NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Vchitect: Efficient and Scalable Video Generation

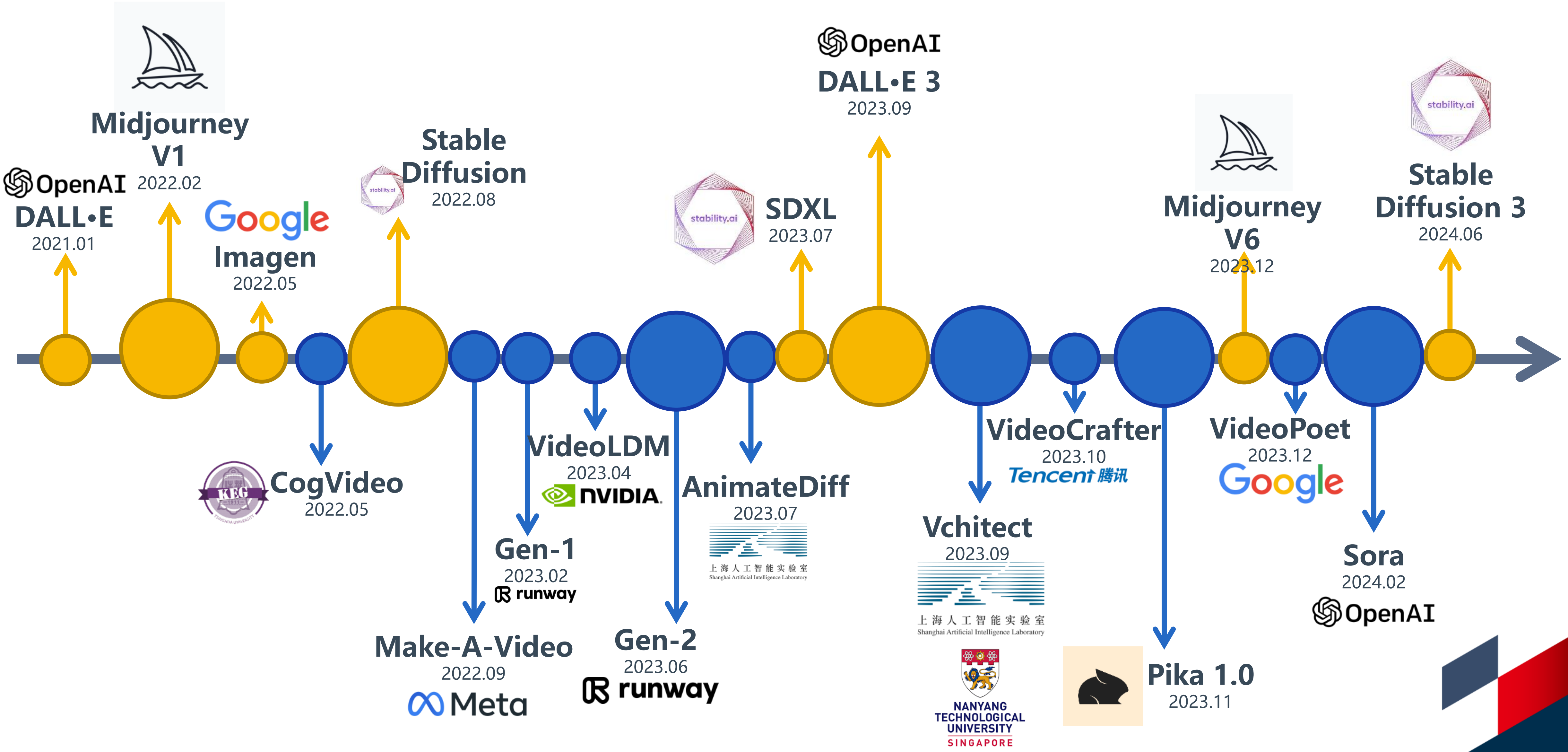
Ziwei Liu（刘子纬）

https://liuziwei7.github.io/

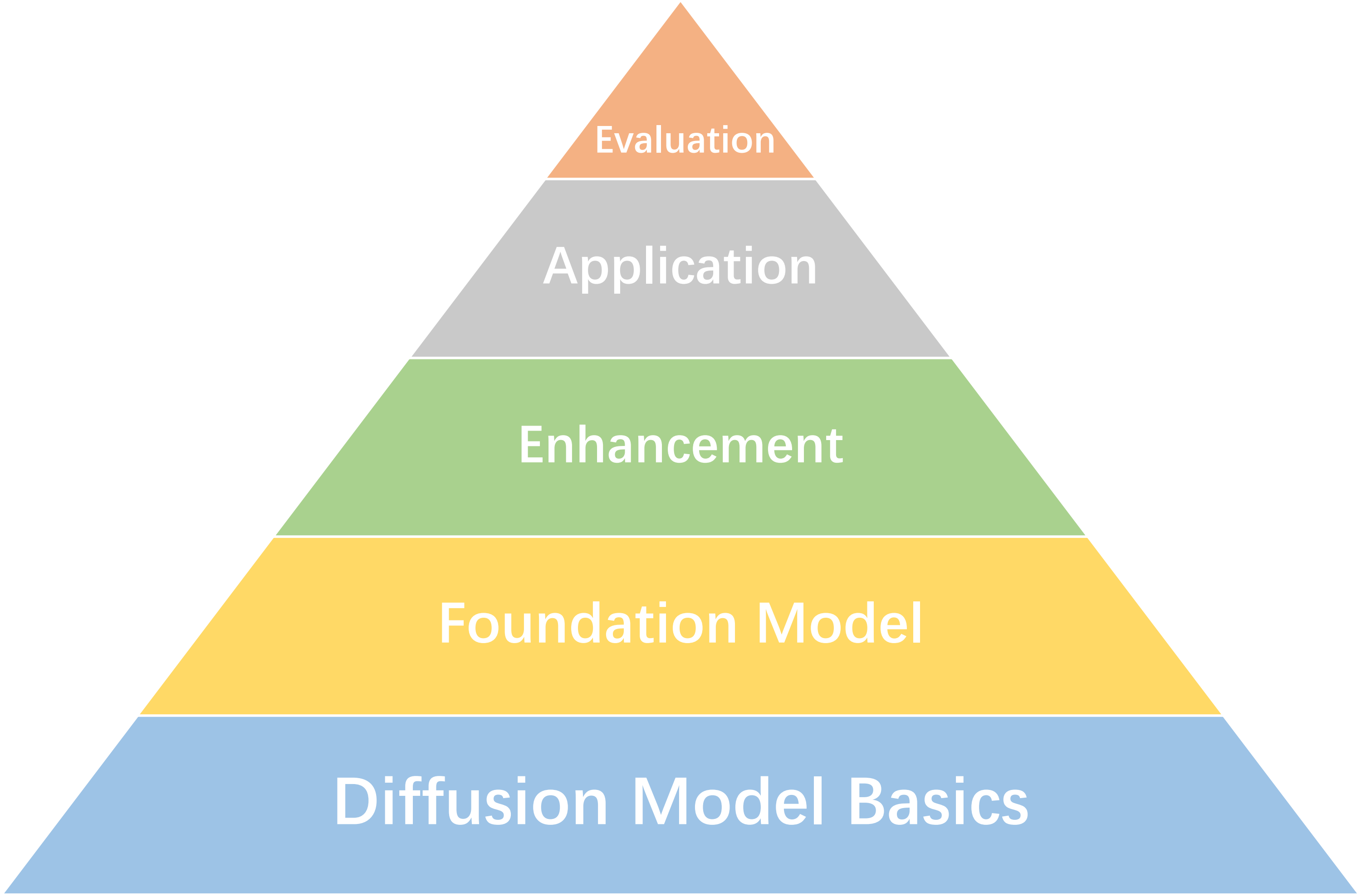
Nanyang Technological University

The Timeline from T2I to T2V

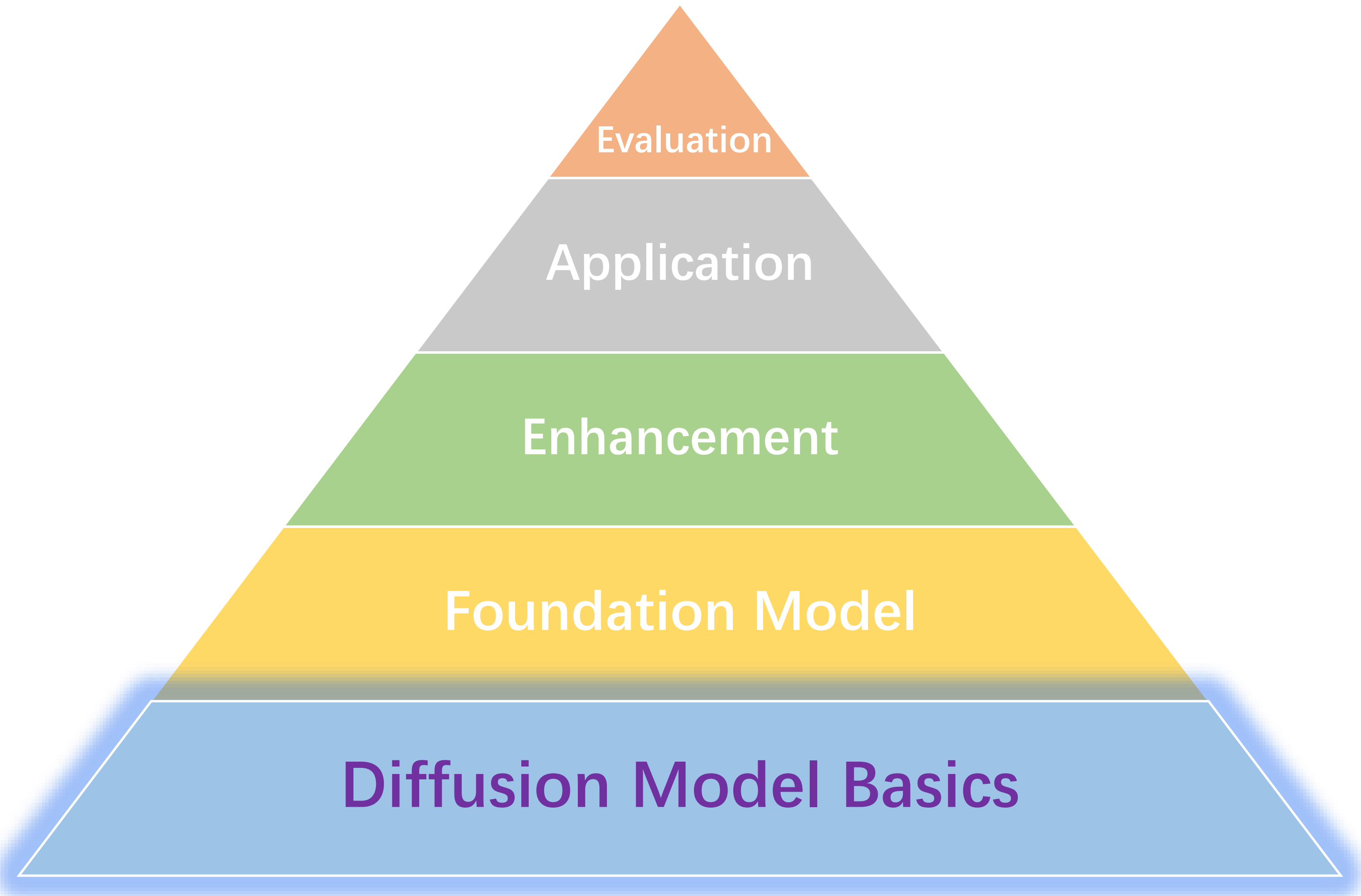
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



