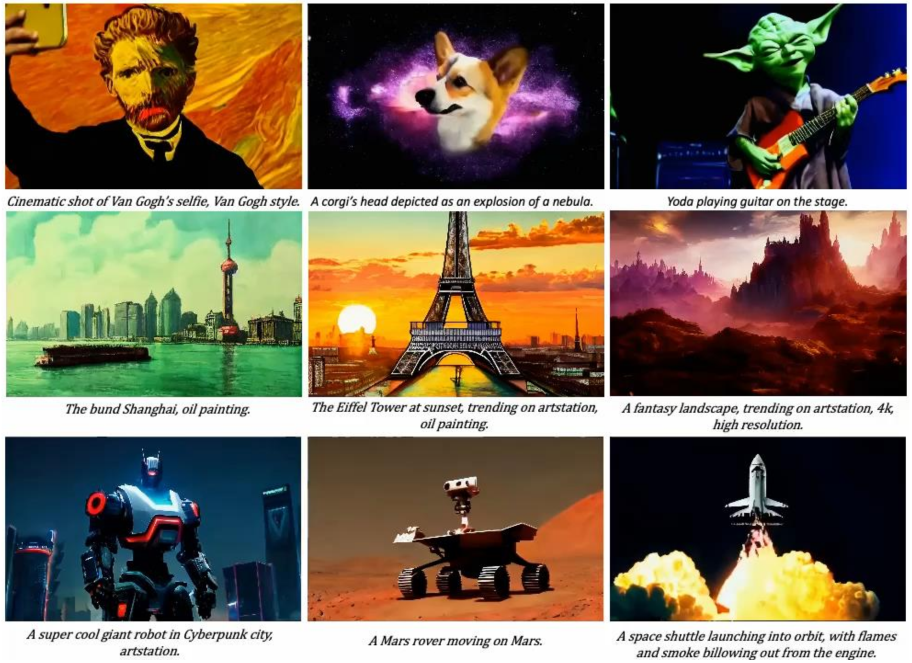
开源影响力

- **Github**: 2K star
- **Replicate**: 16K 次使用

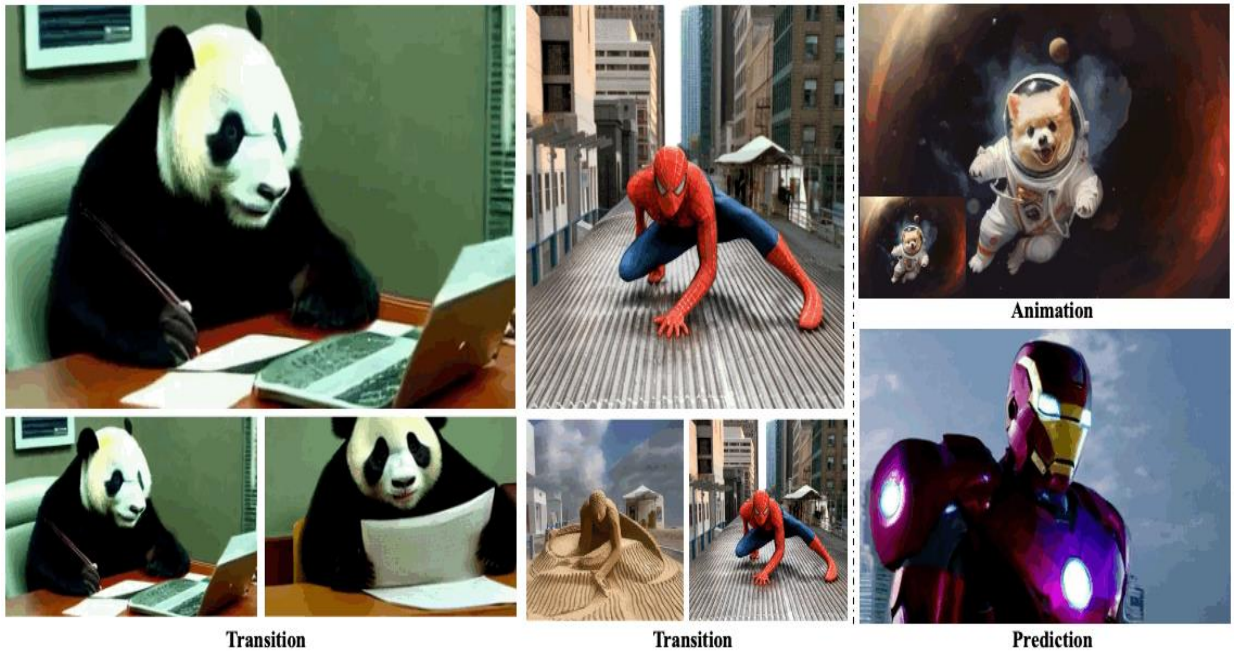
应用

- **央视听媒体大模型**
- **第一部AIGC动画“千秋诗颂”**

Vchitect: A Large-scale Video Generation System



LaVie [Wang, Chen, Ma *et al*., arXiv'23]
Text-to-video generation



SEINE [Chen, Wang *et al.*, arXiv'23]
Image-to-video generation

LaVie

High-quality Video Generation with Cascaded Latent Diffusion Models

Cinematic shot of Van Gogh's selfie, Van Gogh style.

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

Yoda playing guitar on the stage.

The bund Shanghai, oil painting.

The Eiffel Tower at sunset, trending on artstation, oil painting.

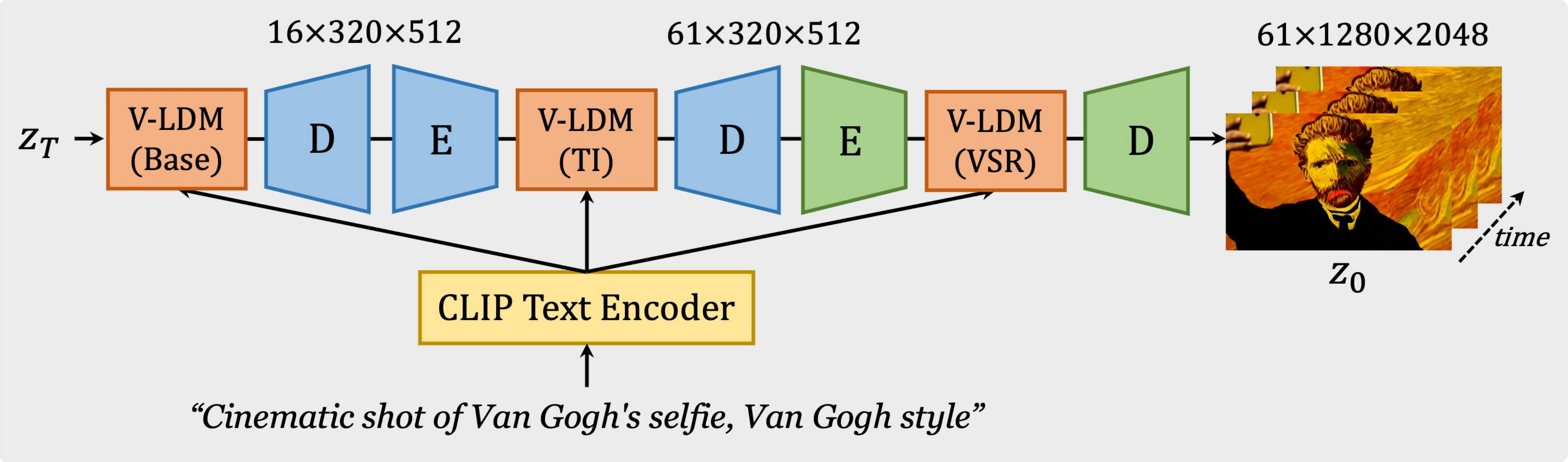
A fantasy landscape, trending on artstation, 4k, high resolution.

A super cool giant robot in Cyberpunk city, artstation.

A Mars rover moving on Mars.

A space shuttle launching into orbit, with flames and smoke billowing out from the engine.

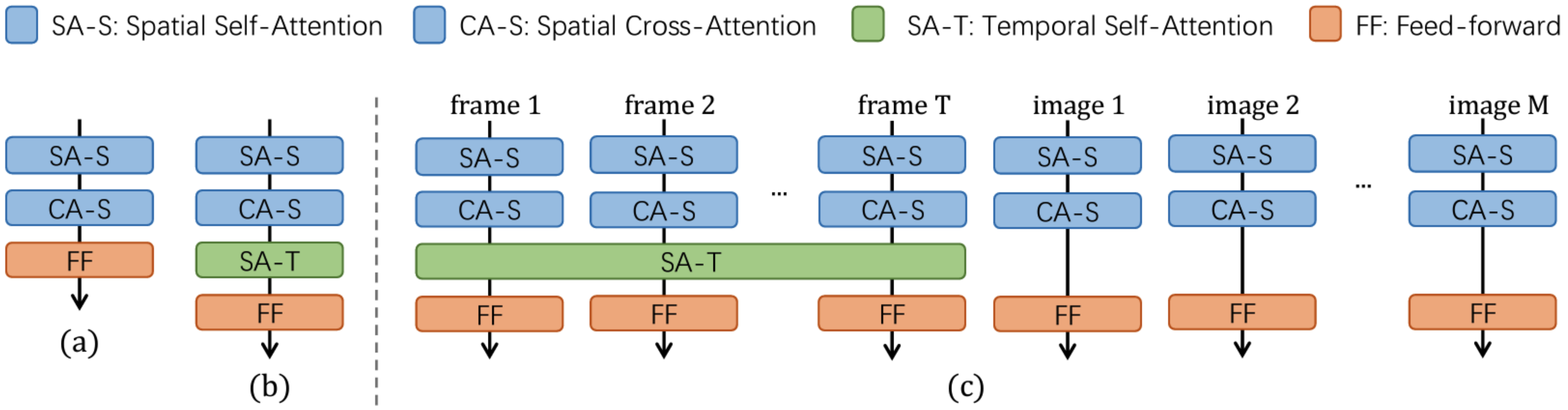
LaVie – Model Design



A cascaded video generation system:

- Base model → 320x512 resolution, 16 frames
- Interpolation model → 320x512, 61 frames
- Super-resolution model → 1280x2048, 61frames
- CLIP Text Encoder

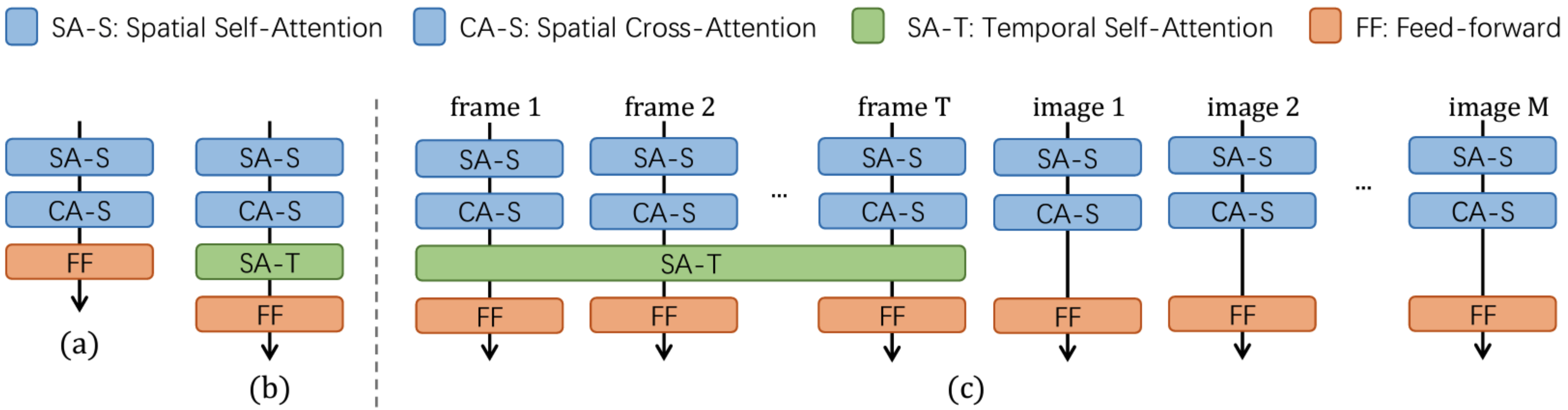
LaVie – Architecture



Pre-trained Stable Diffusion:

- 2D UNet → 3D UNet
- Involving temporal self-attention + relative positional encoding

LaVie – Learning



- Pre-trained Stable Diffusion

  1. Fast convergence

- Joint image-video fine-tuning

  1. Prevent catastrophic forgetting
  2. More creativity, diversity and better visual quality

- Learning objective (image-video joint training):

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathcal{E}(\mathbf{v}_t), t, c_V)\|_2^2\right] + \alpha * \mathbb{E}\left[\|\epsilon - \epsilon_\theta(\mathcal{E}(\mathbf{x}_t), t, c_I)\|_2^2\right]$$