Video Generation



Video Generation

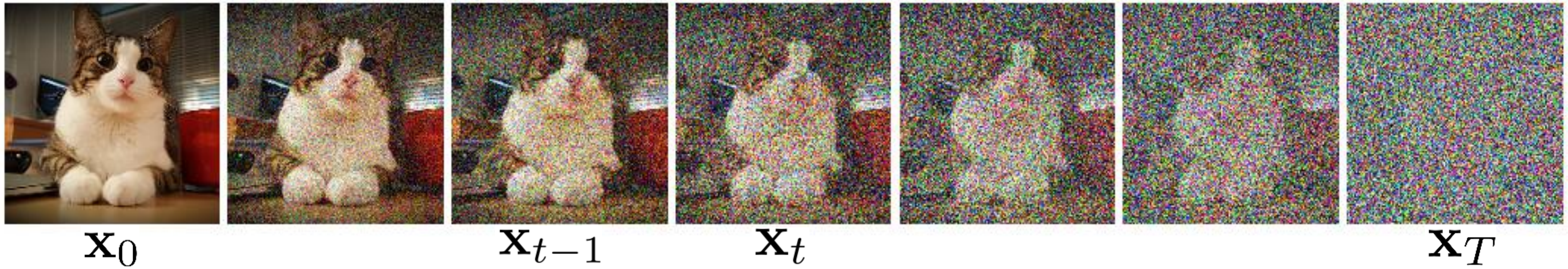


FreeU: Free Lunch in Diffusion U-Net

reverse process / denoising process

gradually denoise to image

Image



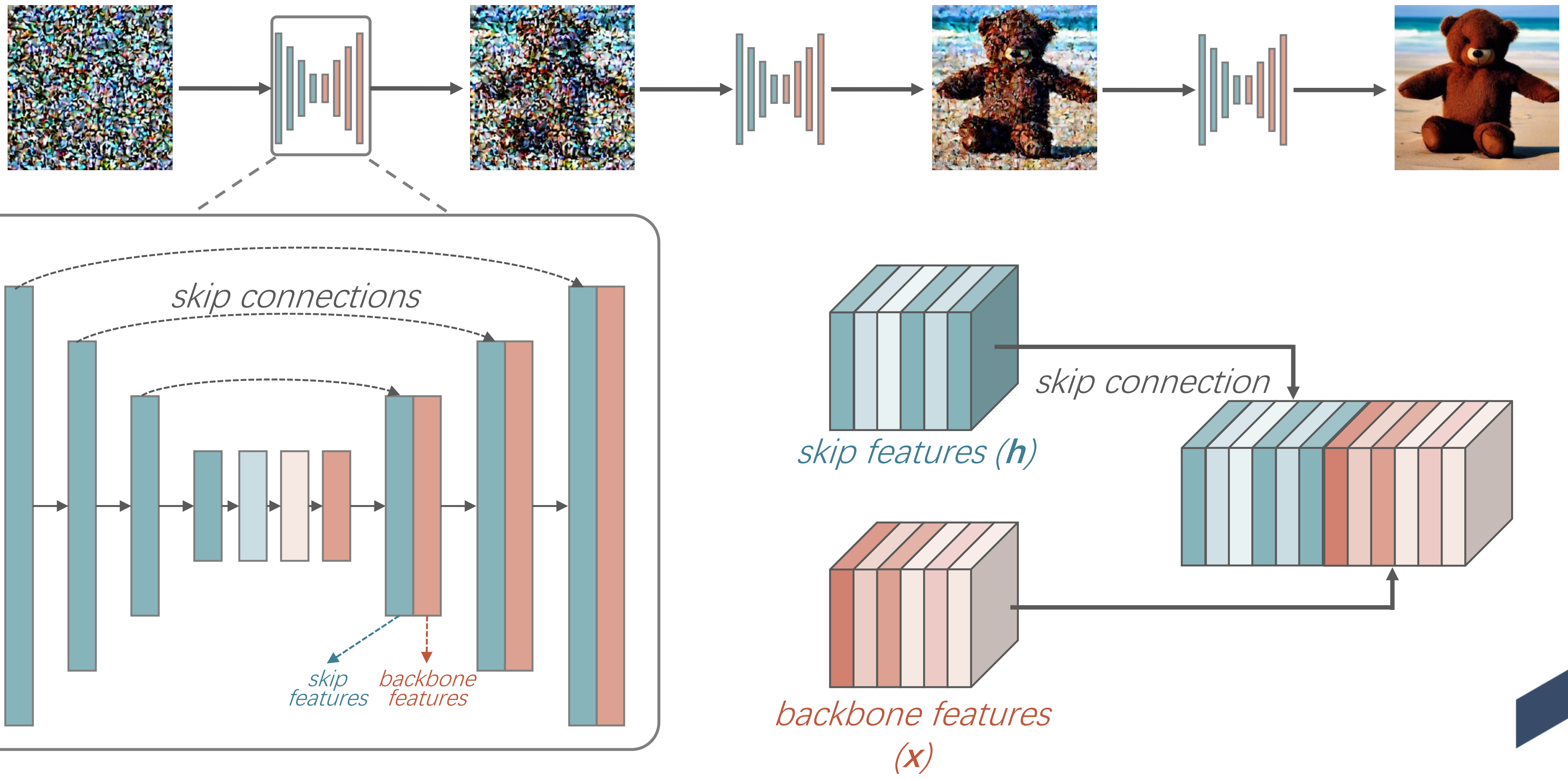
Noise

gradually adds Gaussian noise to the data

forward process / diffusion process

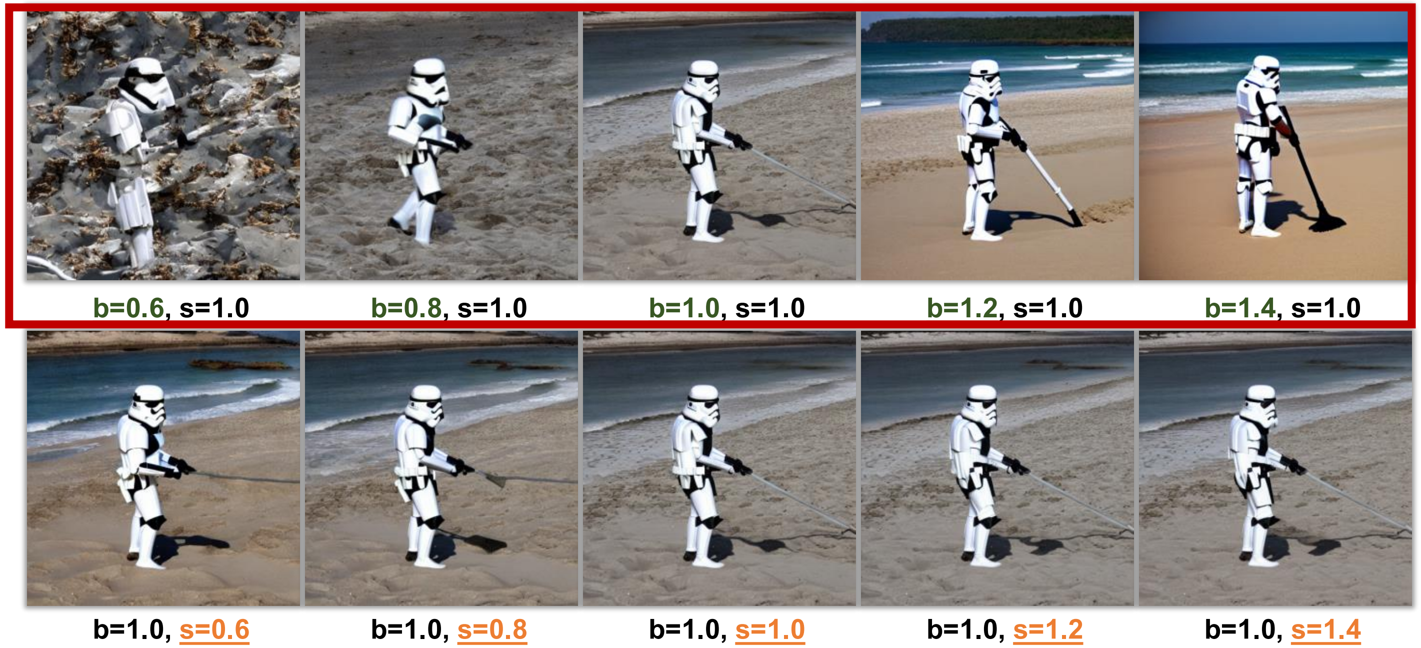
Image Credit: CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

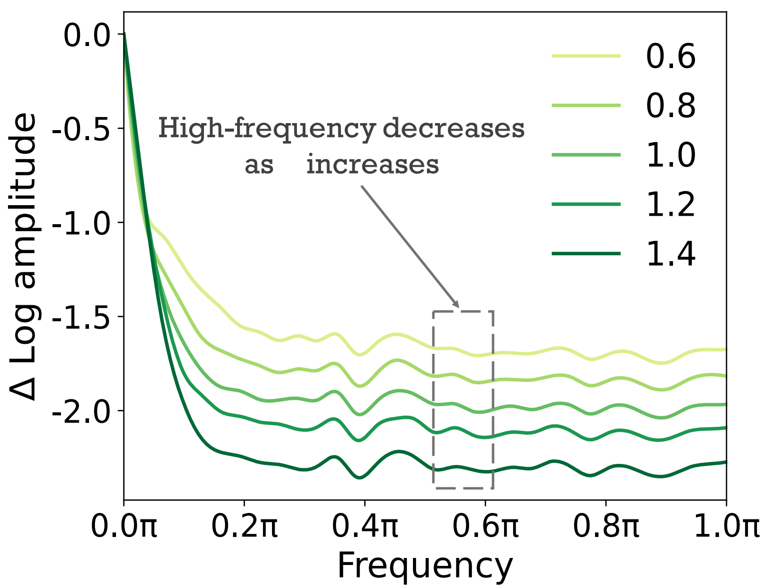
FreeU: Free Lunch in Diffusion U-Net



FreeU: Free Lunch in Diffusion U-Net

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

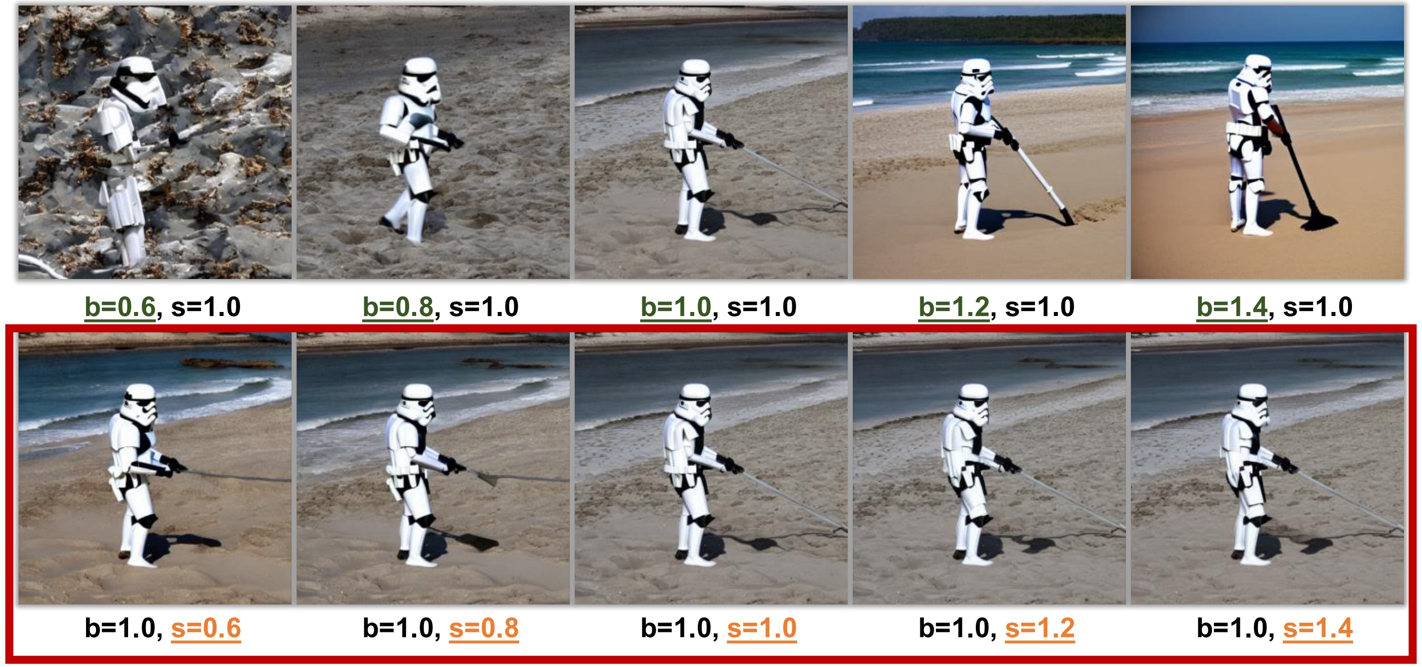


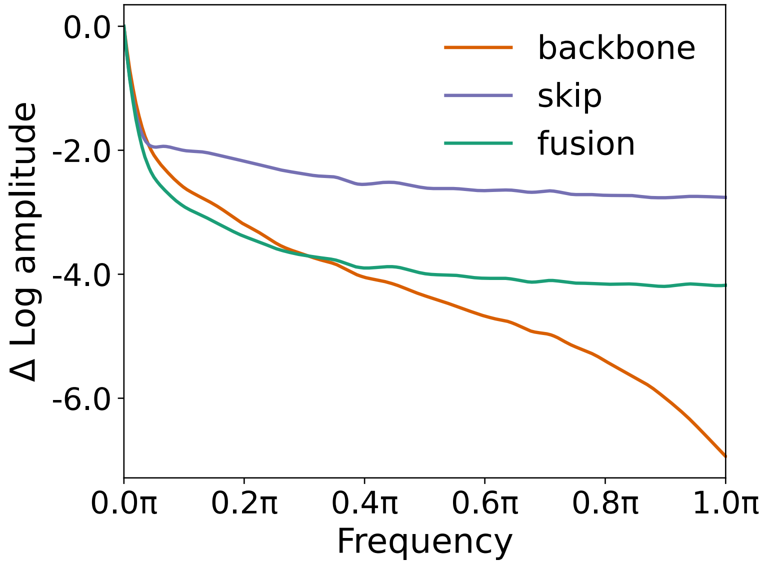


Fourier relative log amplitudes of variations of b

FreeU: Free Lunch in Diffusion U-Net

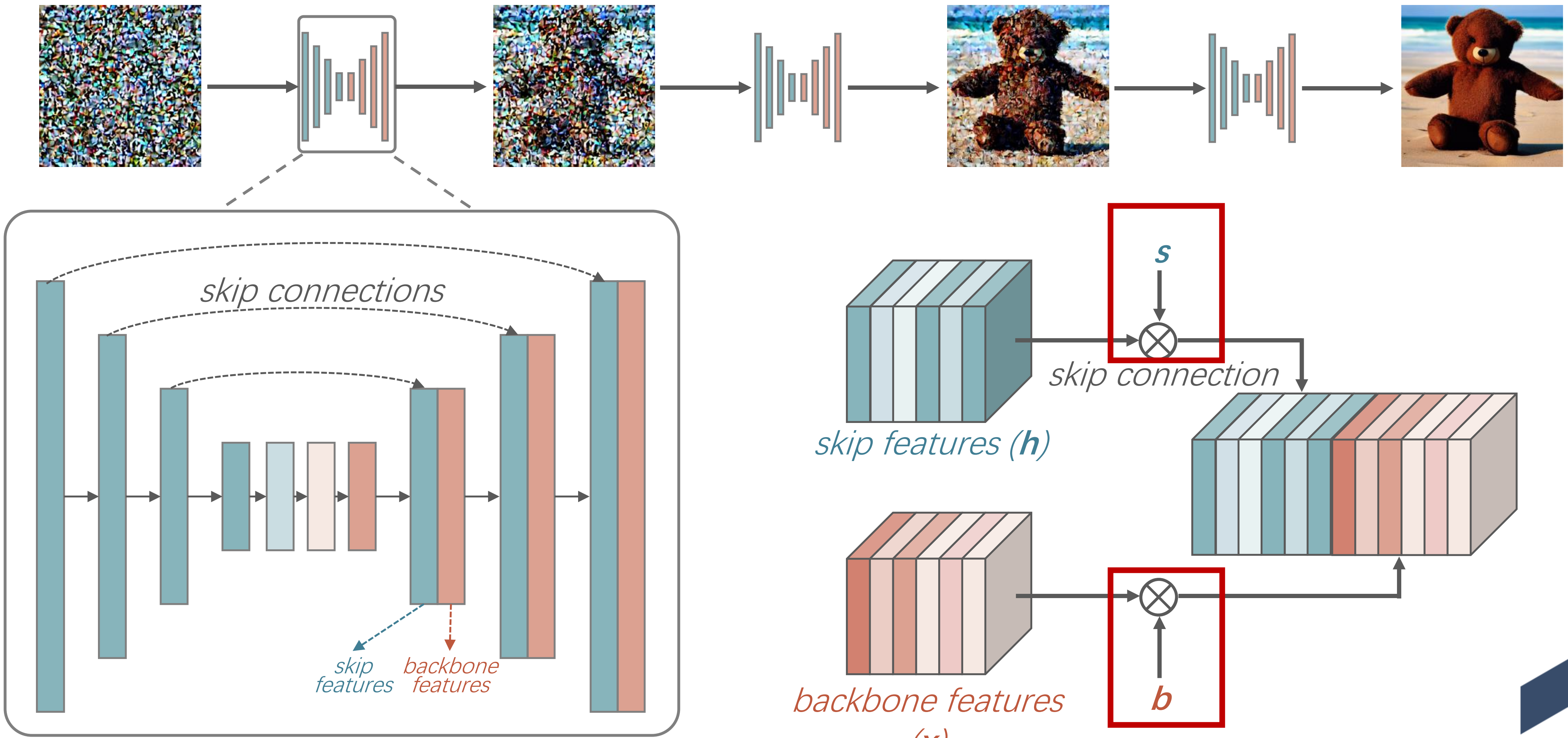
- **Backbone**: primarily contributes to denoising
- **Skip**: introduce high-frequency features into the decoder module