LaVie – Data

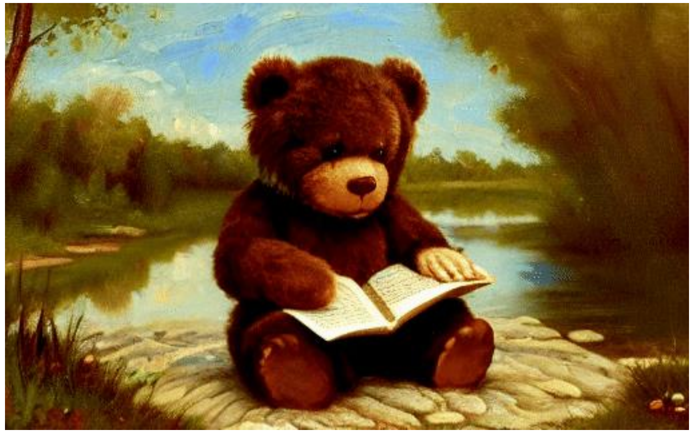
Videos from Vimdeo25M dataset

1. LAION-5B dataset (large-scale image dataset)
2. WebVid10M (large-scale text-video dataset, ~320 x 500, with watermark)
3. Vimeo25M (large-scale text-video dataset)
    1. More detailed captions (provided by VideoChat)
    2. Higher resolution, 1080p, better visual quality
    3. Better aesthetics

LaVie – More results

Two teddy bears playing poker under water

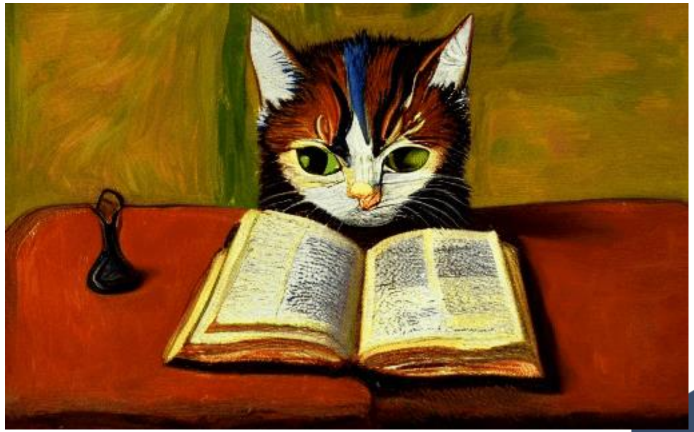
a teddy bears skateboarding under water



a teddy bears reading a book in the park, oil painting style

Elon Musk standing besides a rocket

Iron Man flying in the sky

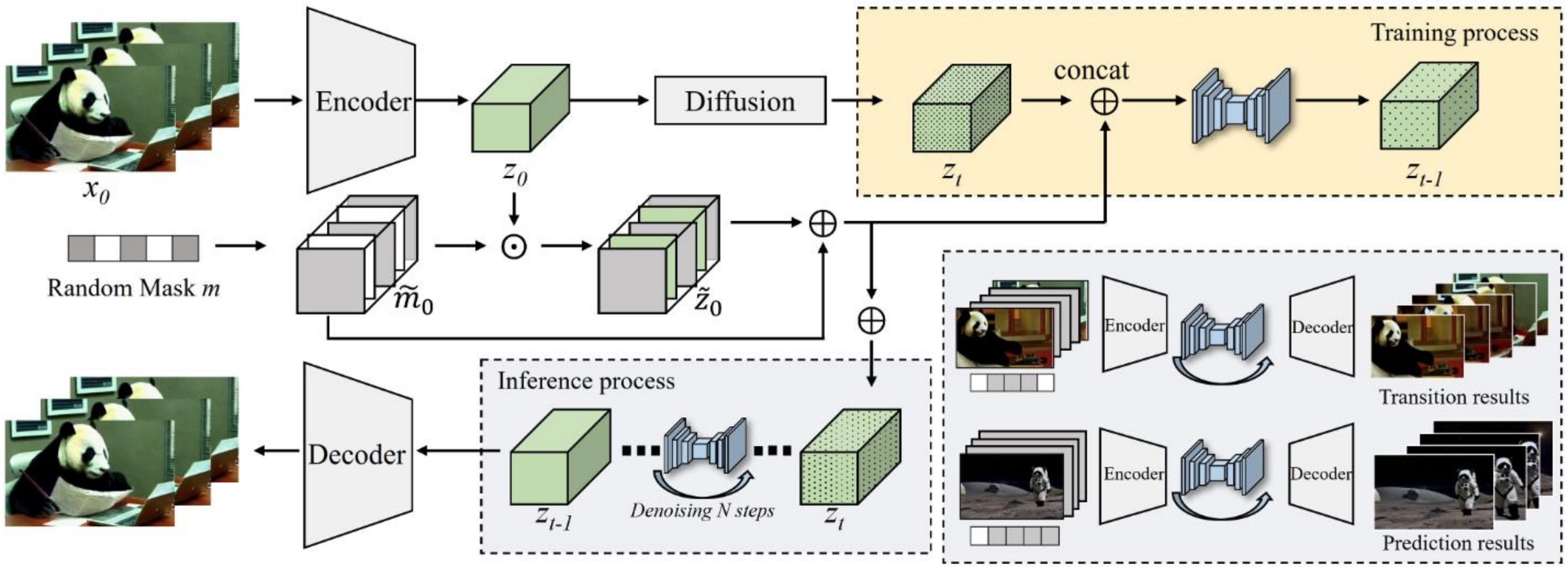


a cat reading a book, Van Gogh style

SEINE

Short-to-Long Video Diffusion Model for Generative Transition and Prediction

SEINE – Architecture & Learning



<u>Training</u>

1. LaVie pretrained
2. Image-conditioned generation
3. Random masks as extra input conditions

<u>Inference:</u>

Different masks →

Transition, Animation, Prediction

SEINE – More results

Image-to-video generation

Transition

Story-based Long Video Generation (LaVie + SEINE)

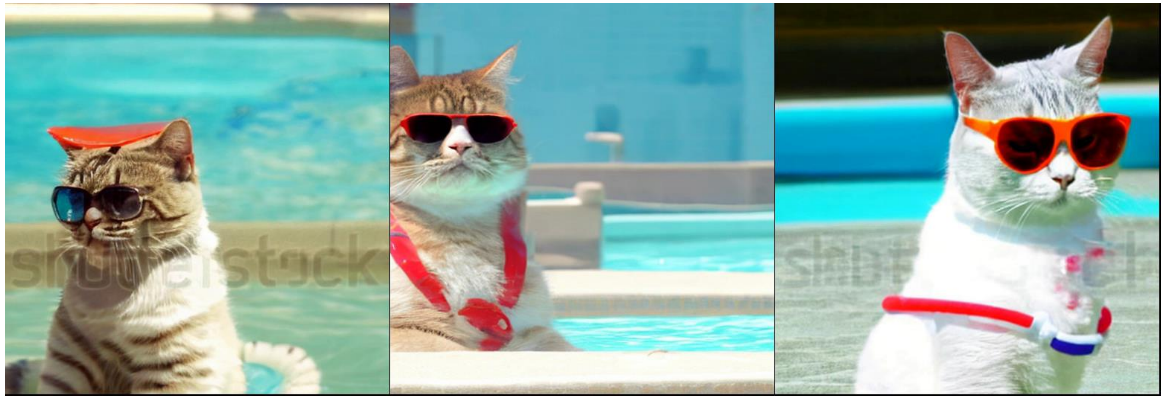
Latte: Latent Diffusion Transformer

A diffusion transformer for general video generation

A dog in astronaut suit and sunglasses floating in space.

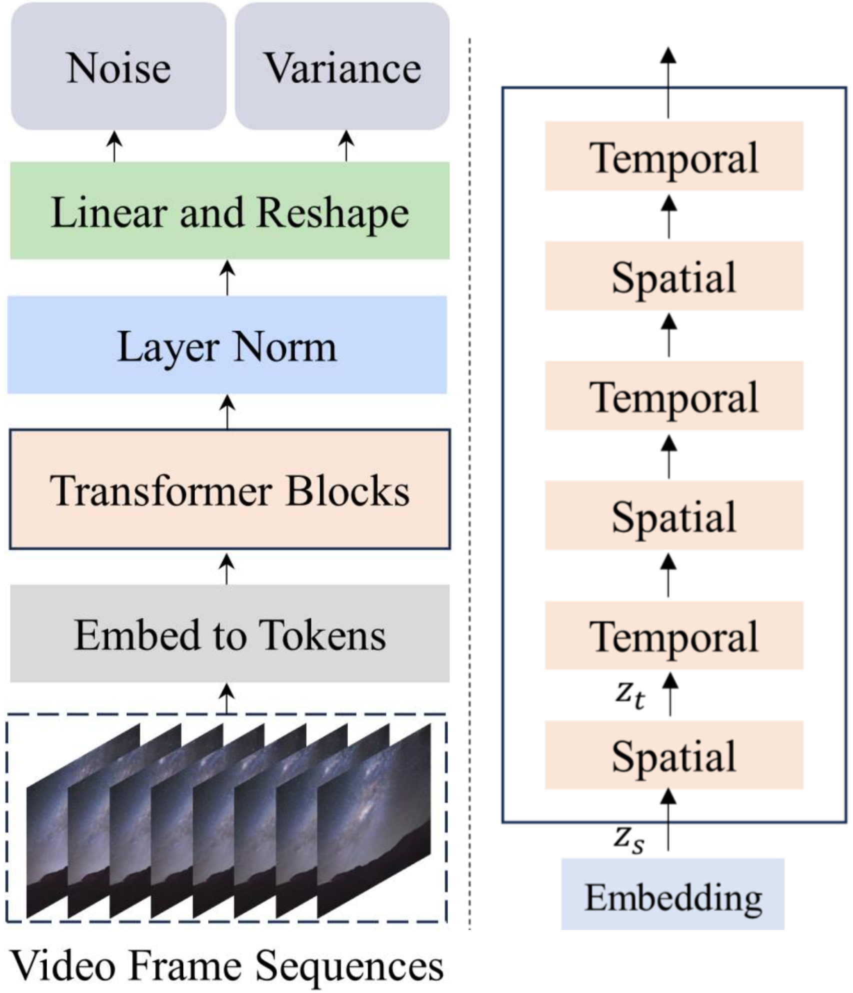
Yellow and black tropical fish dart through the sea.

Yellow and black tropical fish dart through the sea.



a cat wearing sunglasses and working as a lifeguard at pool.

Latte: Latent Diffusion Transformer



Latte architecture

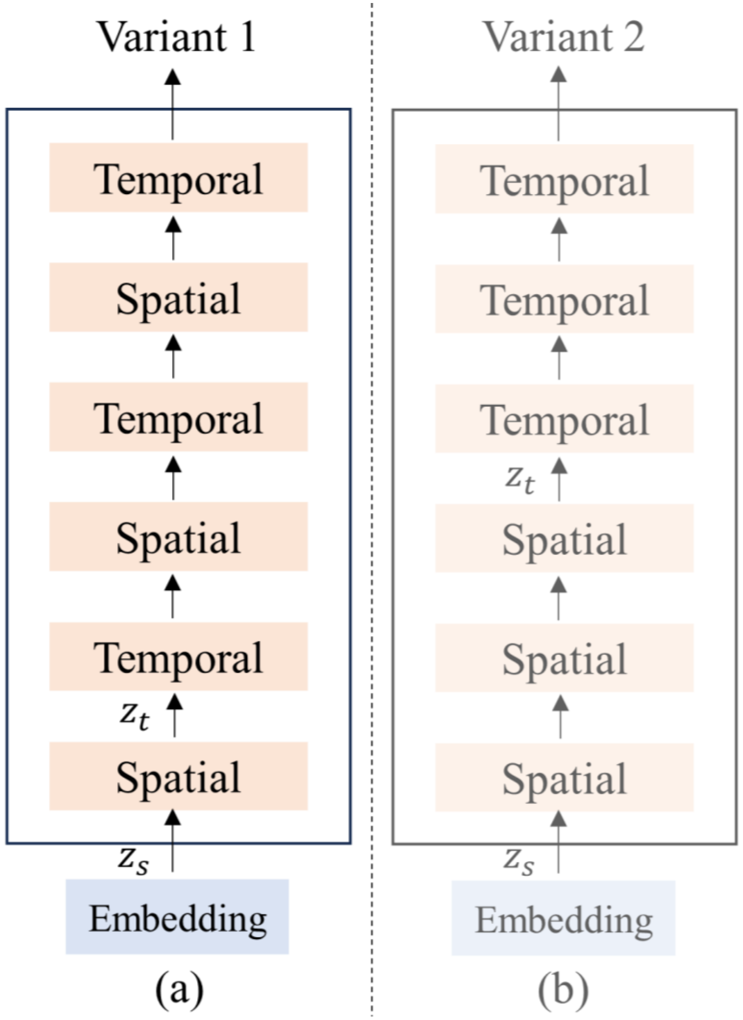
We introduce:

1. Model architecture designs

2. Transformer designs

3. Best practices in model and training

Latte – Model design

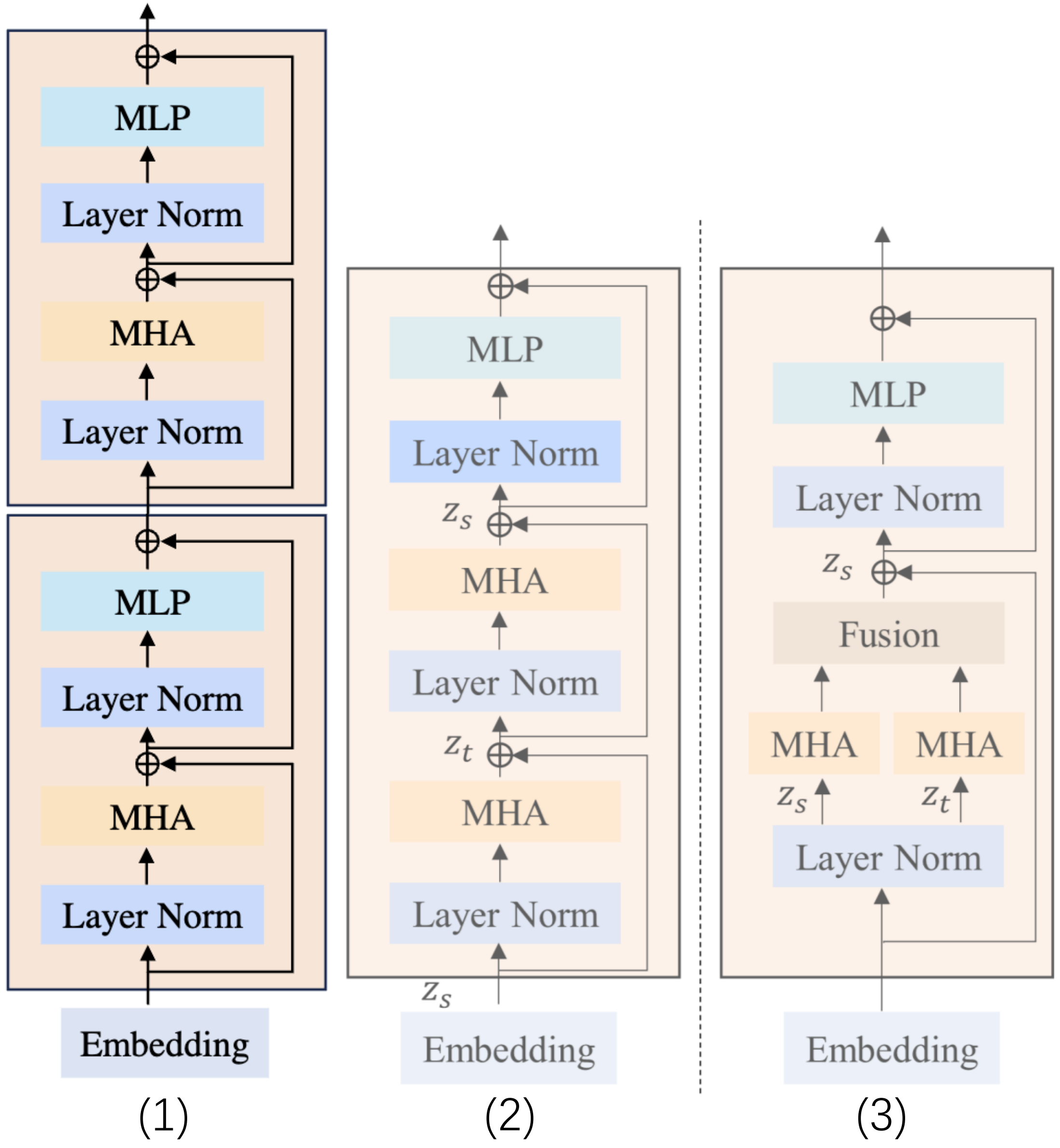


Variant 1:
(Spatial + Temporal) x N blocks

Our choice

Variant 2:
(Spatial x N/2 blocks) + (Temporal x N/2 blocks)

Latte – Transformer block design



1. Separate spatial & temporal transformer blocks
- Spatial block
- Temporal block

Our choice

2. Joint spatio-temporal transformer block
- cascaded spatial and temporal attentions

3. Joint spatio-temporal transformer block
- parallel spatial and temporal attentions

Latte – Best Practice Design

We systematically analyze:

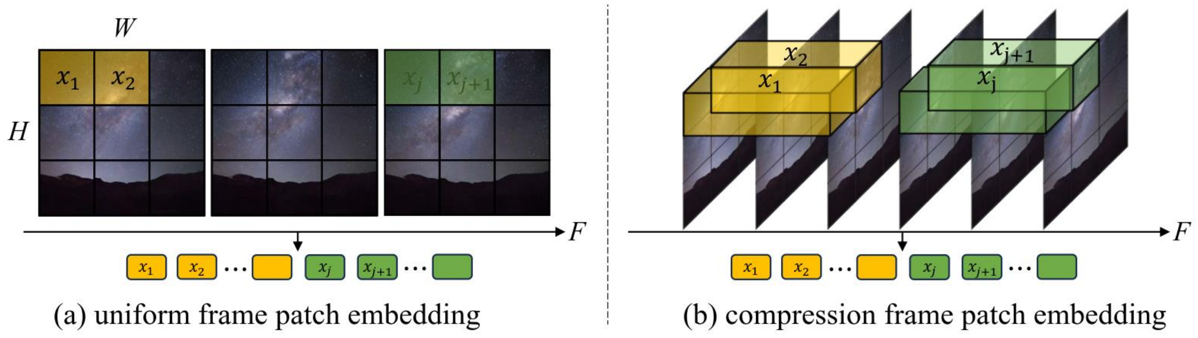
(a) Video sampling interval (rate 2, **3**, 4, 8, 16)

(b) Temporal positional embedding (**absolute** or relative)

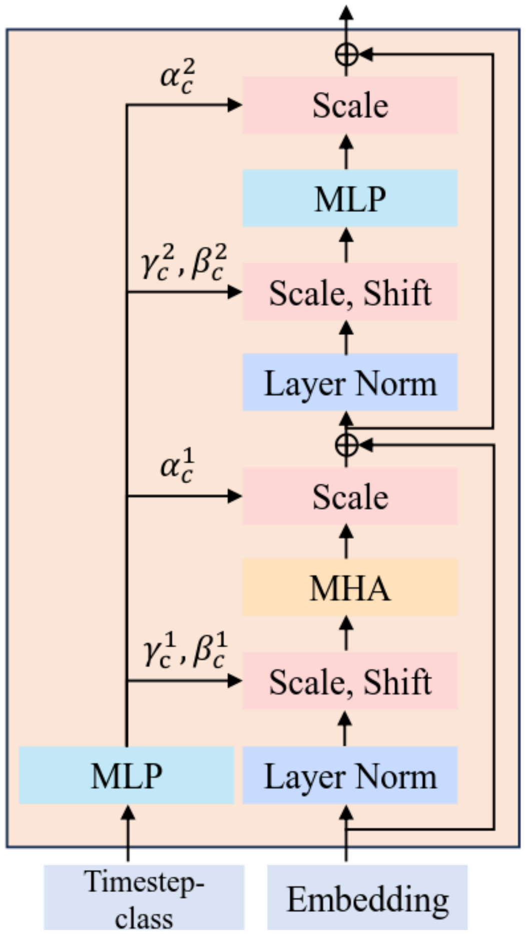
(c) ImageNet pretraining is **NOT NECESSARY**

(d) Video clip patch embedding (**uniform** or compression)

(f) Timestep-class information injection (**S-AdaLN** or all-tokens)



Video clip patch embedding



Timestep-class information injection

Latte – Quantitative analysis

Method	IS ↑	FID ↓
MoCoGAN	10.09	23.97
VideoGPT	12.61	22.7
MoCoGAN-HD	23.39	7.12
DIGAN	23.16	19.1
StyleGAN-V	23.94	9.445
PVDM	60.55	29.76
Latte (ours)	68.53	5.02
Latte+IMG (ours)	**73.31**	**3.87**

Frame-level quality comparison

Method	FaceForensics	SkyTimelapse	UCF101	Taichi-HD
MoCoGAN	124.7	206.6	2886.9	-
VideoGPT	185.9	222.7	2880.6	-
MoCoGAN-HD	111.8	164.1	1729.6	128.1
DIGAN	62.5	83.11	1630.2	156.7
StyleGAN-V	47.41	79.52	1431.0	-
PVDM	355.92	75.48	1141.9	540.2
MoStGAN-V	39.70	65.30	1380.3	-
LVDM	-	95.20	372.0	99.0
Latte (ours)	34.00	59.82	477.97	159.60
Latte+IMG (ours)	**27.08**	**42.67**	**333.61**	**97.09**