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

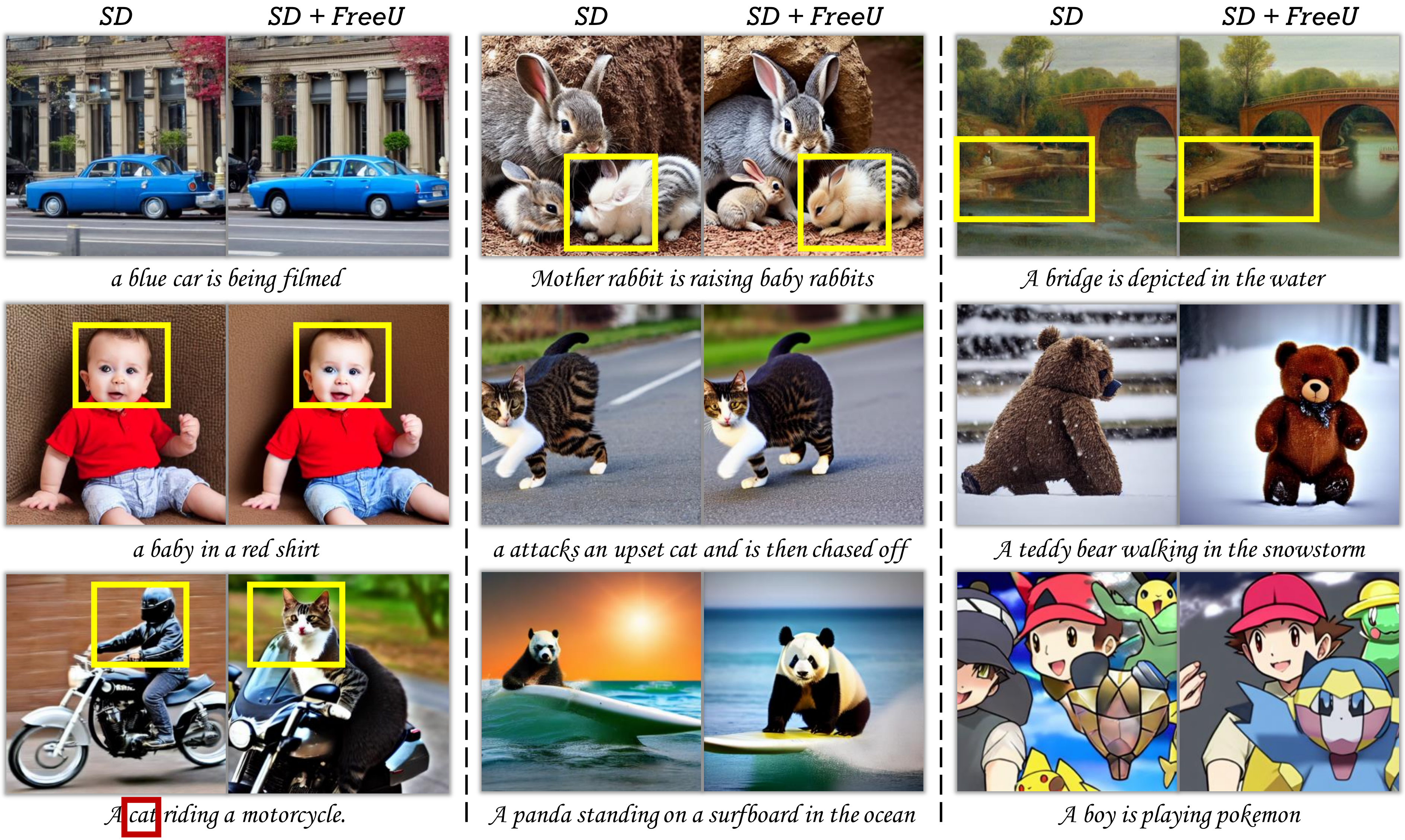
FreeU: Free Lunch in Diffusion U-Net



(a) UNet Architecture

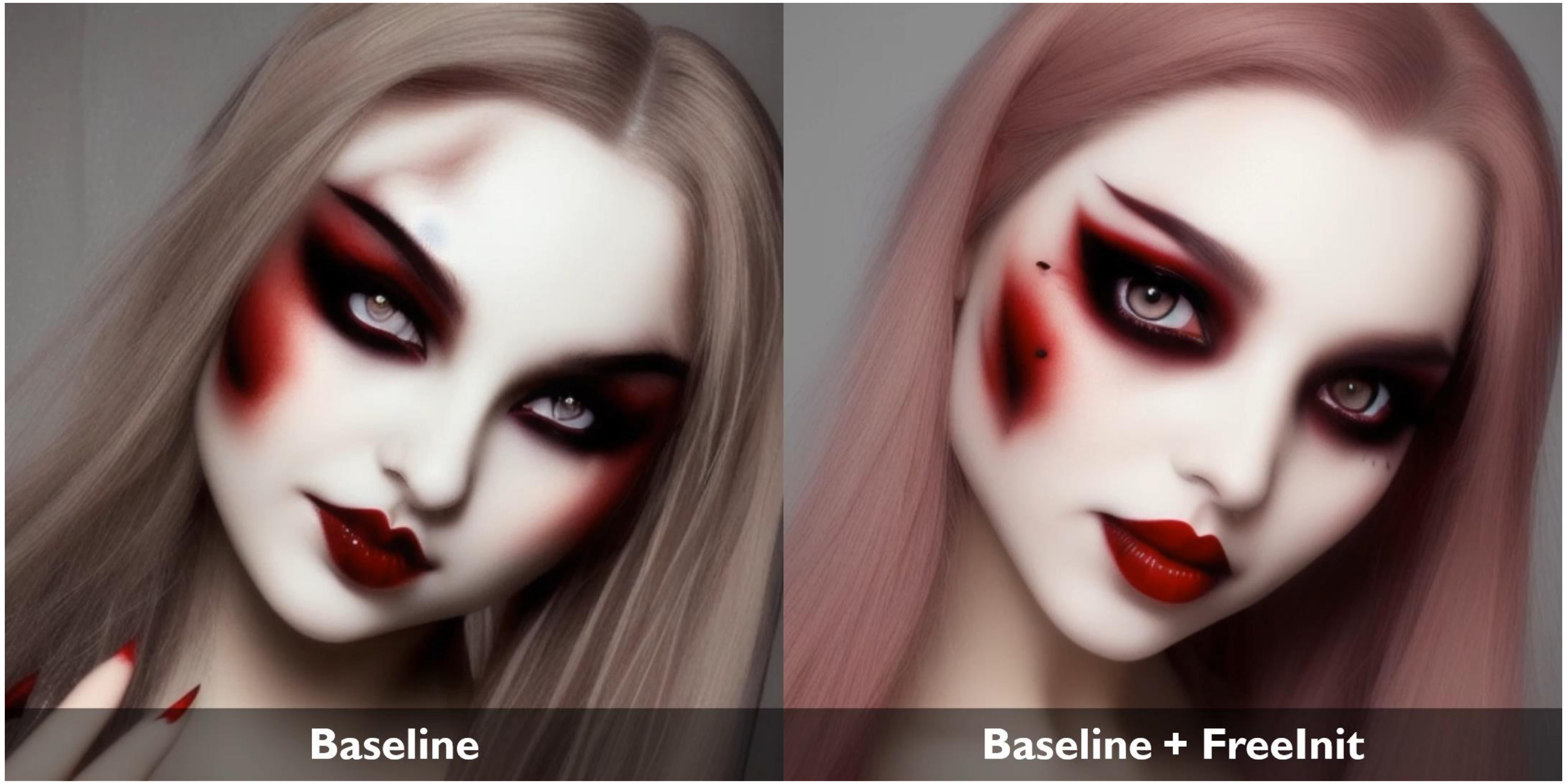
(b) FreeU Operations

Visual Results: Text-to-Image



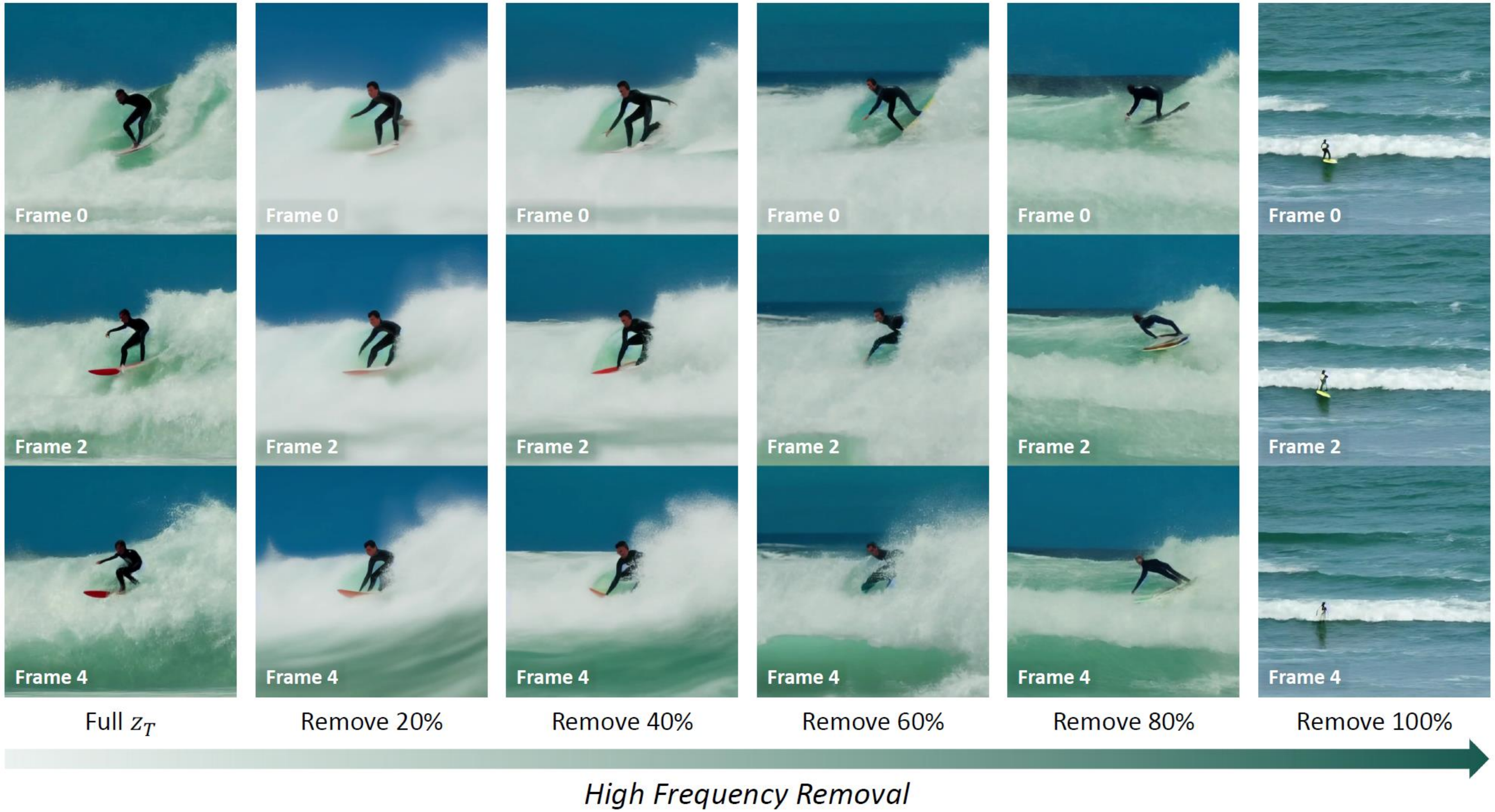
FreeInit

Bridging initialization gap in video diffusion models



- A training-free method for enhancing temporal consistency
- Support arbitrary video diffusion models

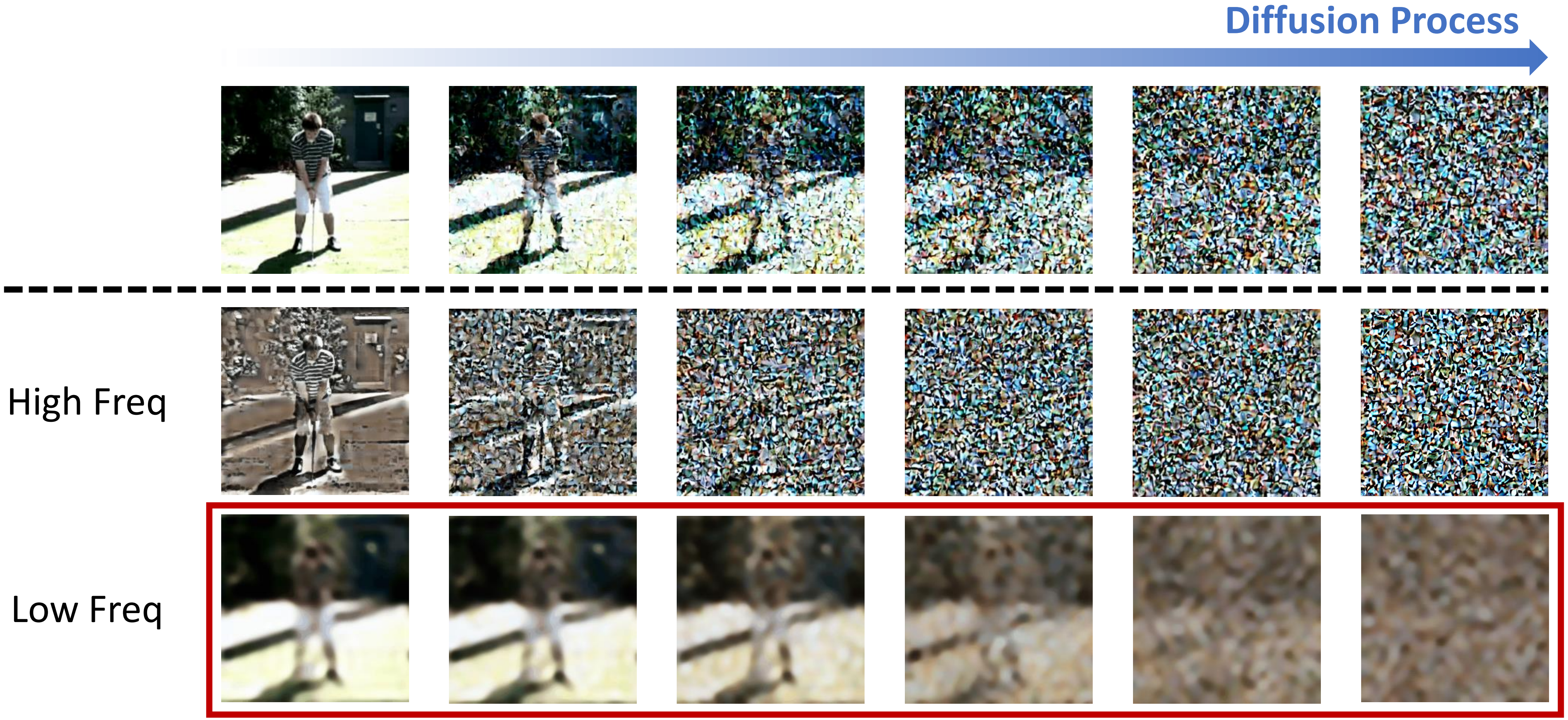
Observation: Initialization Gap



Observation 1: Low-frequency in Initial Noise *Matters*!

Spatio-temporal low-frequency components of the initial noise dominate the overall distribution.

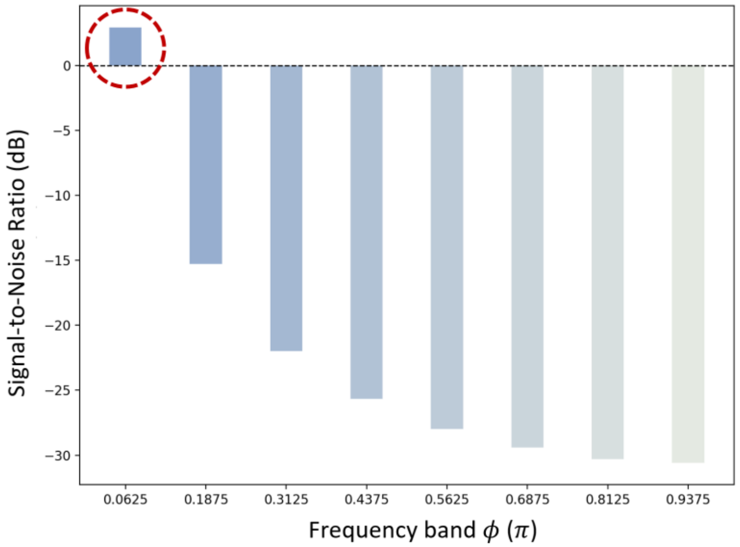
Observation: Initialization Gap



Observation 2: Information Leakage at Training:

The diffusion process cannot fully corrupt 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

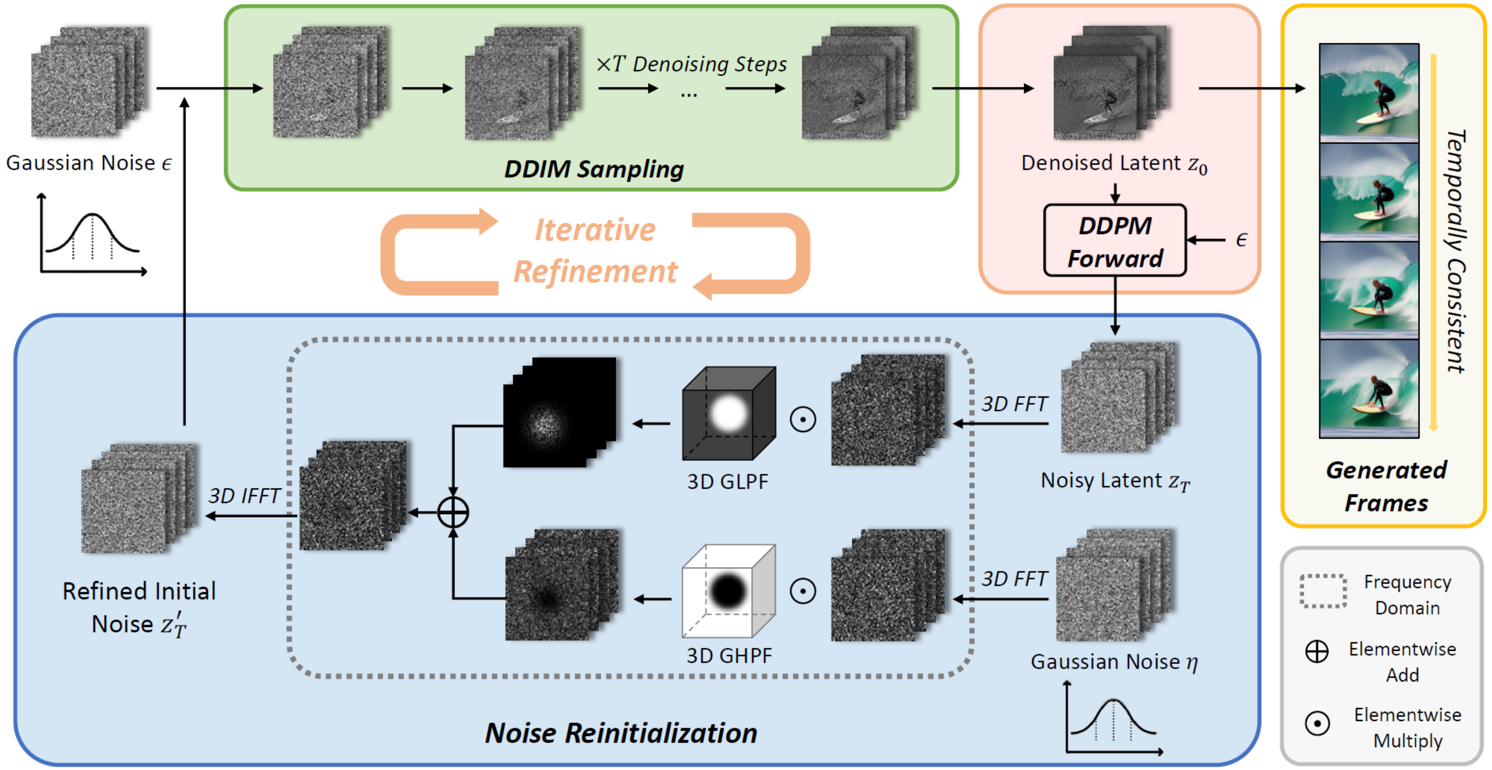
This causes an implicit *training-inference gap*:

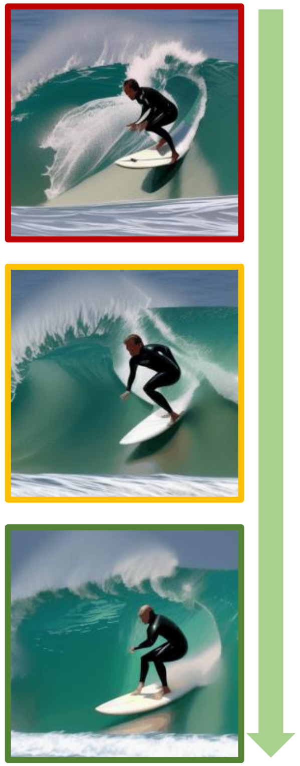
- At training, the initial noise contain temporal correlations at low-frequency band
- While at inference, the initial noise is pure Gaussian White Noise, lacking temporal correlations

Method

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

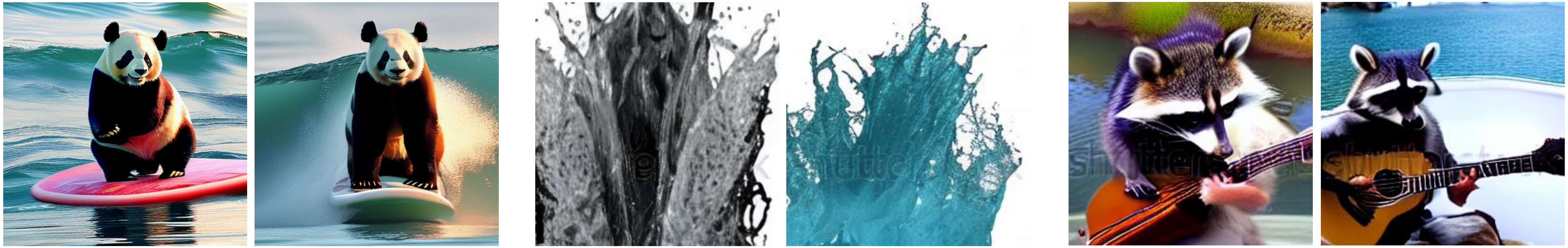




Gradually Enhanced Generation Quality

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 Gap: The model is trained on 16 frames, but is required to generate 64 frames.

Direct 16 Frames

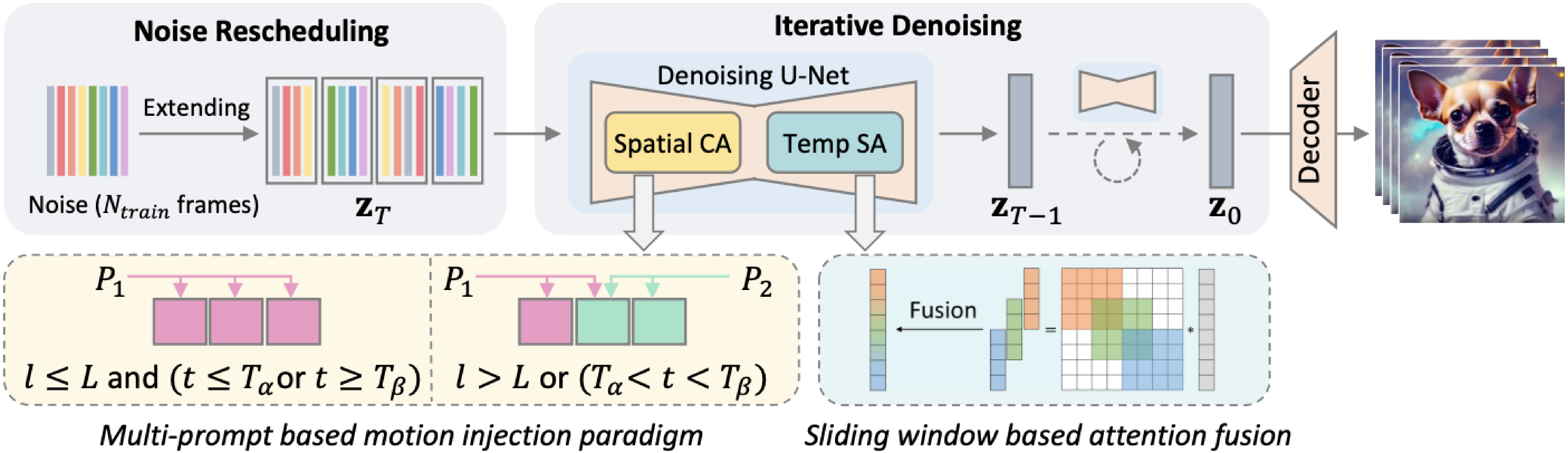
Direct 64 Frames

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

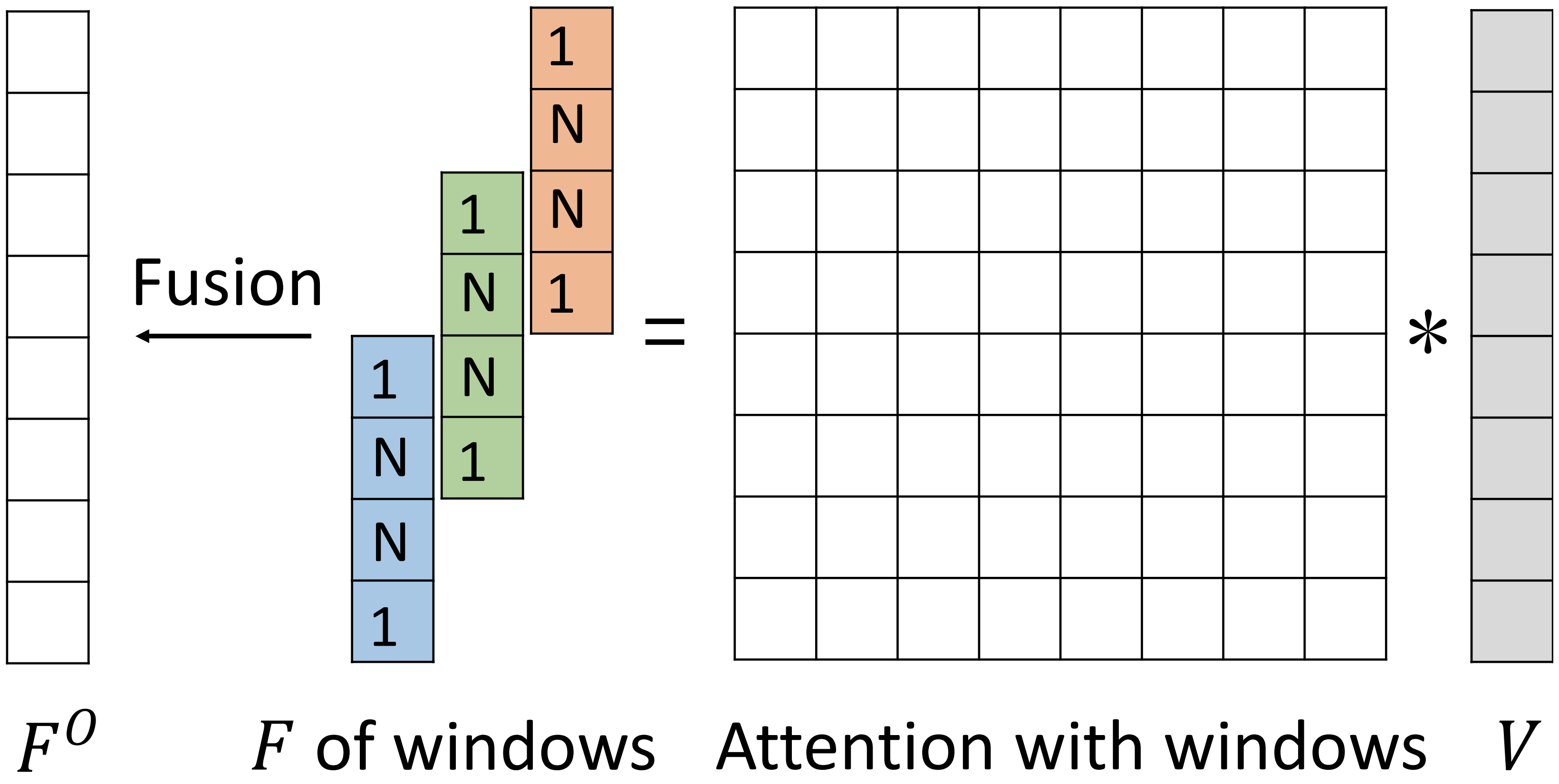
"A video of milk pouring over strawberries, blueberries, and blackberries. "