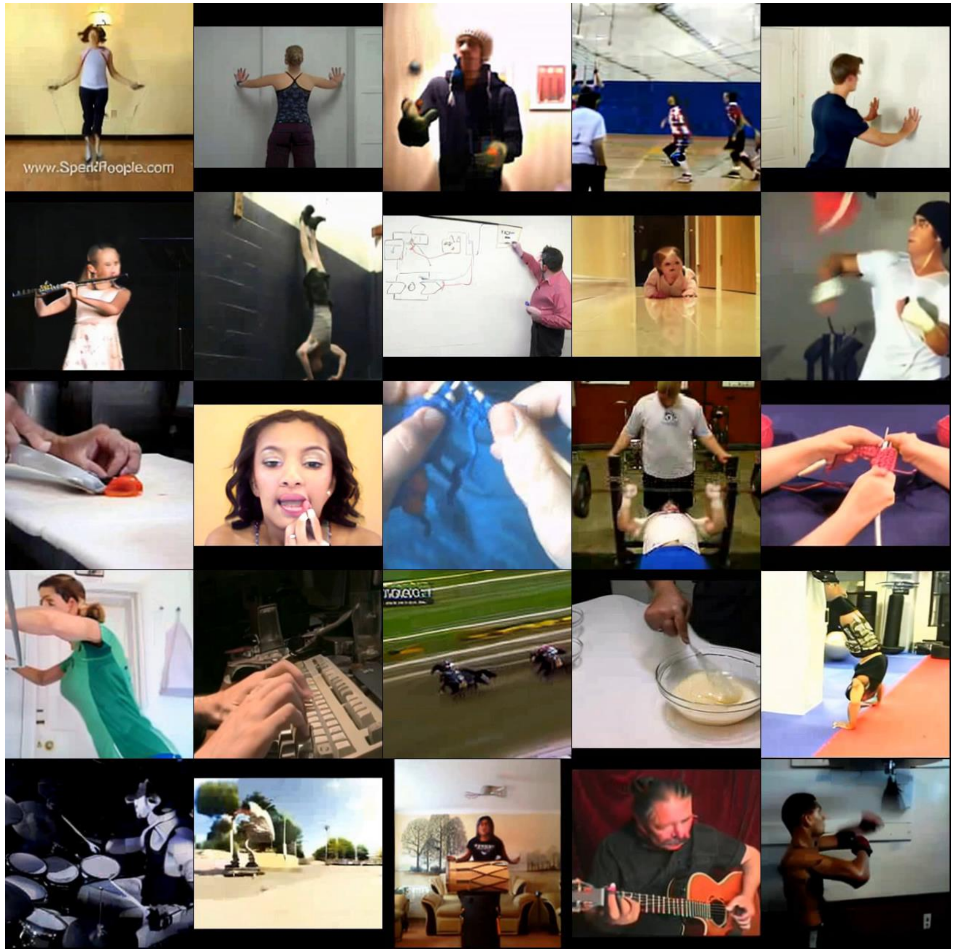
Video-level quality comparison

Latte – Results

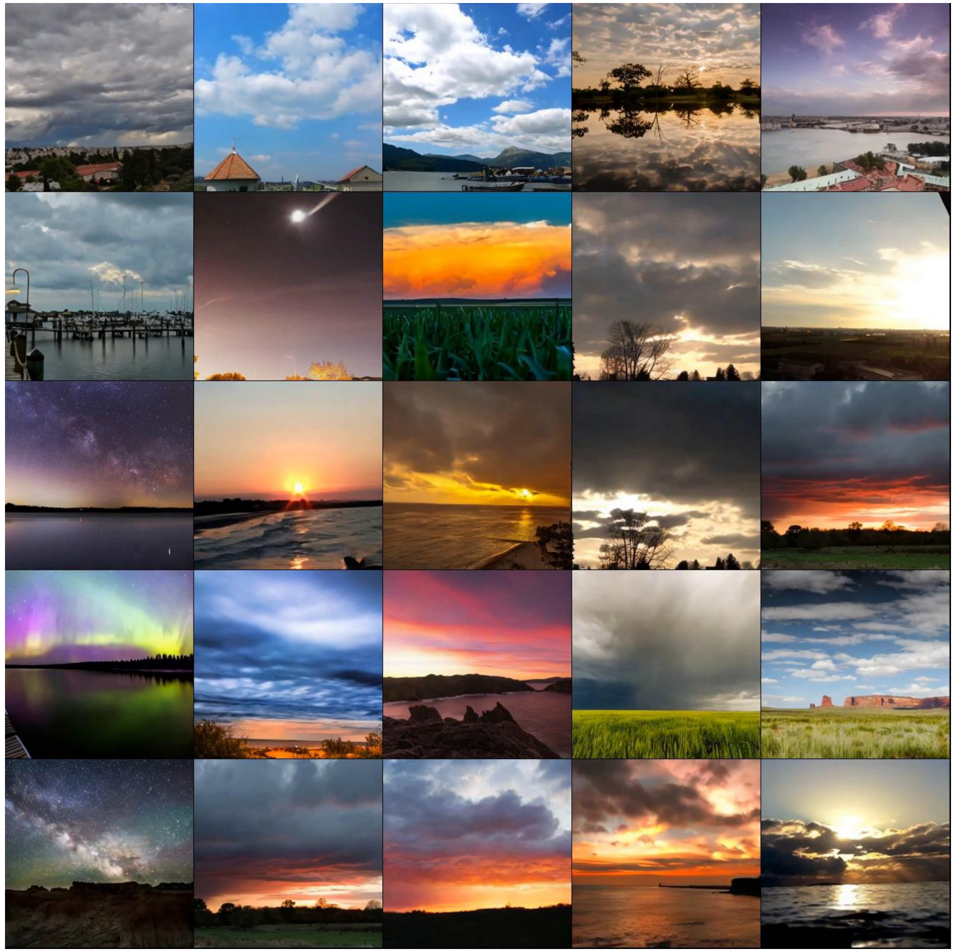
FaceForensics

Taichi-HD

Latte – Results



UCF101

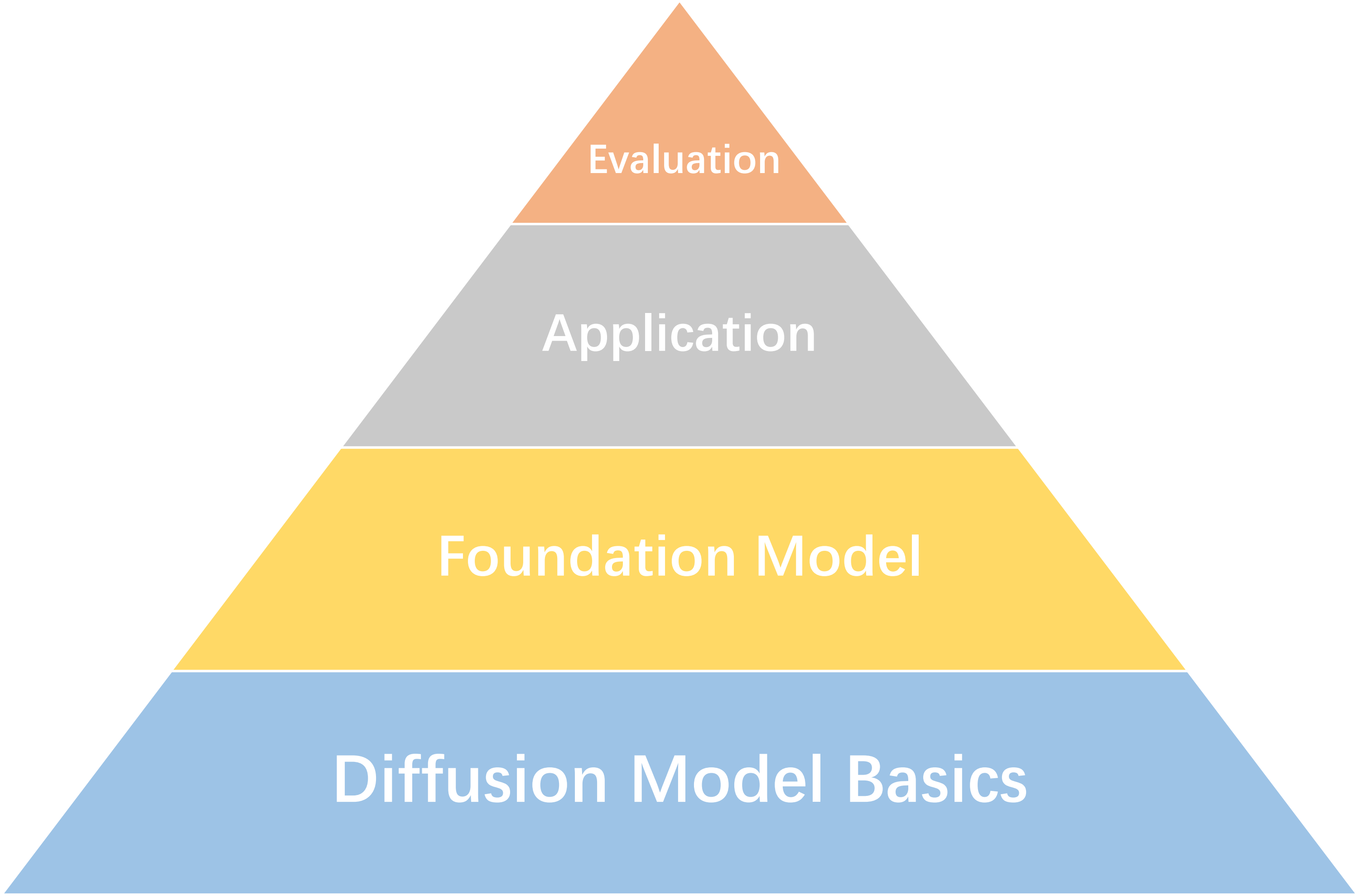


SkyTimelapse

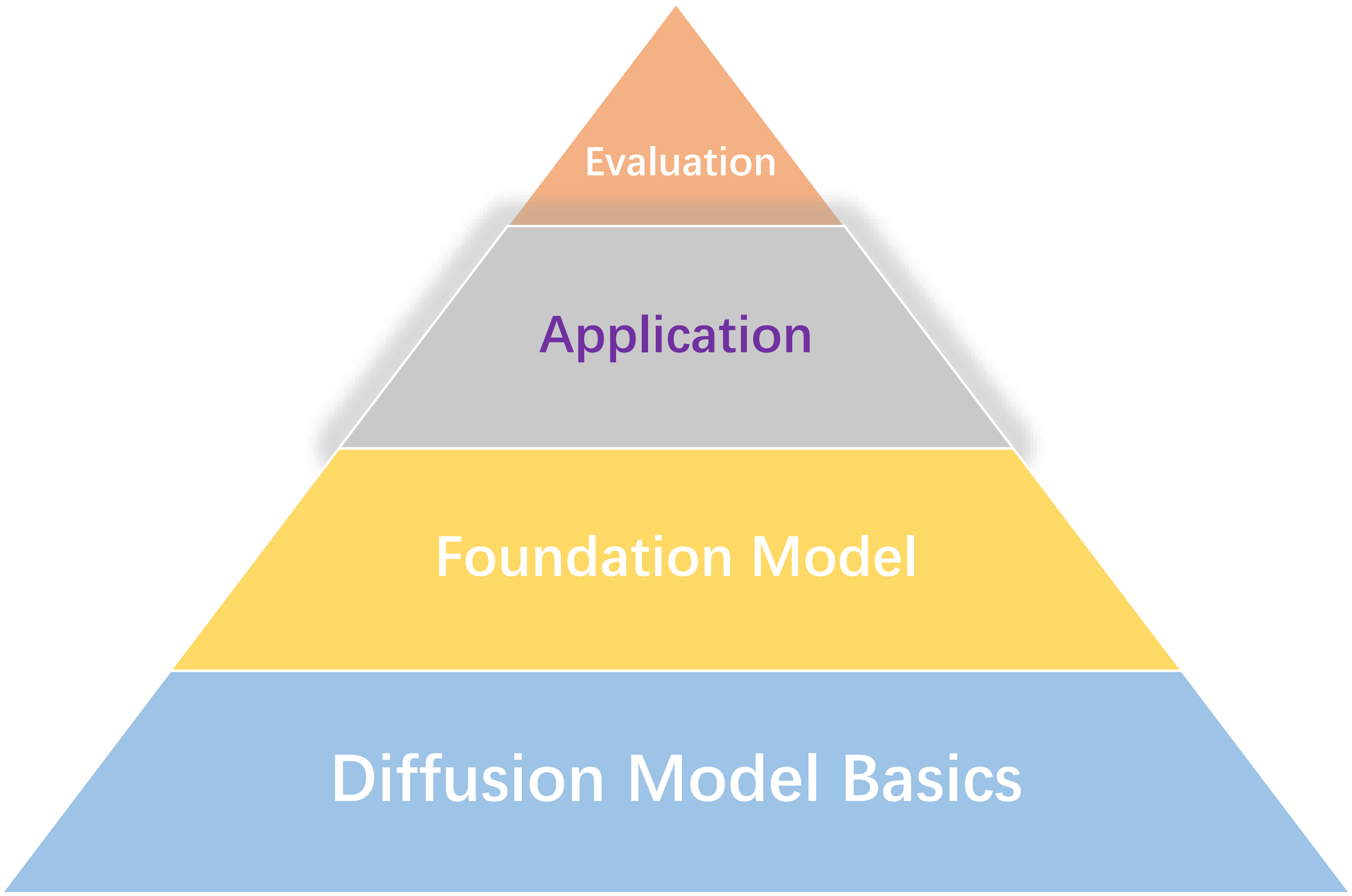
Vchitect Foundation Models

Vchitect Foundation Models

Video Generation



Video Generation



VideoBooth

Diffusion-based Video Generation with Image Prompts

- **Merely using text prompts is not enough to customize video generation**
  - It is hard to enumerate all desired attributes
  - The model is incapable of capturing all attributes accurately from texts

VideoBooth

A photo of a dog

Dog

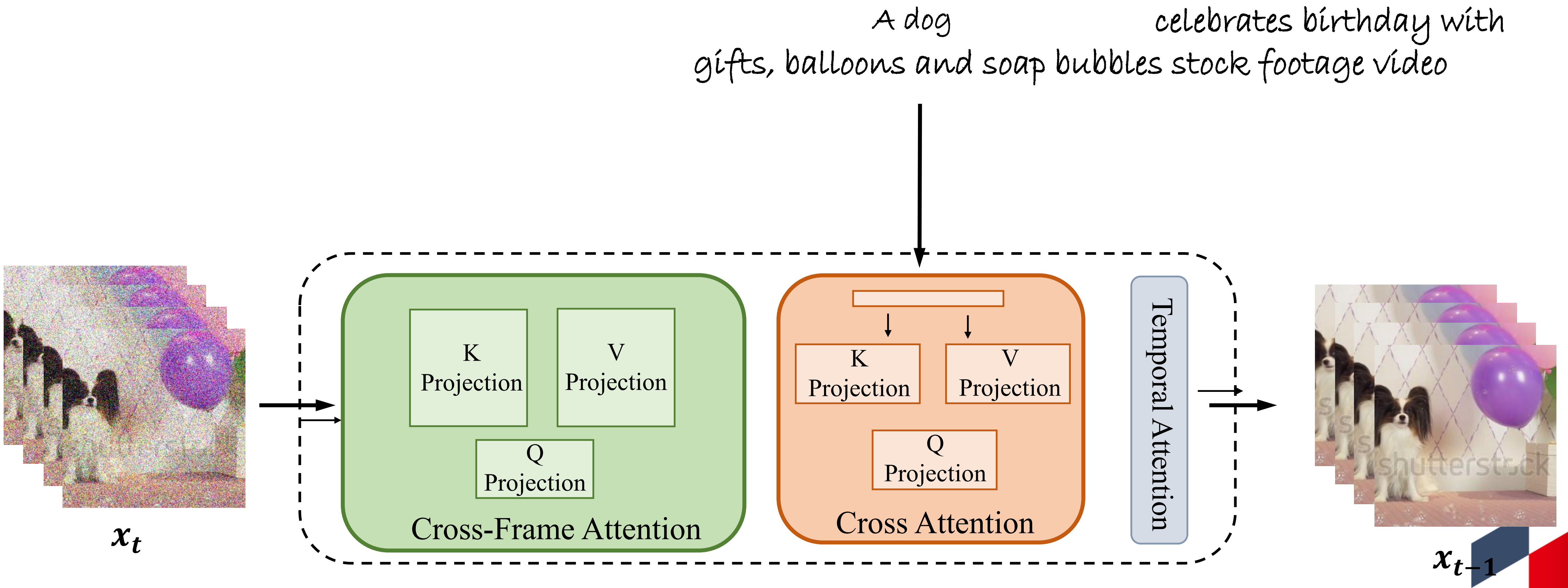
Dog drinking from bowl of water

Dog swimming in lake happily

Dog in park

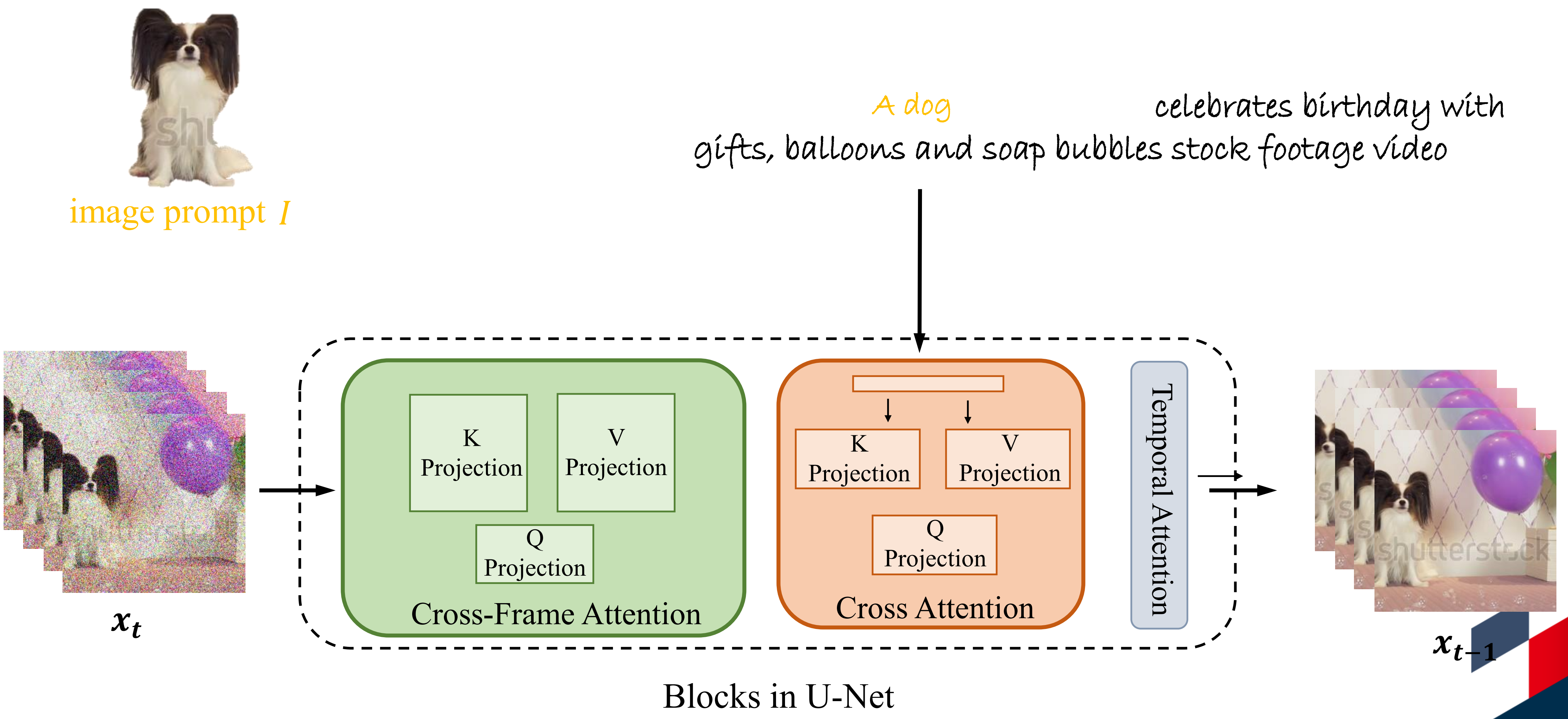
Portrait of a dog, looks out the car window

VideoBooth - Method

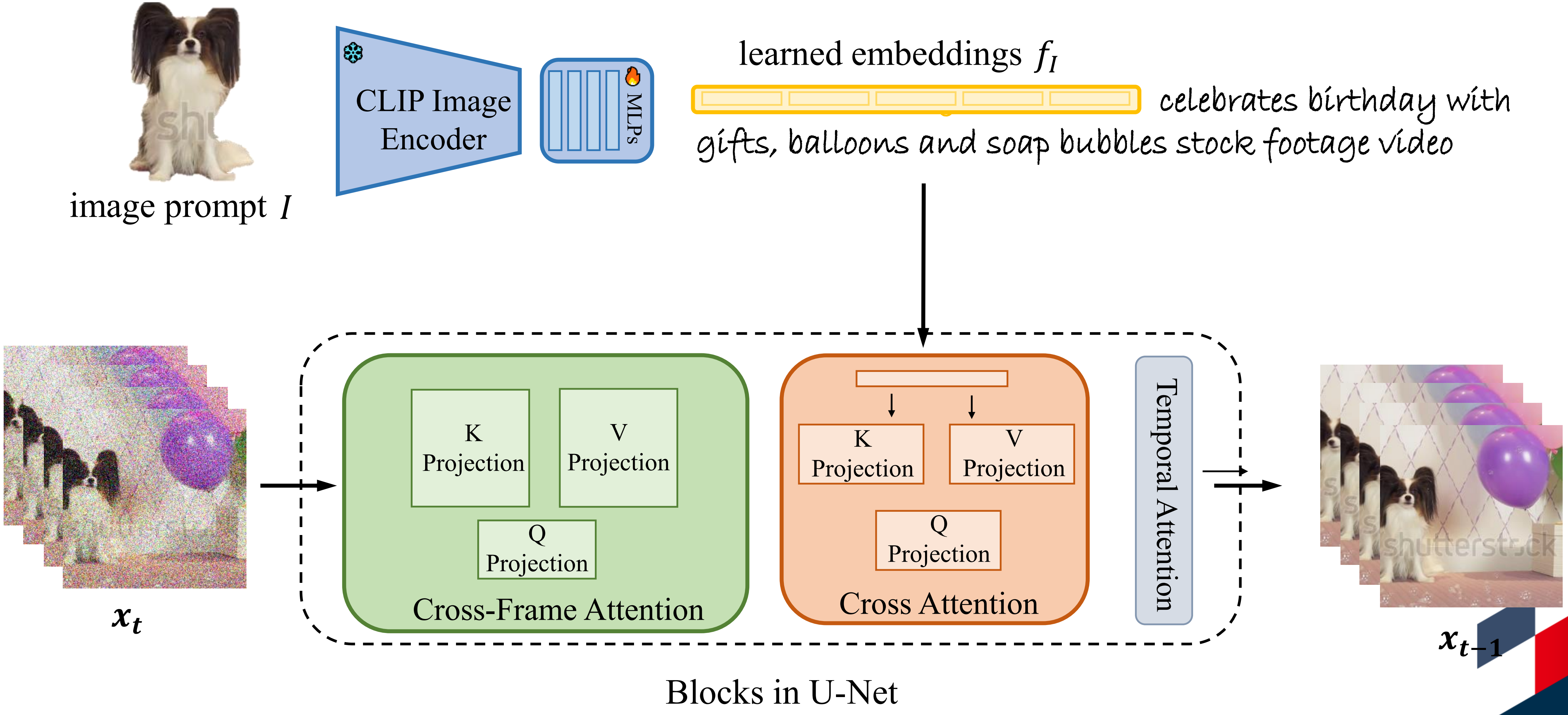


Blocks in U-Net

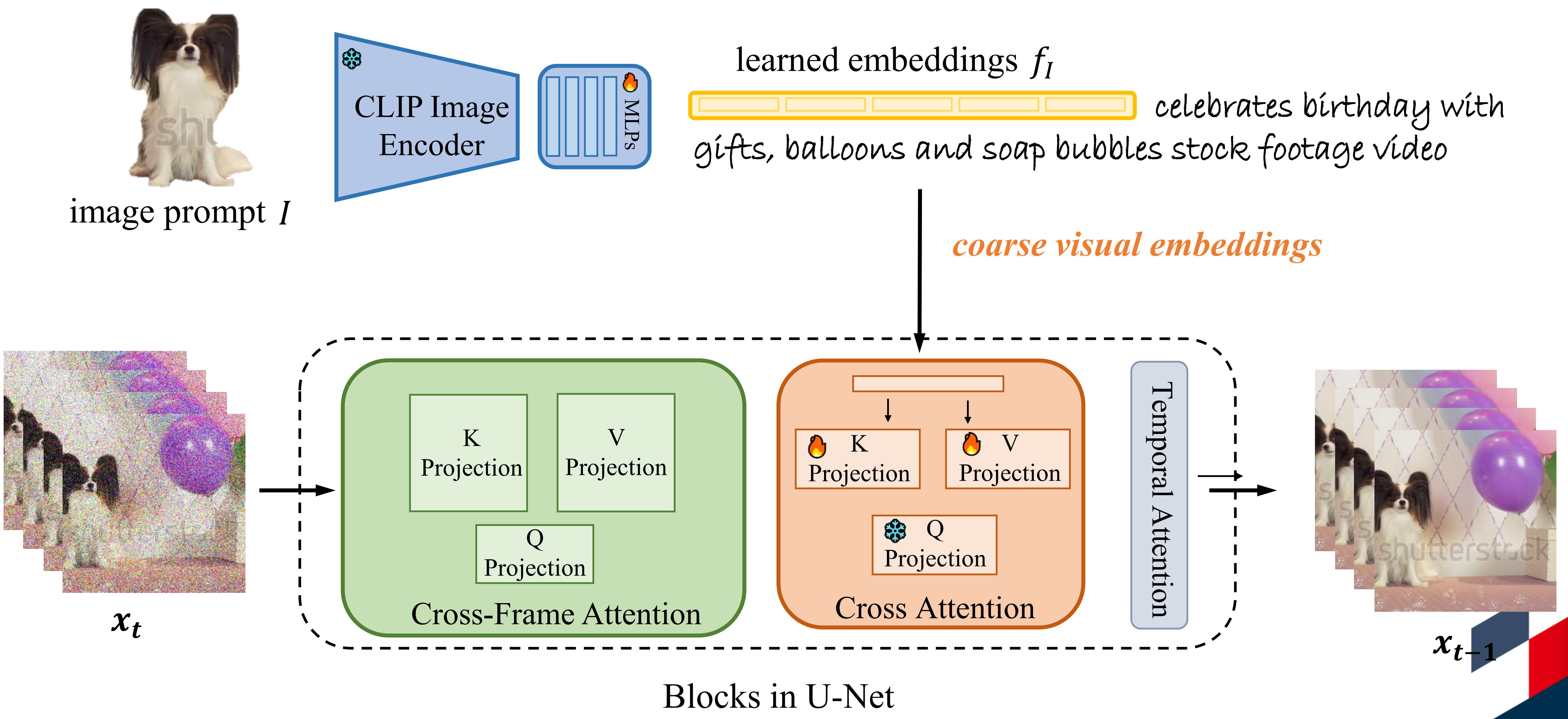
VideoBooth - Method



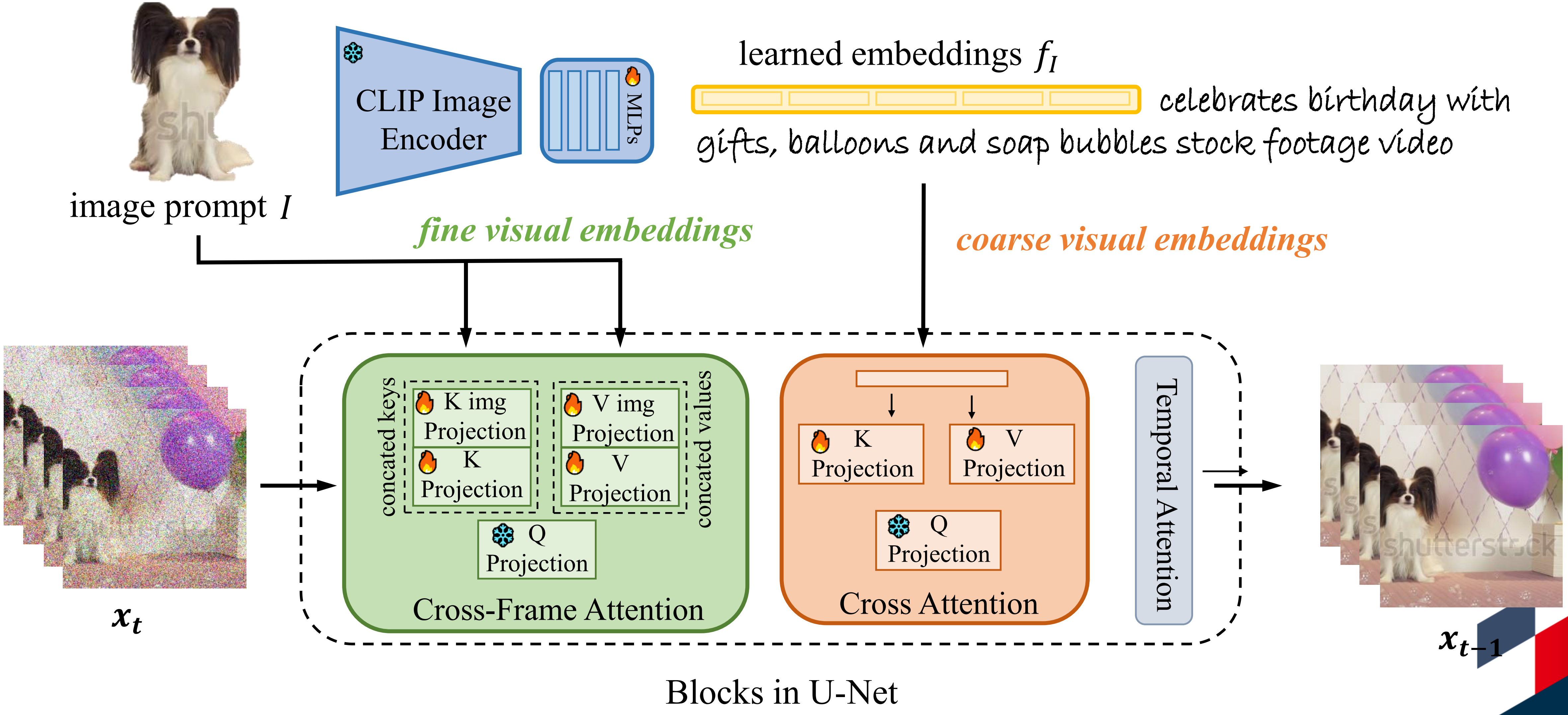
VideoBooth - Method



VideoBooth - Method



VideoBooth - Method



VideoBooth - Results

Image Prompt

Text Prompt

dog laying on ground

Textual Inversion

DreamBooth

ELITE

VideoBooth (Ours)

VideoBooth - Results

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Image Prompt

Text Prompt

close up of cat on top of a vintage chair

Textual Inversion

DreamBooth

ELITE

VideoBooth (Ours)