Method Overview

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

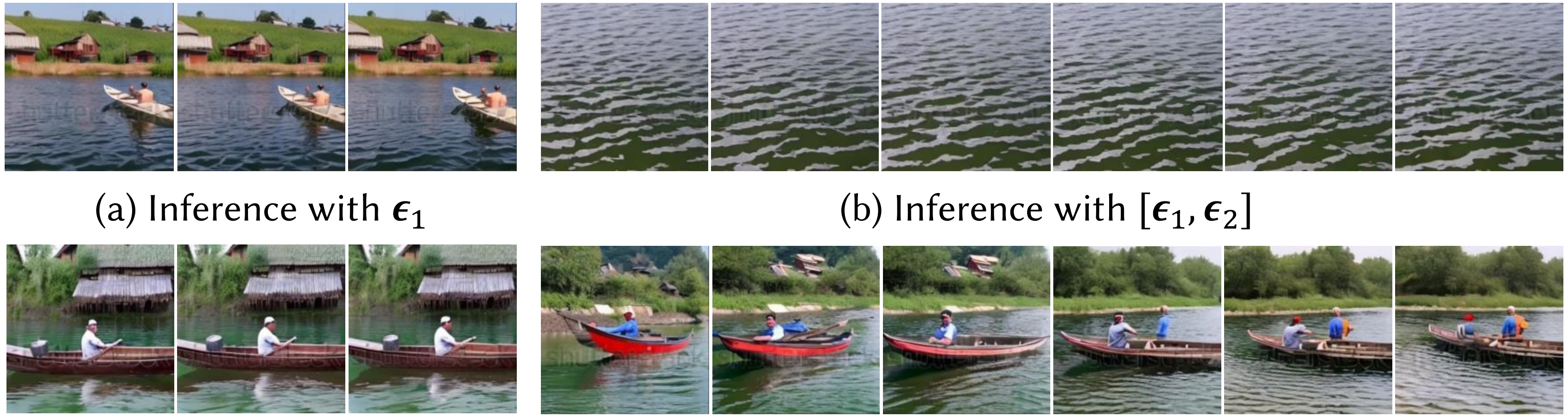


Local Window Fusion



Only apply to temporal attention, negligible additional costs

Noise Rescheduling

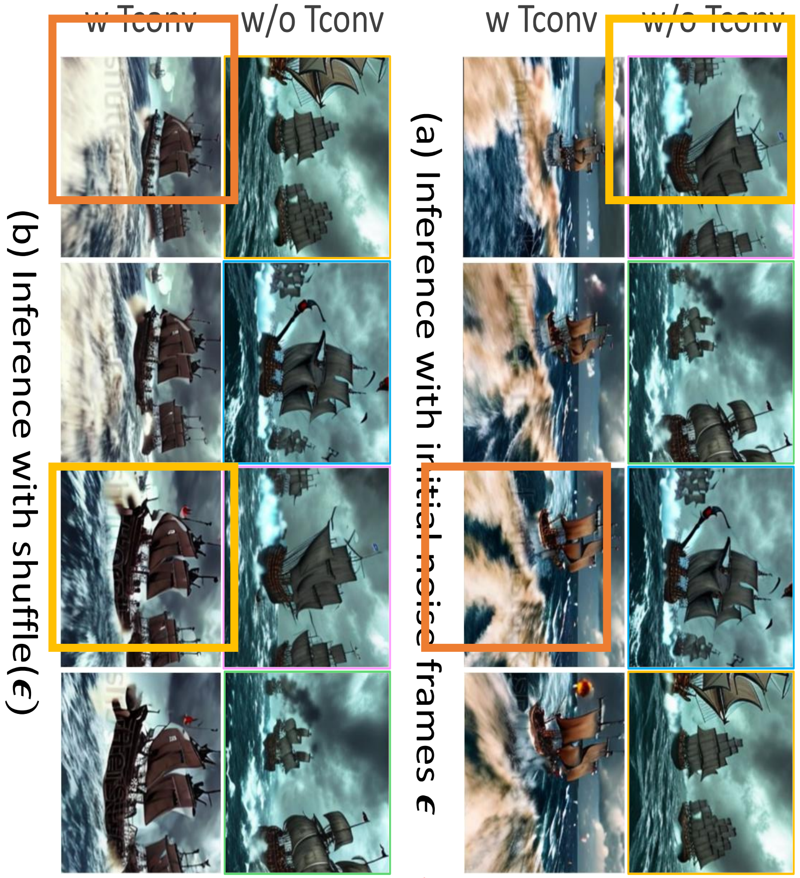


(a) Inference with $\boldsymbol{\epsilon}_1$

(b) Inference with $[\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2]$

(c) Inference with $\boldsymbol{\epsilon}_2$

(d) Sliding window inference with $[\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2]$



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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

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

GenL

Ours w/o Motion Injection

Ours

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

Results

Direct Inference	Sliding	GenL	Ours

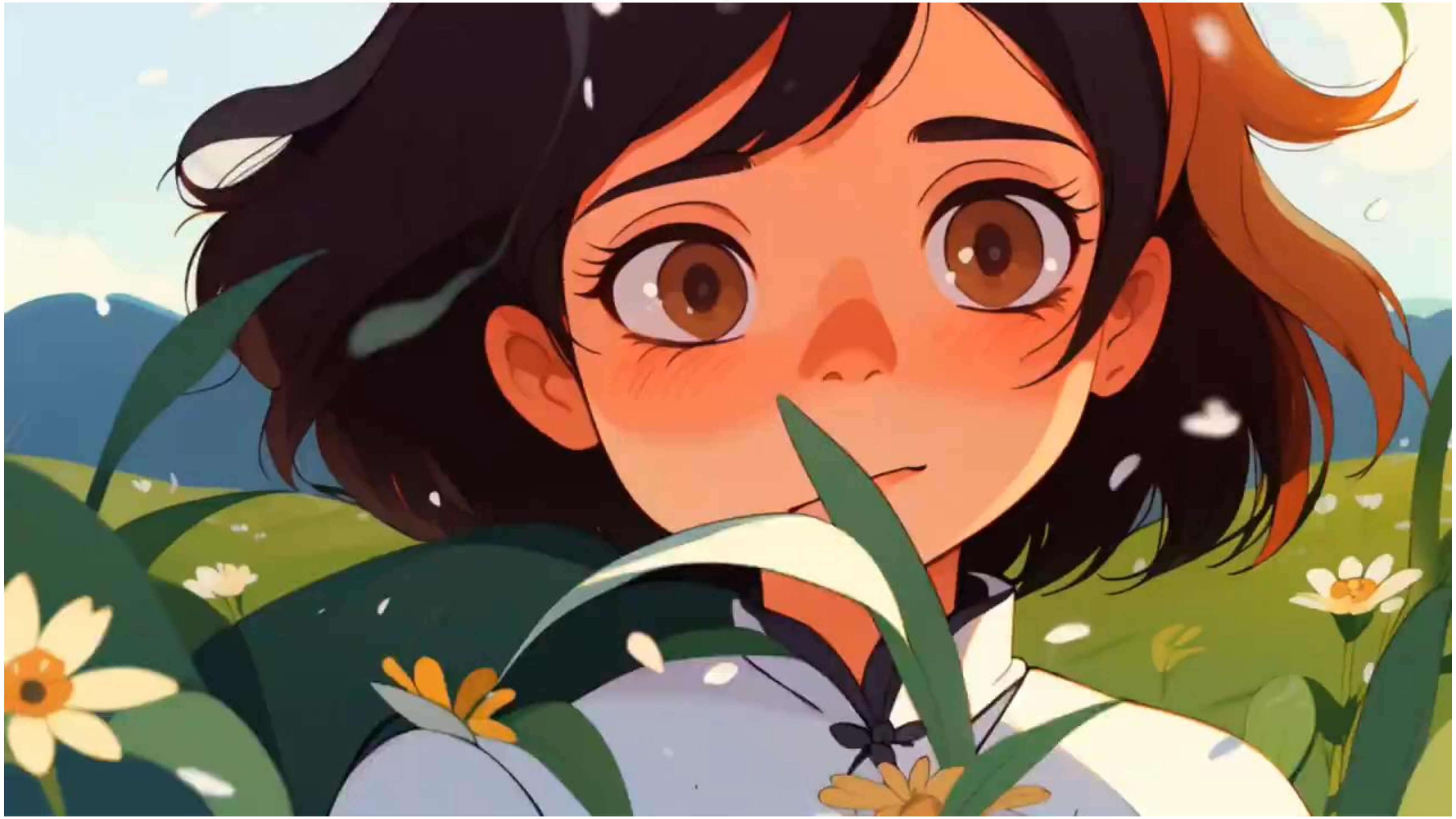



"A chihuahua in astronaut suit floating in space, cinematic lighting, glow effect"

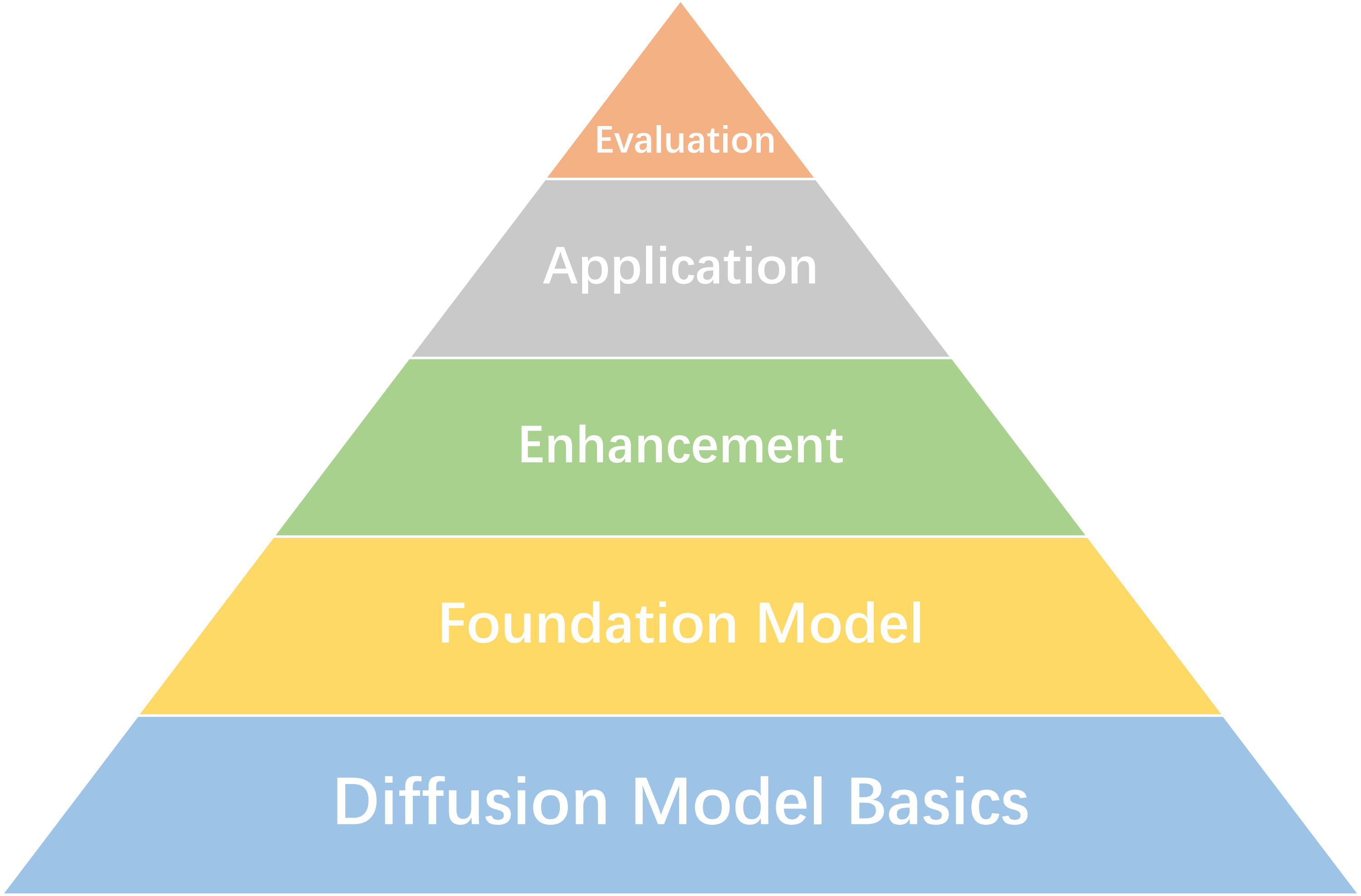



"A very happy fuzzy panda dressed as a chef eating pizza in the New York street food truck"