VideoBooth - Results

Image Prompt

Text Prompt

car in the bush

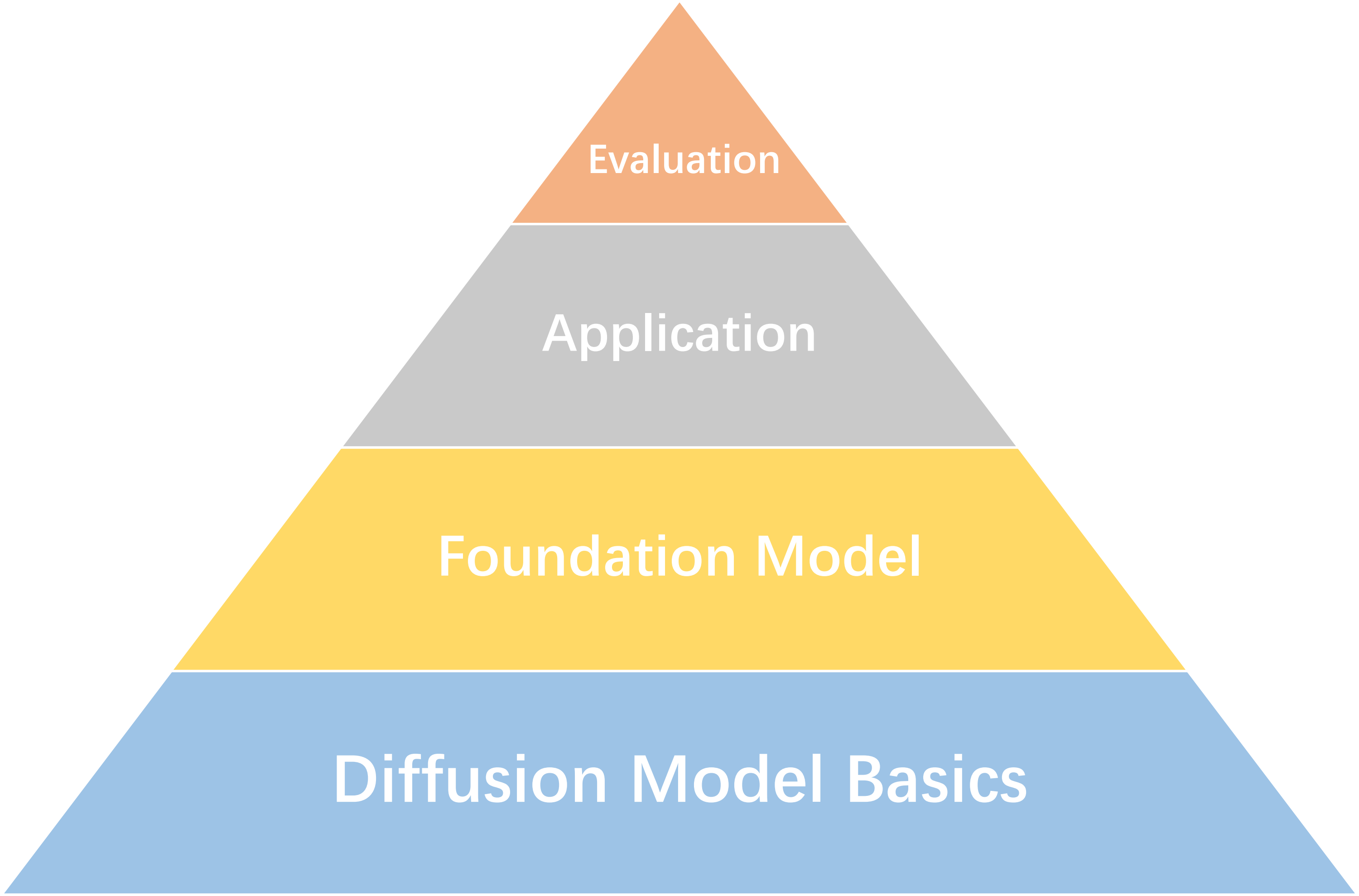
Textual Inversion

DreamBooth

ELITE

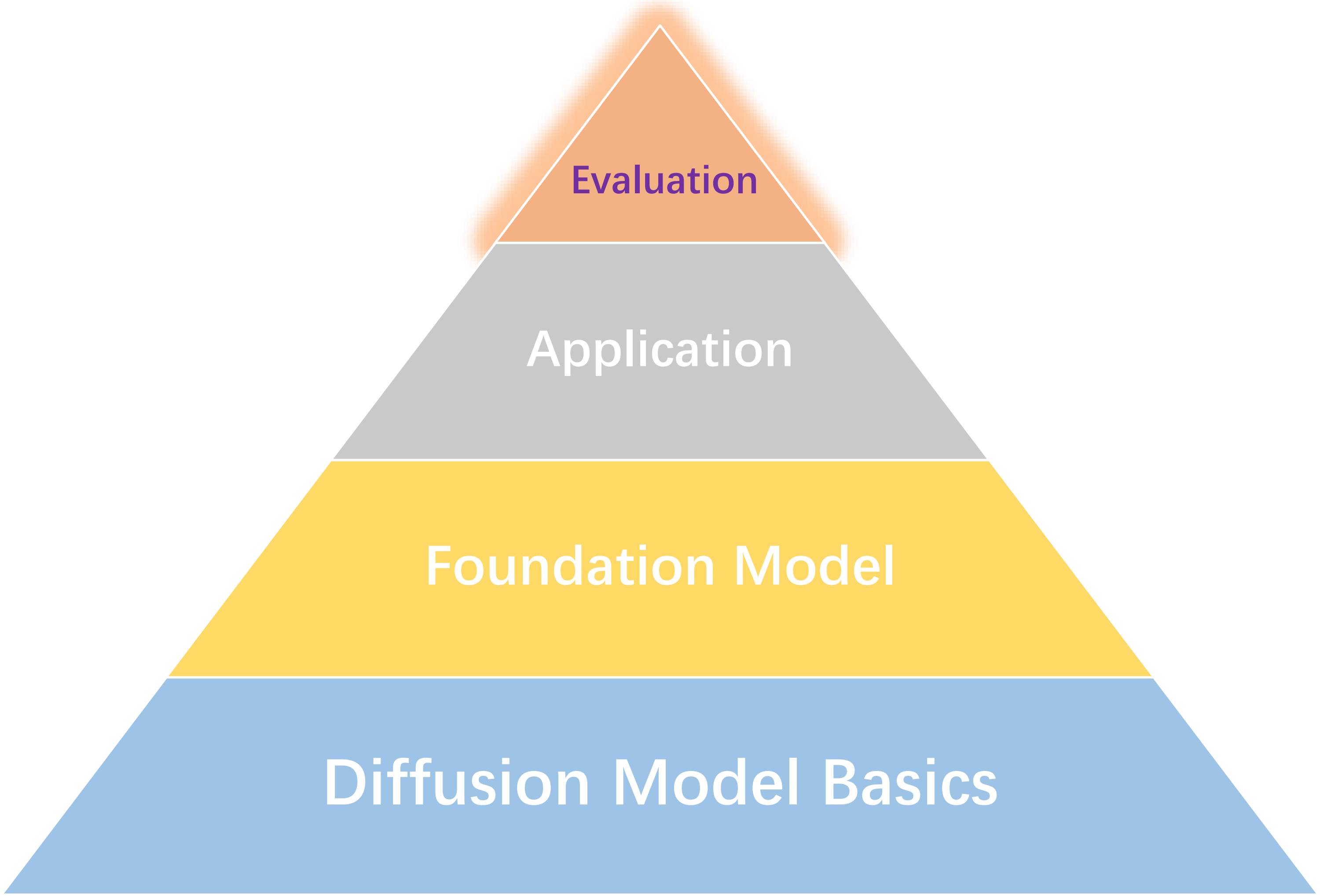
VideoBooth (Ours)

Video Generation



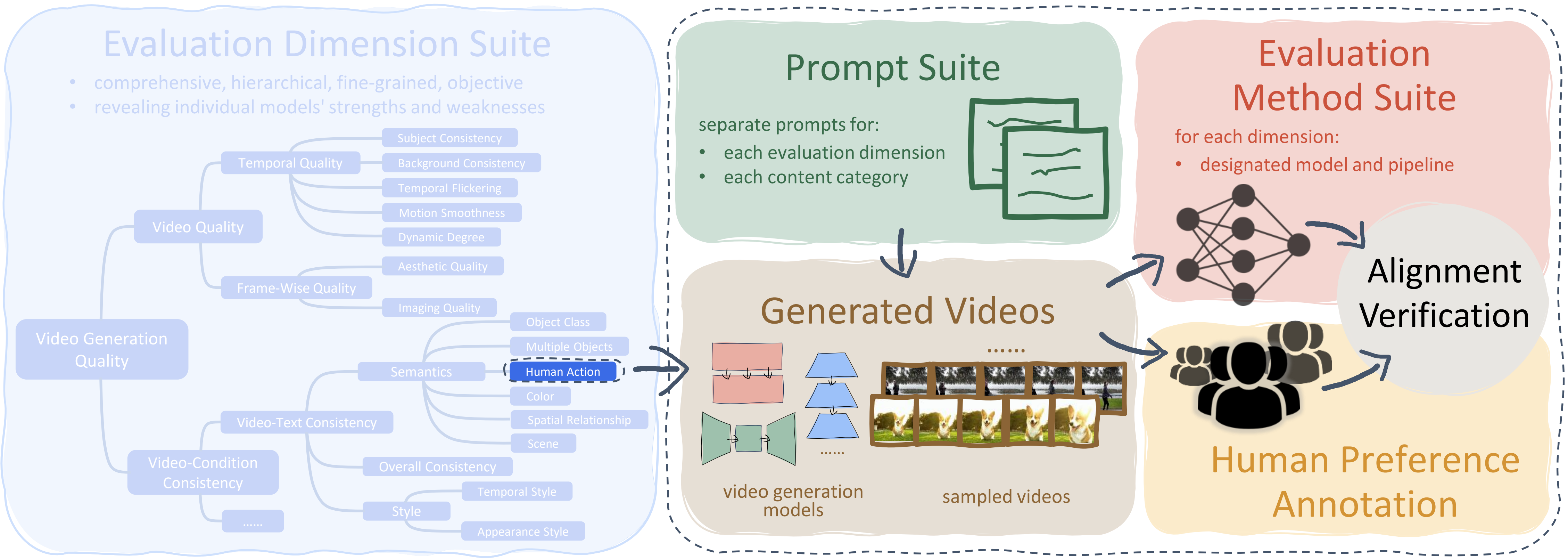
Video Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



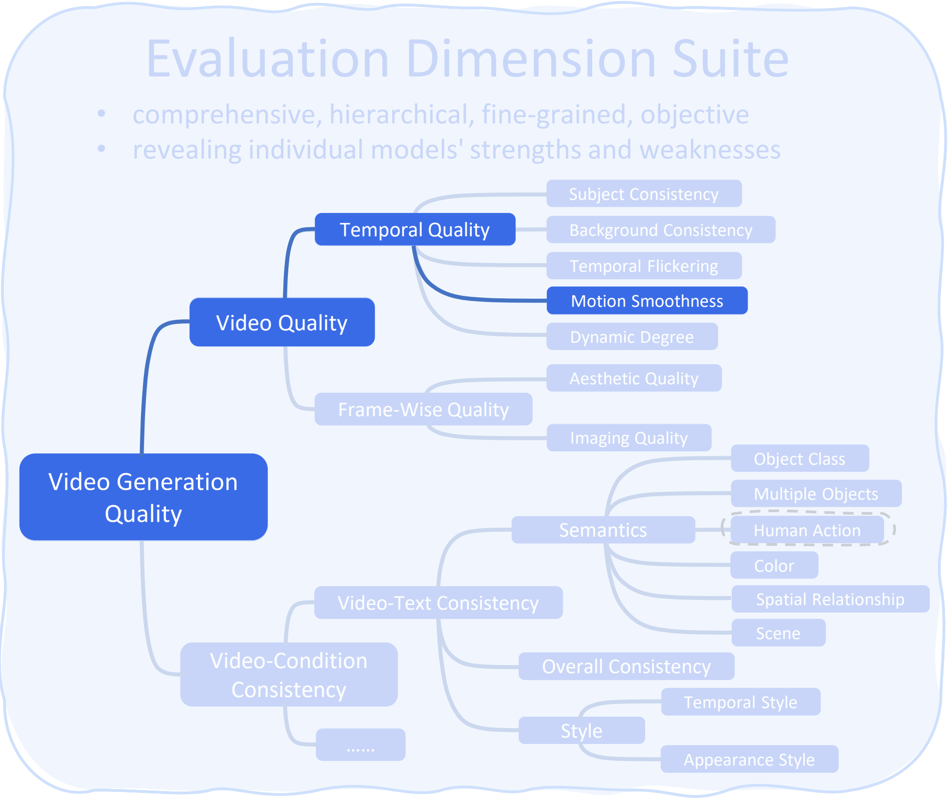
Overview of VBench

Comprehensive Benchmark Suite for Video Generative Models

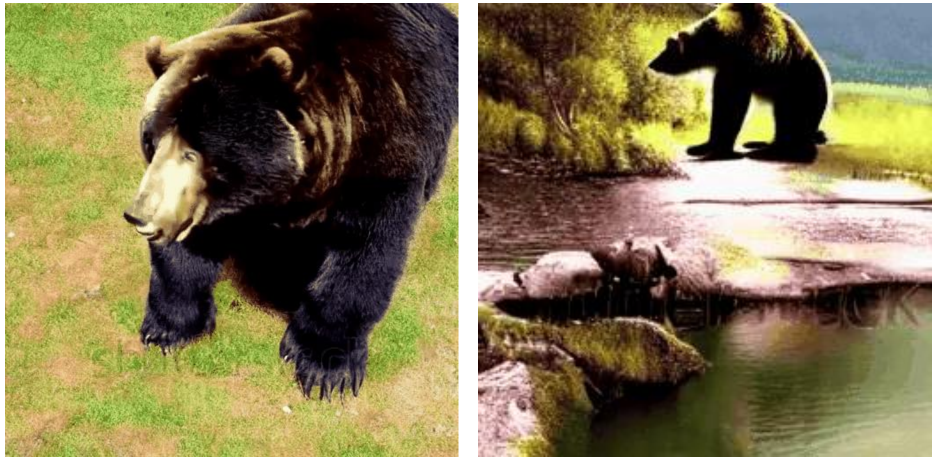


Evaluation Dimension: *Motion Smoothness*

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

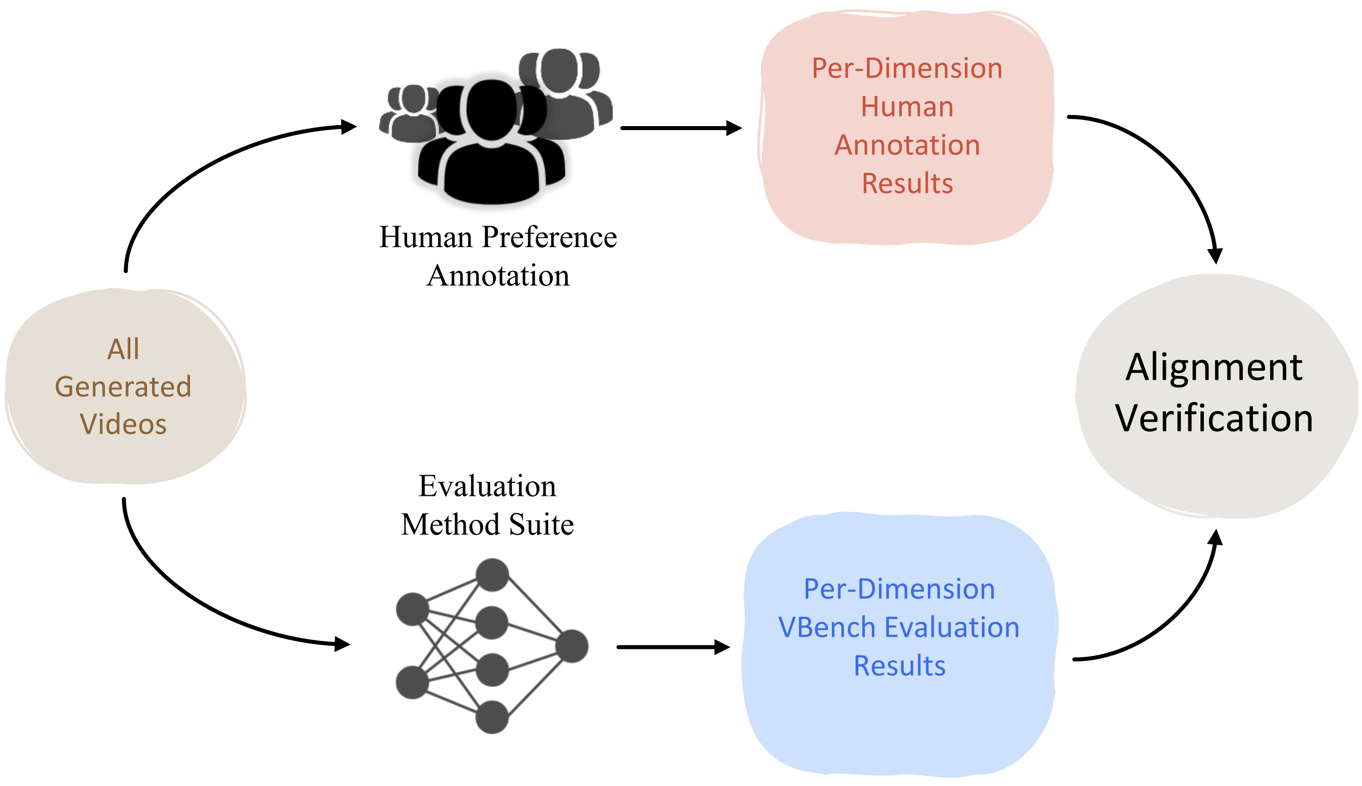


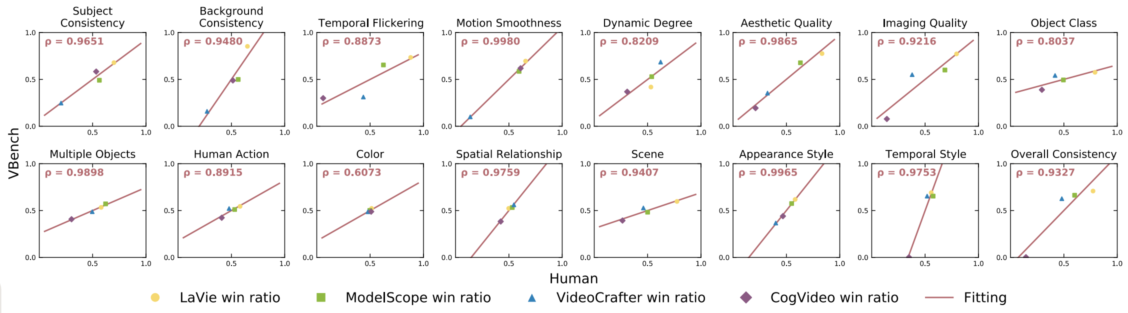
score 96.04% *(better)* **score 88.47%**



whether the motion in the generated video is smooth

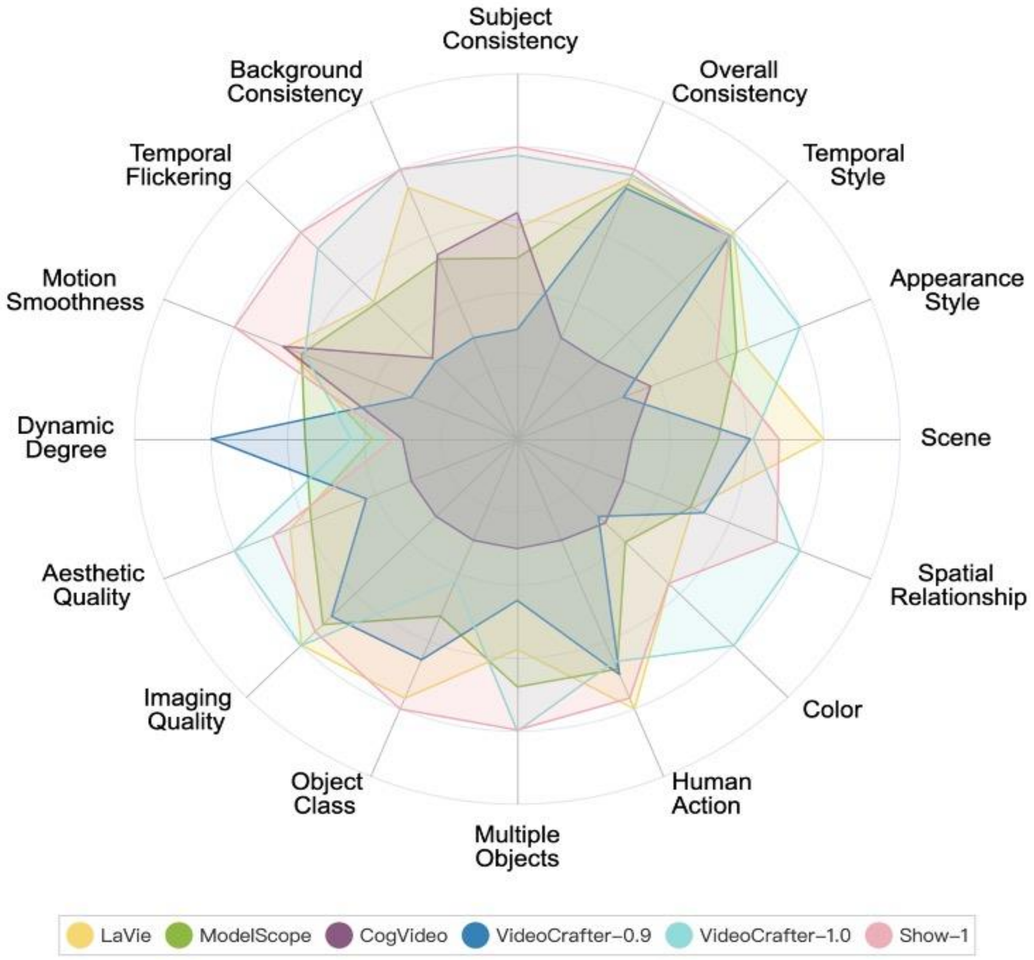
Human Alignment of VBench





VBench evaluations across all dimensions closely match human perceptions.

Insight 1. Trade-off between different dimensions



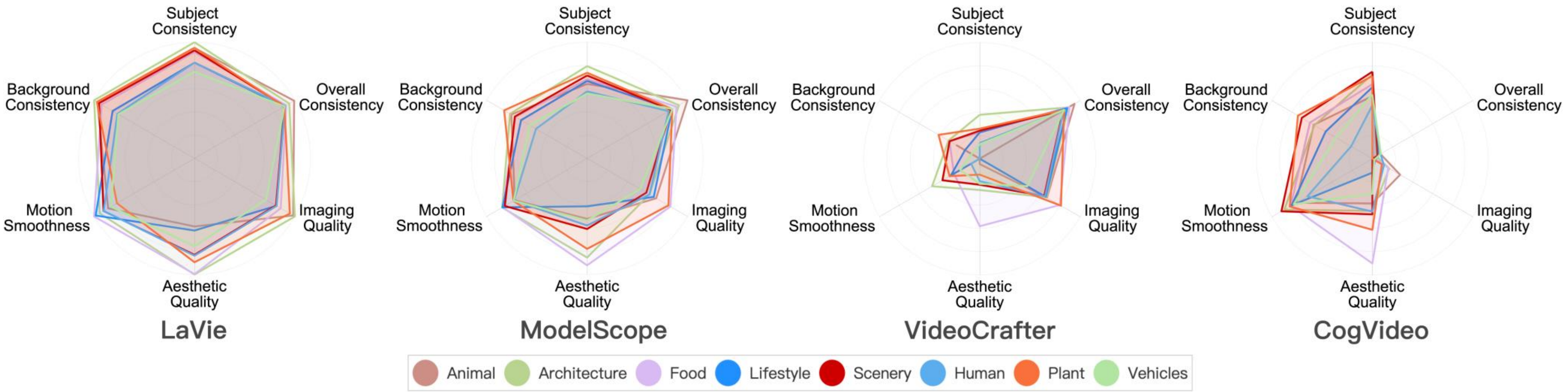
「时序连贯性」以及「视频的动态程度」：不要二选一，而应同时提升

我们发现时序连贯性（例如 Subject Consistency、Background Consistency、Motion Smoothness）与视频中运动的幅度（Dynamic Degree）之间有一定的权衡关系。比如说，Show-1 和 VideoCrafter-1.0 在背景一致性和动作流畅度方面表现很好，但在动态程度方面得分较低；这可能是因为生成「没有动起来」的画面更容易显得「在时序上很连贯」。另一方面，VideoCrafter-0.9 在与时序一致性的维度上弱一些，但在 Dynamic Degree 上得分很高。

这说明，同时做好「时序连贯性」和「较高的动态程度」确实挺难的；未来不应只关注其中一方面的提升，而应该同时提升「时序连贯性」以及「视频的动态程度」这两方面，这才是有意义的。

Insight 2.1. Do evaluate different content categories

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

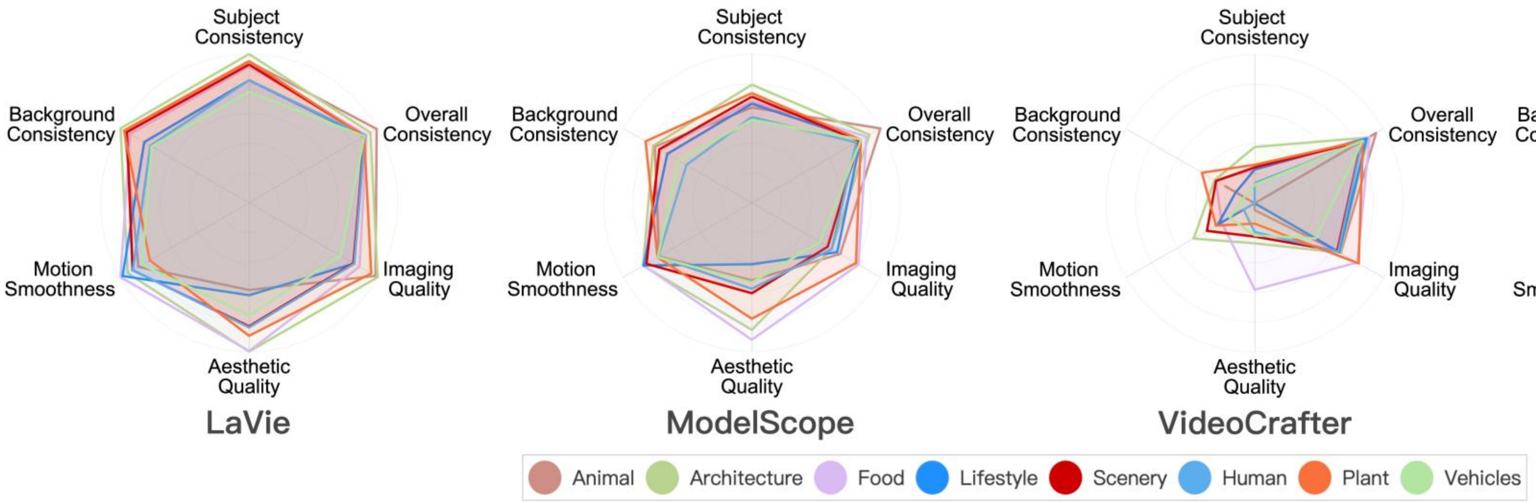


分场景内容进行评测，发掘各家模型潜力

有些模型在不同类别上表现出的性能存在较大差异，比如在美学质量（Aesthetic Quality）上，CogVideo 在 「Food」 类别上表现不错，而在 「LifeStyle」 类别得分较低。如果通过训练数据的调整，CogVideo 在 「LifeStyle」 这些类别上的美学质量是否可以提升上去，进而提升模型整体的视频美学质量?