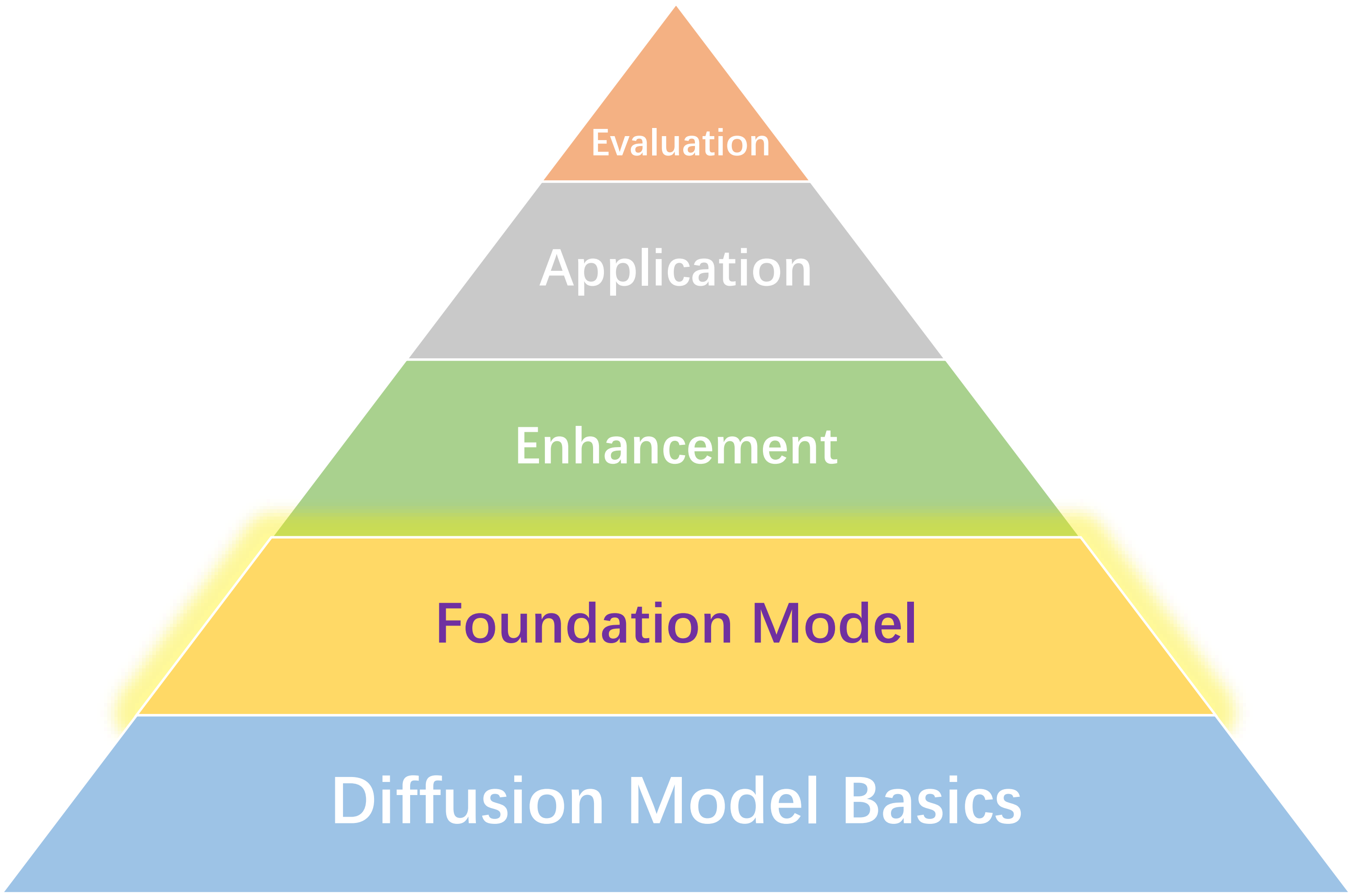
Results



Video Generation



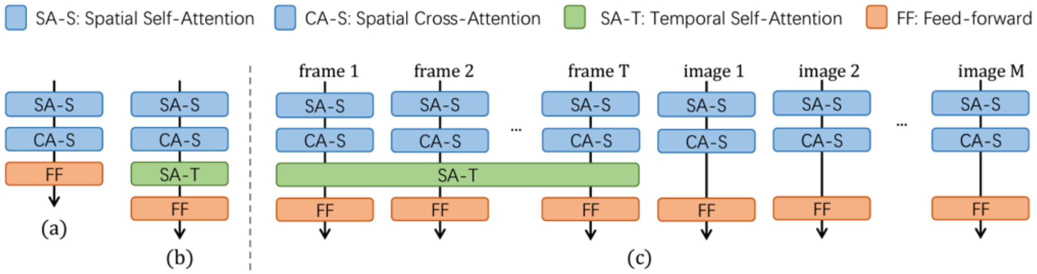
Video Generation



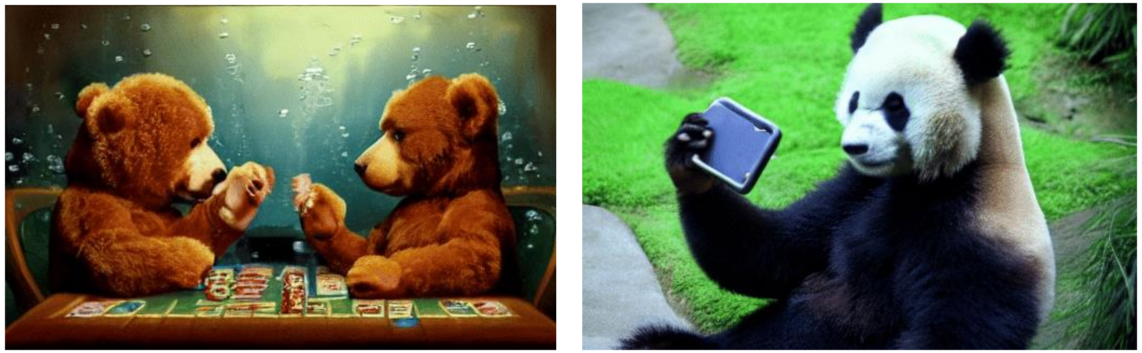
Vchitect: A Large-scale Video Generation System

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

- **Storytelling, multiple-shots, minute-level 4K video generation**
- Achieves smooth transitions, cohesive storytelling, high-definition quality, leading across various metrics



Architecture

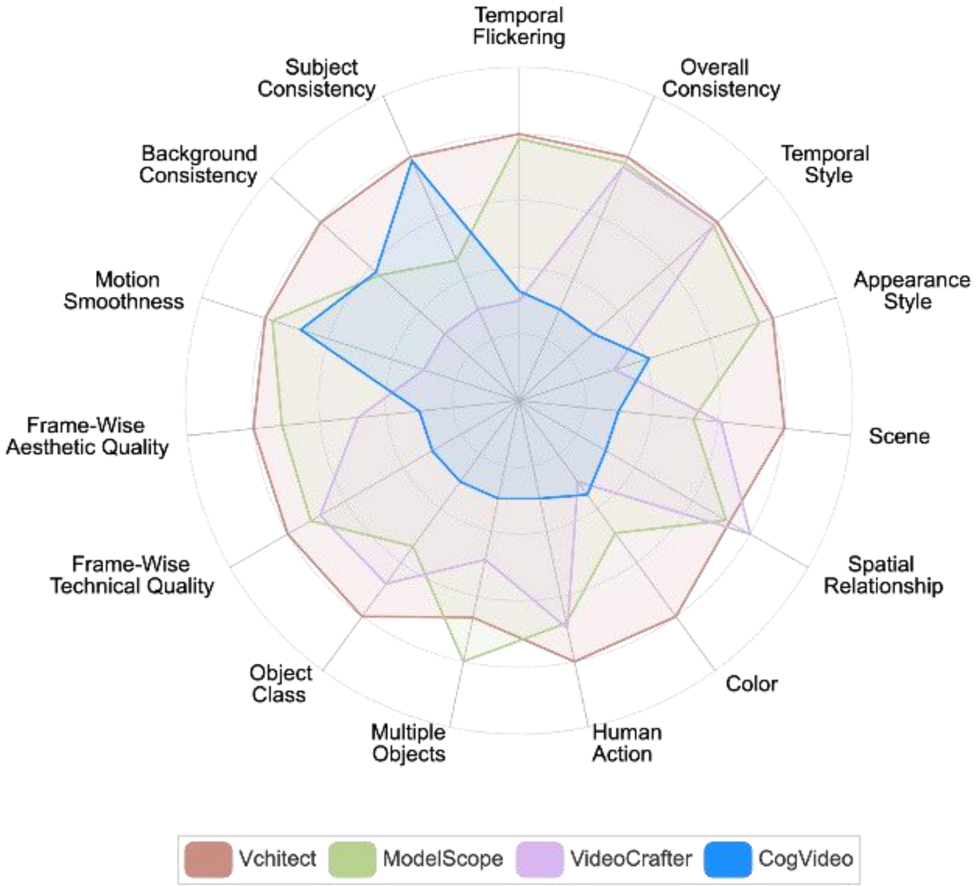


Text to Video

Long Video Generation

Image to Video – Transition & Animation

Vchitect: A Large-scale Video Generation System



Comparison with Open-sourced Models

Vchitect

EmuVideo

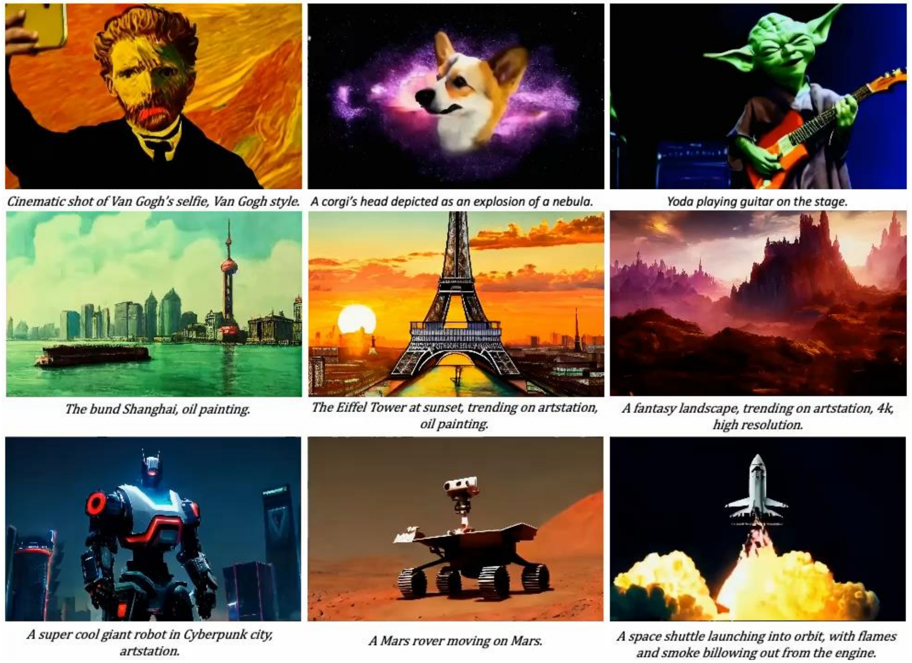
Vchitect

Lumiere

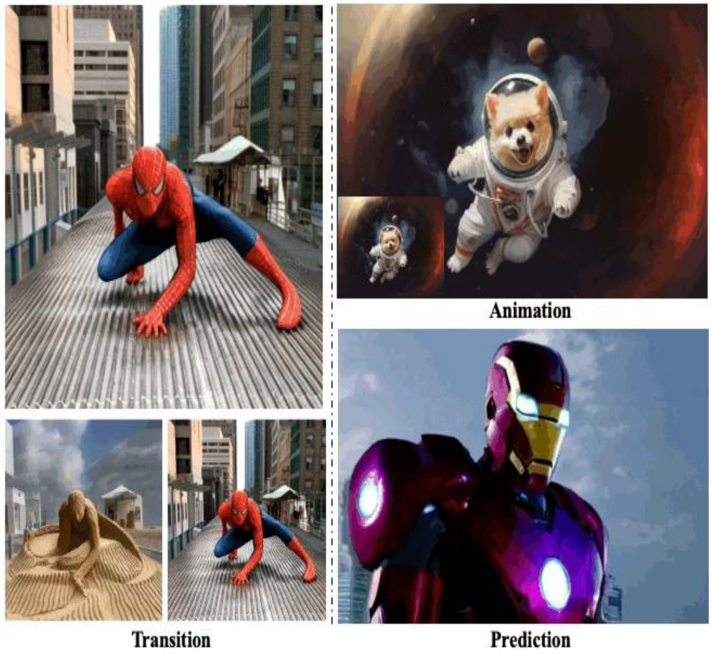
Comparison with Close-sourced Models

Vchitect: A Large-scale Video Generation System

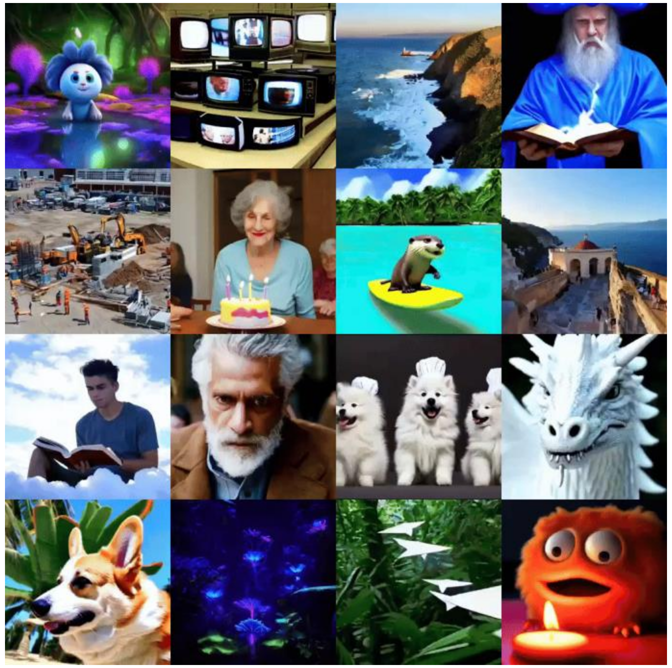
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE



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



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

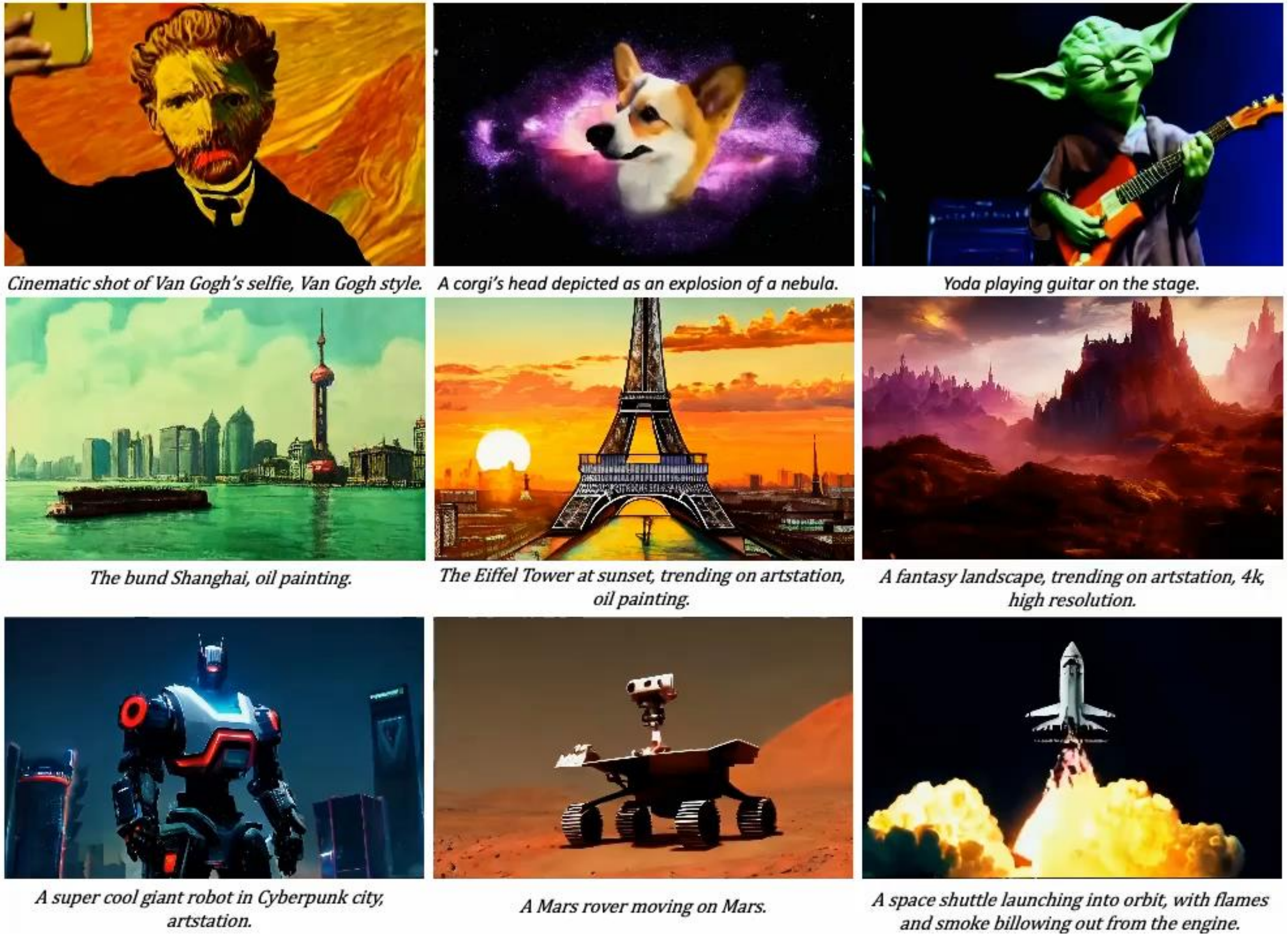


LATTE [Ma, Wang *et al.*, arXiv'24]
Latent Diffusion Transformer

LaVie

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

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.