这也告诉我们，在评估视频生成模型时，需要考虑模型在不同类别或主题下的表现，挖掘模型在某个能力维度的上限，进而针对性地提升 「拖后腿」 的场景类别。

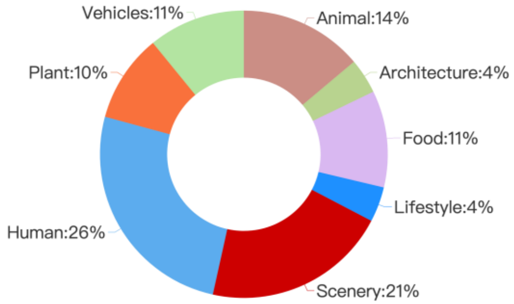
Insight 2.2. Do evaluate different content categories

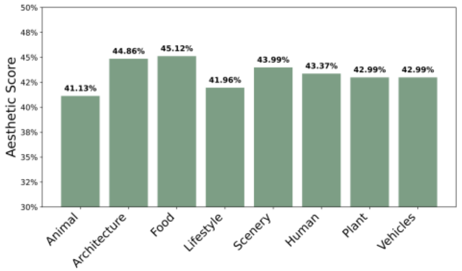


百万量级的数据集：提升数据质量优先于数据量

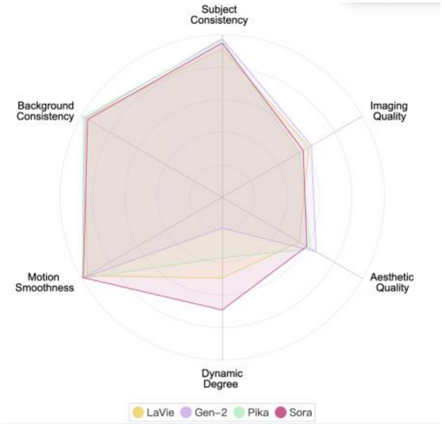
「Food」 类别虽然在 WebVid-10M 中仅占据 11%，但在评测中几乎总是拥有最高的美学质量分数。于是我们进一步分析了 WebVid-10M 数据集不同类别内容的美学质量表现，发现 「Food」 类别在 WebVid-10M 中也有最高的美学评分。

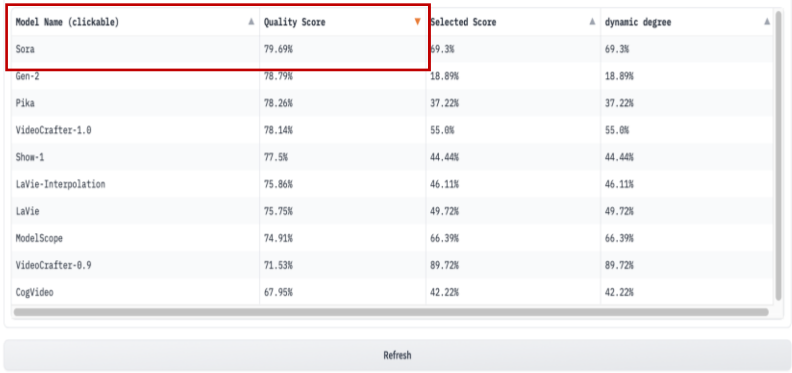
这意味着，在百万量级数据的基础上，筛选 / 提升数据质量比增加数据量更有帮助。





Sora





Model Name (clickable)	Quality Score	Selected Score	dynamic degree
Sora	79.69%	69.3%	69.3%
Gen-2	78.79%	18.89%	18.89%
Pika	78.26%	37.22%	37.22%
VideoCrafter-1.0	78.14%	55.0%	55.0%
Show-1	77.5%	44.44%	44.44%
LaVie-Interpolation	75.86%	46.11%	46.11%
LaVie	75.75%	49.72%	49.72%
ModelScope	74.91%	66.39%	66.39%
VideoCrafter-0.9	71.53%	89.72%	89.72%
CogVideo	67.95%	42.22%	42.22%

Refresh

- ****Preliminary Evaluation** on VBench**
  - Disclaimer: No Sora access, only evaluated dimensions not sensitive to text prompts, and it will only be completely fair after we sample using VBench text prompts
- **Rank 1**
  - especially dynamic degree

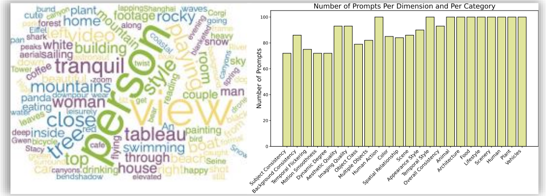
Sora

Gen-2

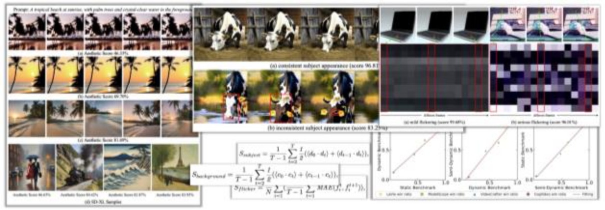
Pika

More details

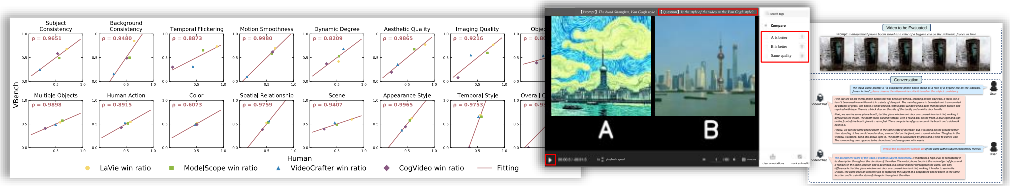
Prompt Suite



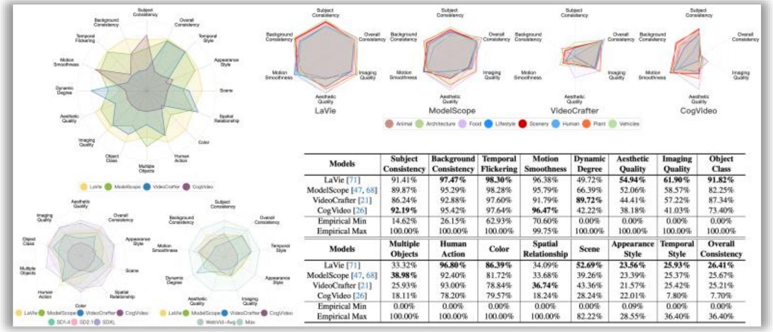
Evaluation Method Suite



Human Preference Annotation & Alignment



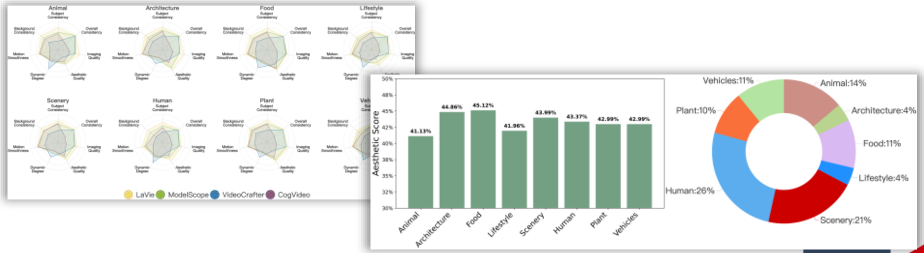
Comprehensive Experiments

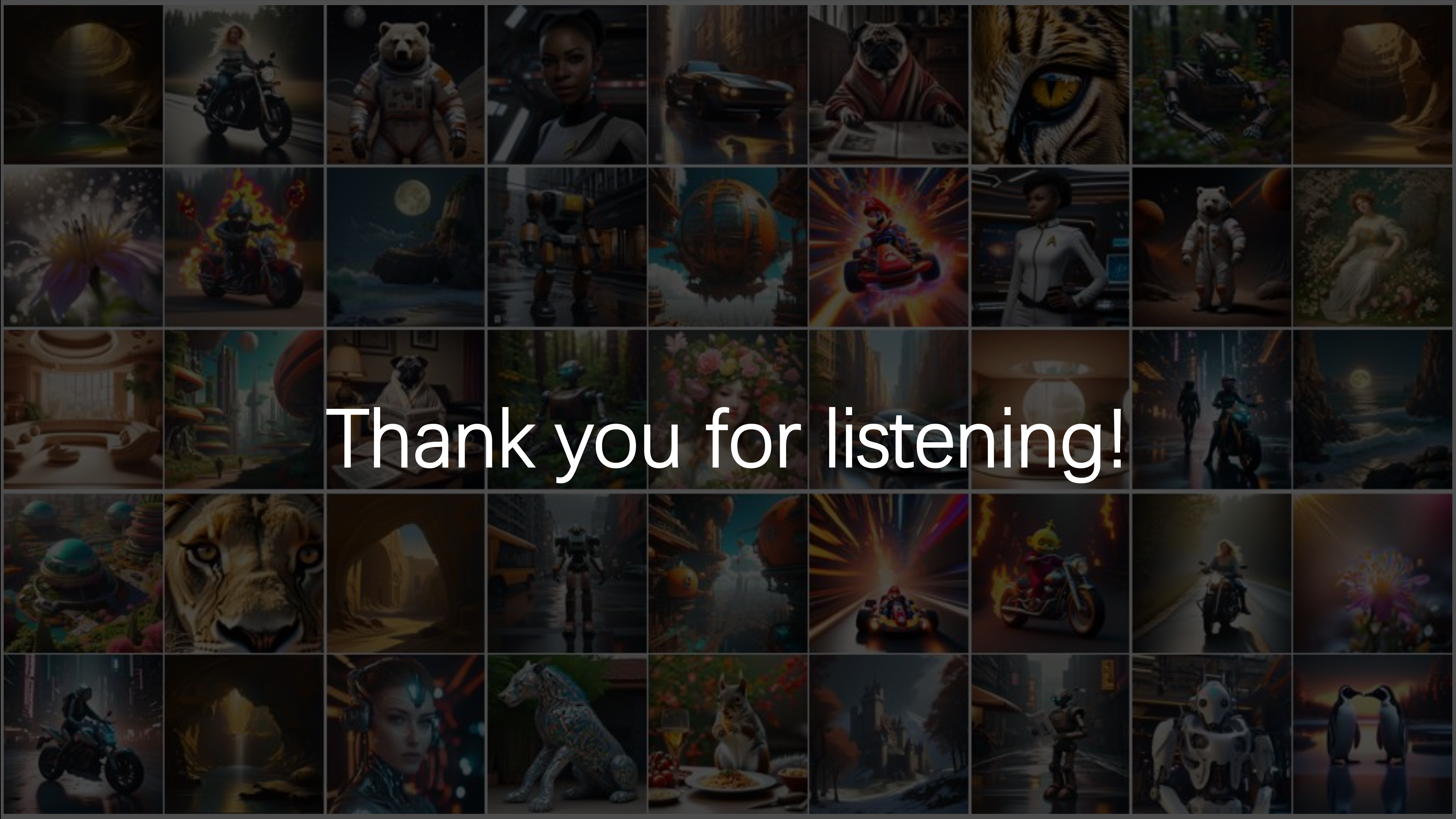


Models	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class
LaVie [71]	91.41%	**97.47%**	**98.30%**	96.38%	49.72%	**54.94%**	**61.90%**	**91.82%**
ModelScope [47, 68]	89.87%	95.29%	98.28%	95.79%	66.39%	52.06%	58.57%	82.25%
VideoCrafter [21]	86.24%	92.88%	97.60%	91.79%	**89.72%**	44.41%	57.22%	87.34%
CogVideo [26]	**92.19%**	95.42%	97.64%	**96.47%**	42.22%	38.18%	41.03%	73.40%
Empirical Min	14.62%	26.15%	62.93%	70.60%	0.00%	0.00%	0.00%	0.00%
Empirical Max	100.00%	100.00%	100.00%	99.75%	100.00%	100.00%	100.00%	100.00%

Models	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
LaVie [71]	33.32%	**96.80%**	**86.39%**	34.09%	**52.69%**	**23.56%**	**25.93%**	**26.41%**
ModelScope [47, 68]	**38.98%**	92.40%	81.72%	33.68%	39.26%	23.39%	25.37%	25.67%
VideoCrafter [21]	25.93%	93.00%	78.84%	**36.74%**	43.36%	21.57%	25.42%	25.21%
CogVideo [26]	18.11%	78.20%	79.57%	18.24%	28.24%	22.01%	7.80%	7.70%
Empirical Min	0.00%	0.00%	0.00%	0.00%	0.00%	0.09%	0.00%	0.00%
Empirical Max	100.00%	100.00%	100.00%	100.00%	82.22%	28.55%	36.40%	36.40%

In-Depth Insights




Thank you for listening!