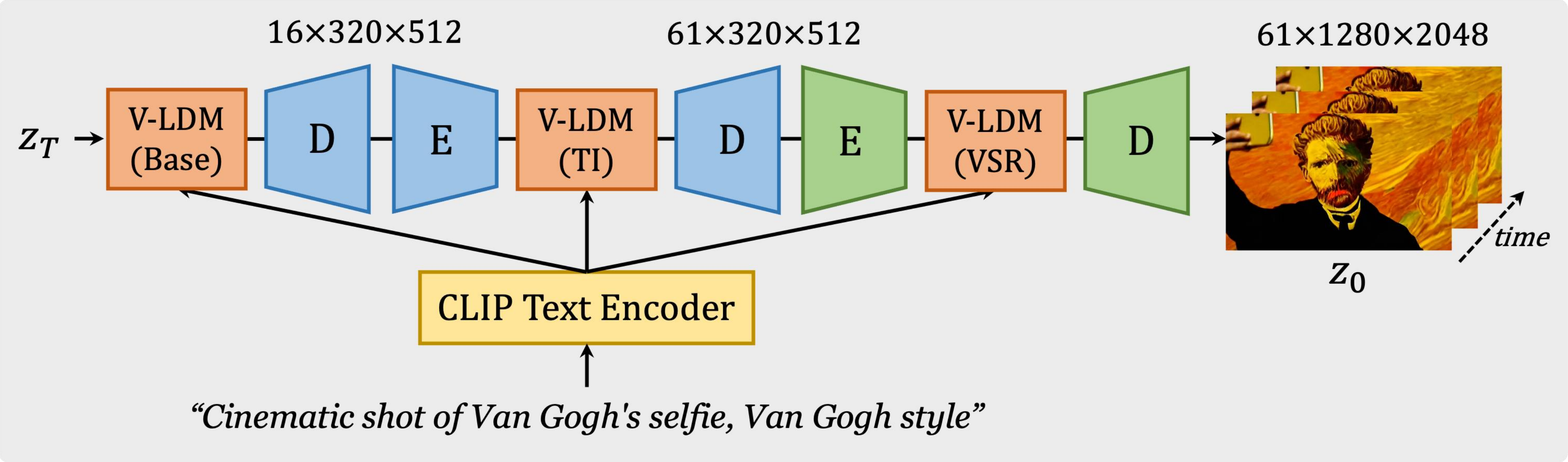
A fantasy landscape, trending on artstation, 4k, high resolution.

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.

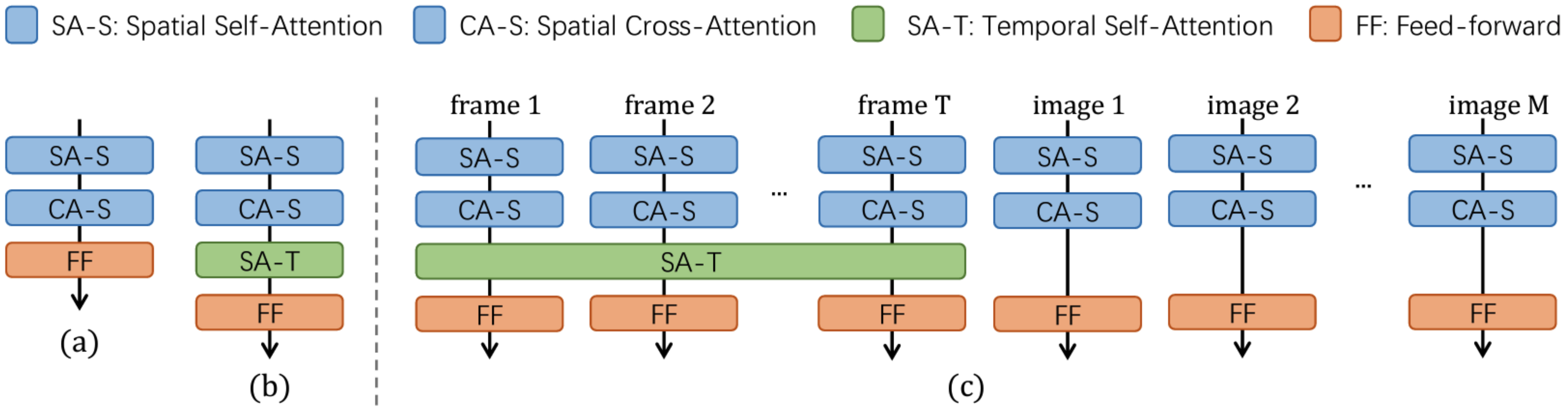
LaVie – Model Design



<u>A cascaded video generation system:</u>

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

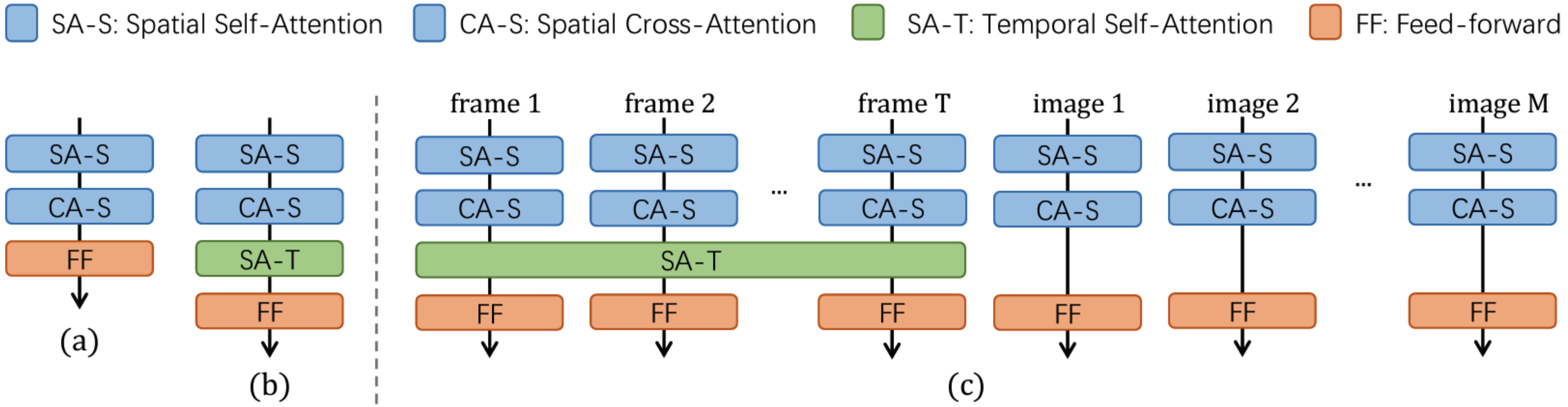
LaVie – Architecture



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

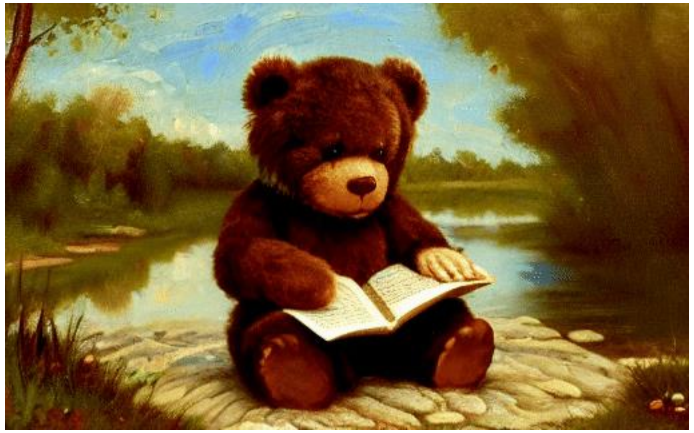
Videos from Vimdeo25M dataset

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)
    - More detailed captions (provided by VideoChat)
    - Higher resolution, 1080p, better visual quality
    - Better aesthetics

LaVie – More results

Two teddy bears playing poker under water

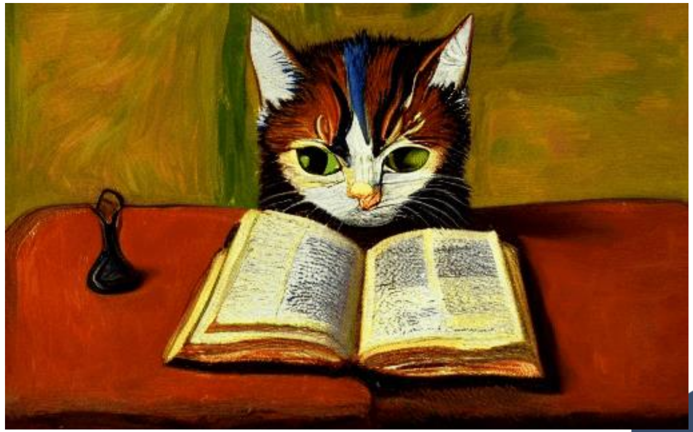
a teddy bears skateboarding under water



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

Elon Musk standing besides a rocket

Iron Man flying in the sky

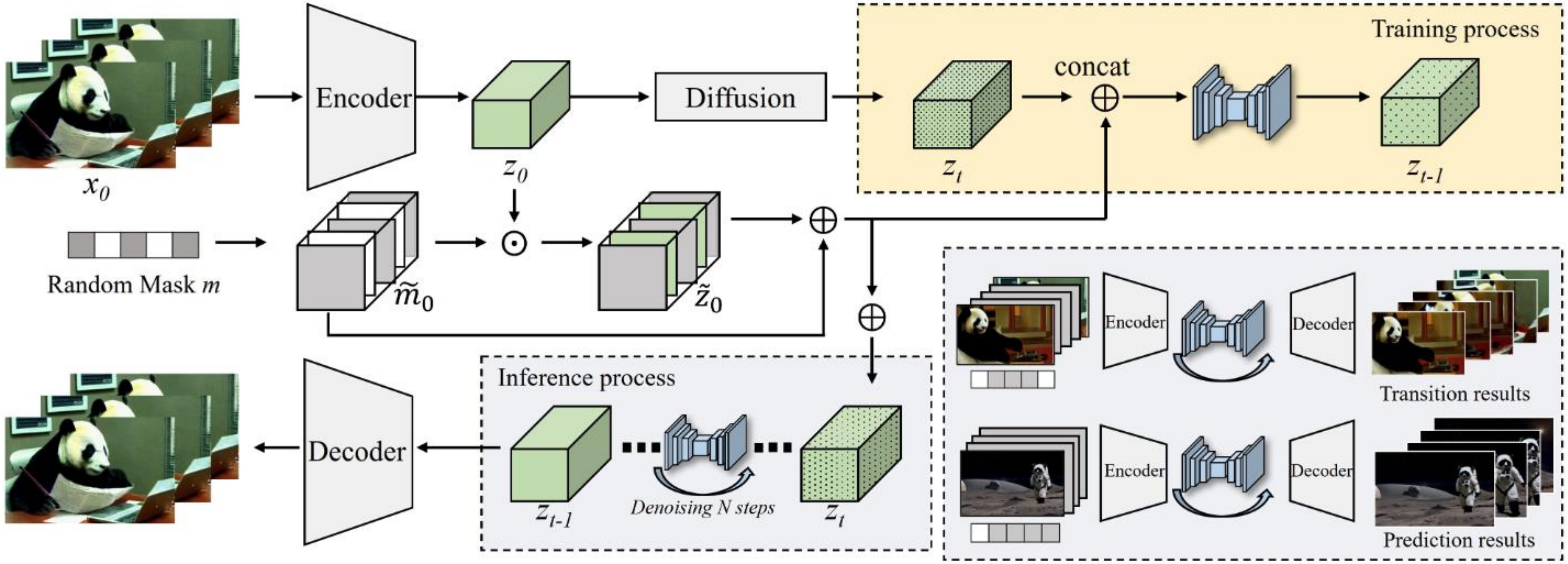


a cat reading a book, Van Gogh style

SEINE

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

SEINE – Architecture & Learning



Training

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

Inference:

Different masks →

Transition, Animation, Prediction

SEINE – More results


Image-to-video generation

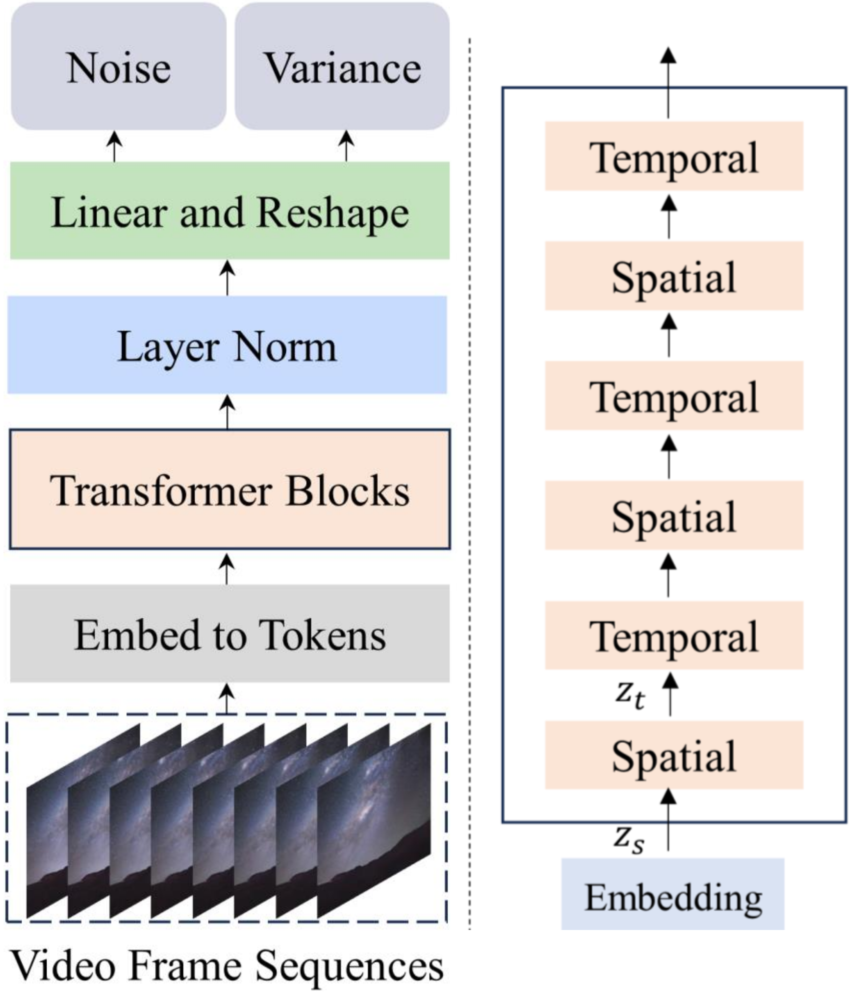


Transition

Story-based Long Video Generation (LaVie + SEINE)

Latte: Latent Diffusion Transformer

A diffusion transformer for general video generation



Latte architecture

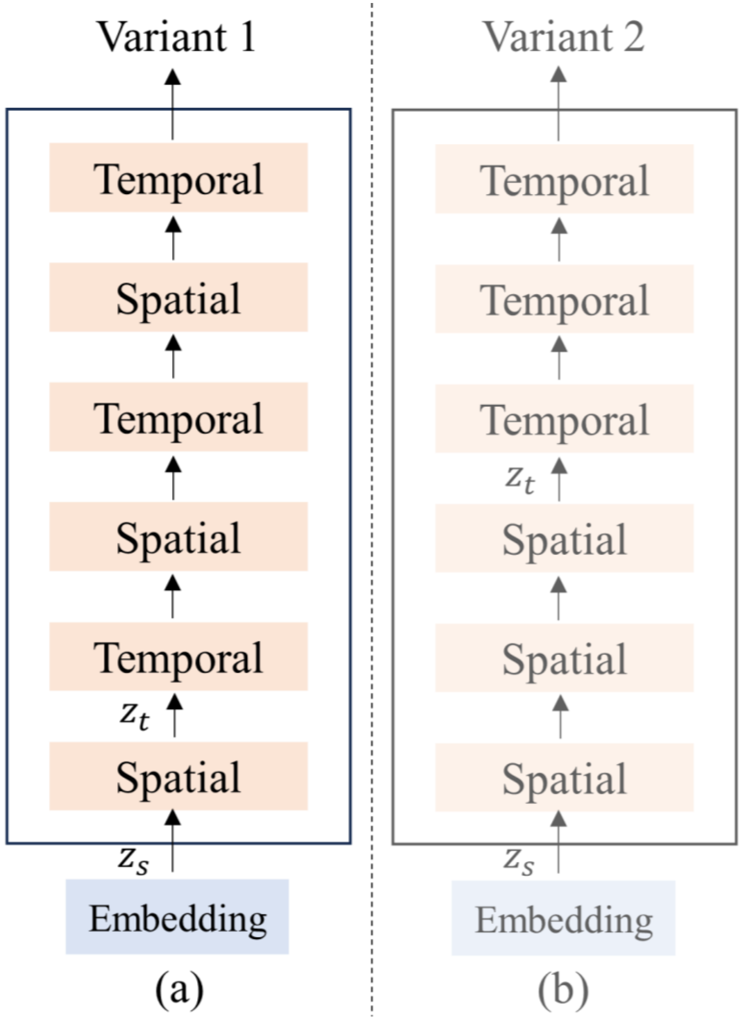
We introduce:

1. Model architecture designs

2. Transformer designs

3. Best practices in model and training

Latte – Model design

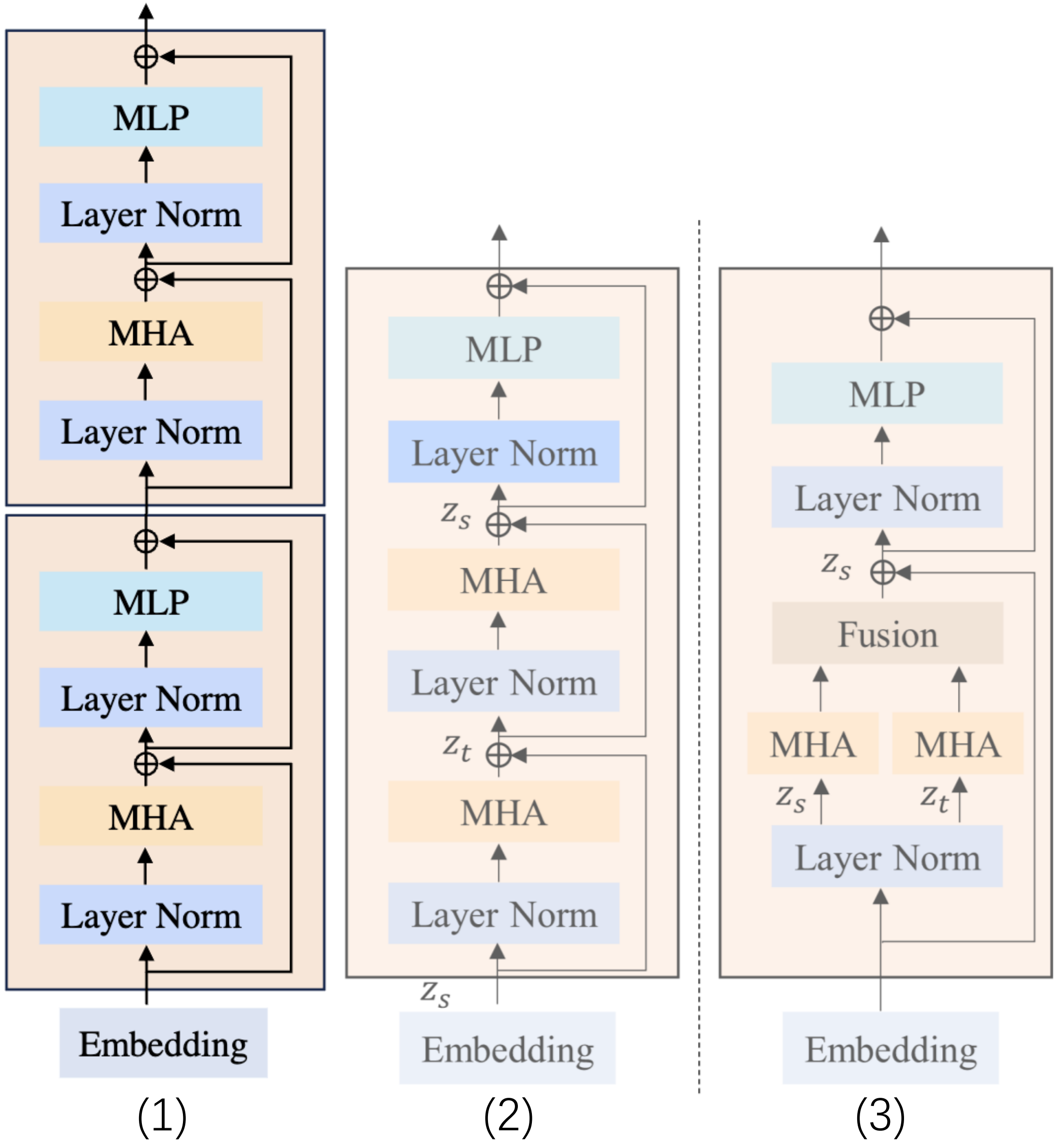


Variant 1:
(Spatial + Temporal) x N blocks

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

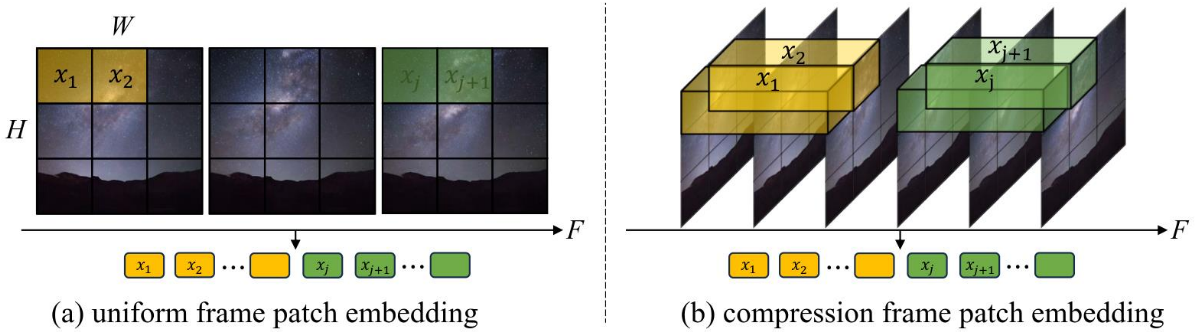
3. Joint spatio-temporal transformer block

- Parallel spatial and temporal attentions

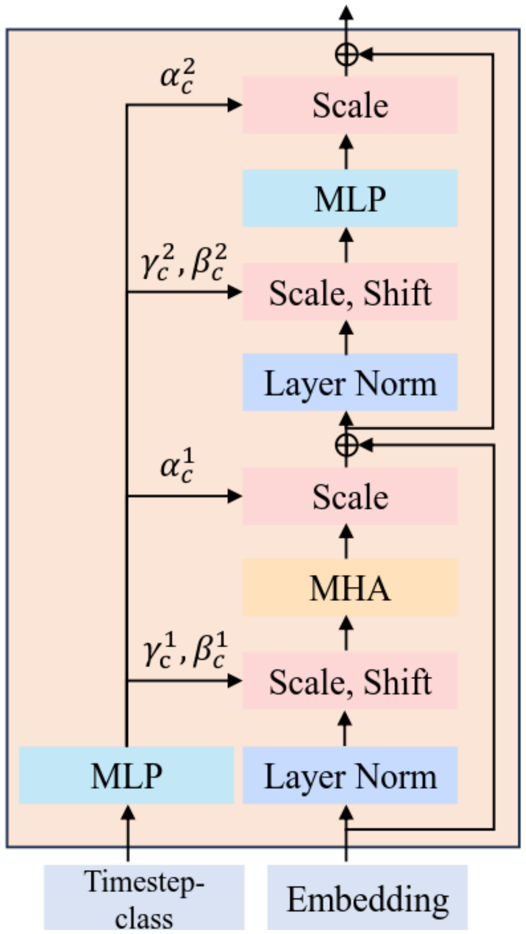
Latte – Best Practice Design

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**

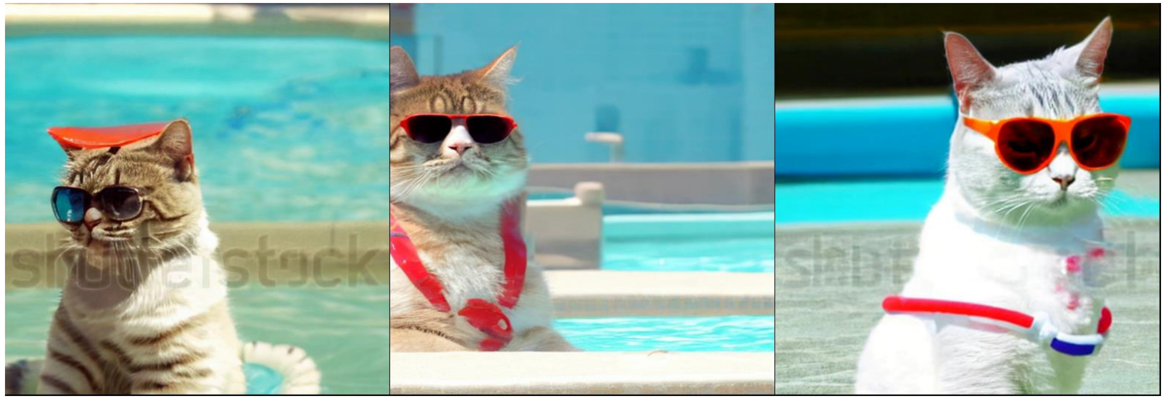
Video-level quality comparison

Latte – Results

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.

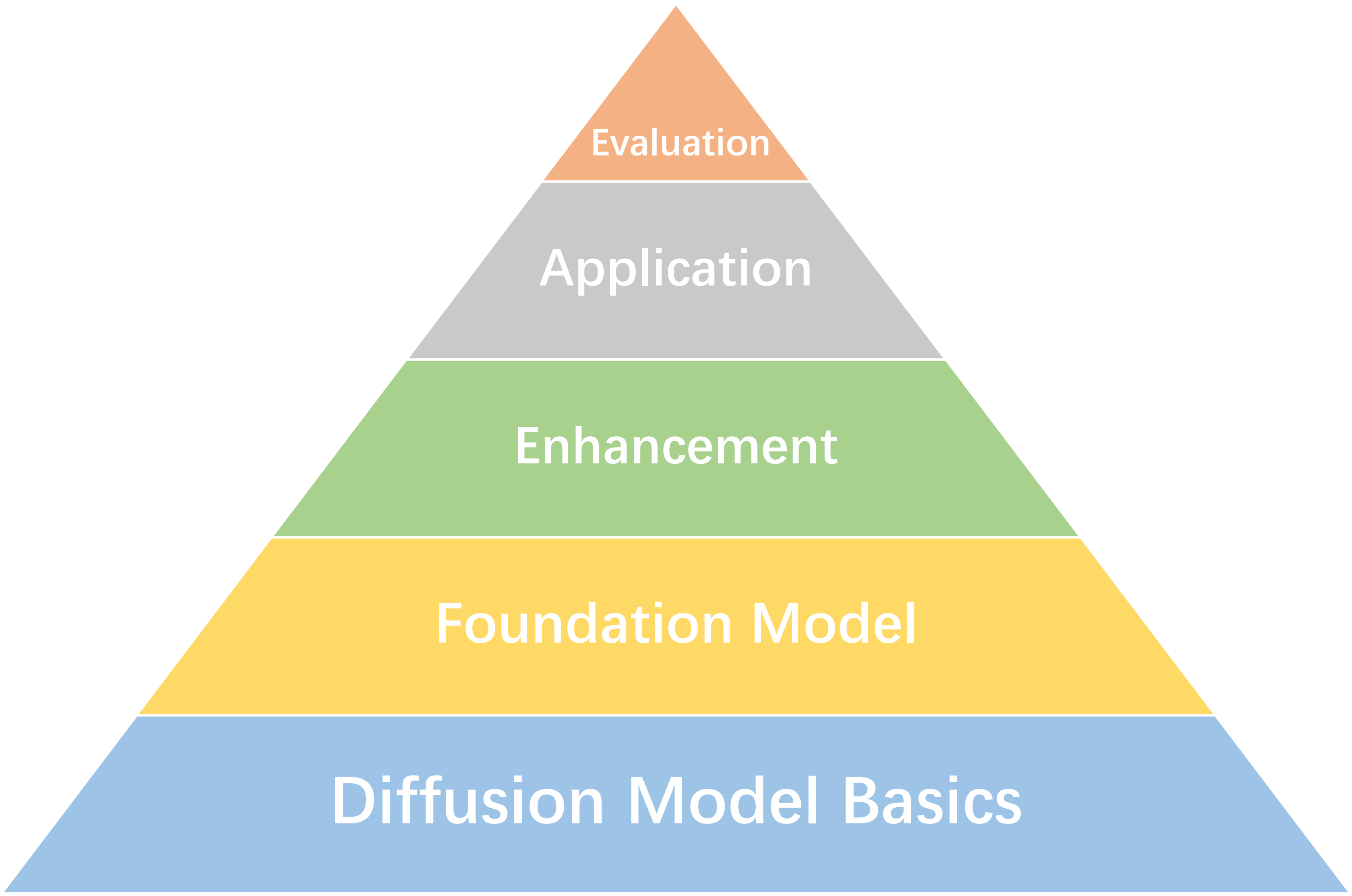


a cat wearing sunglasses and working as a lifeguard at pool.

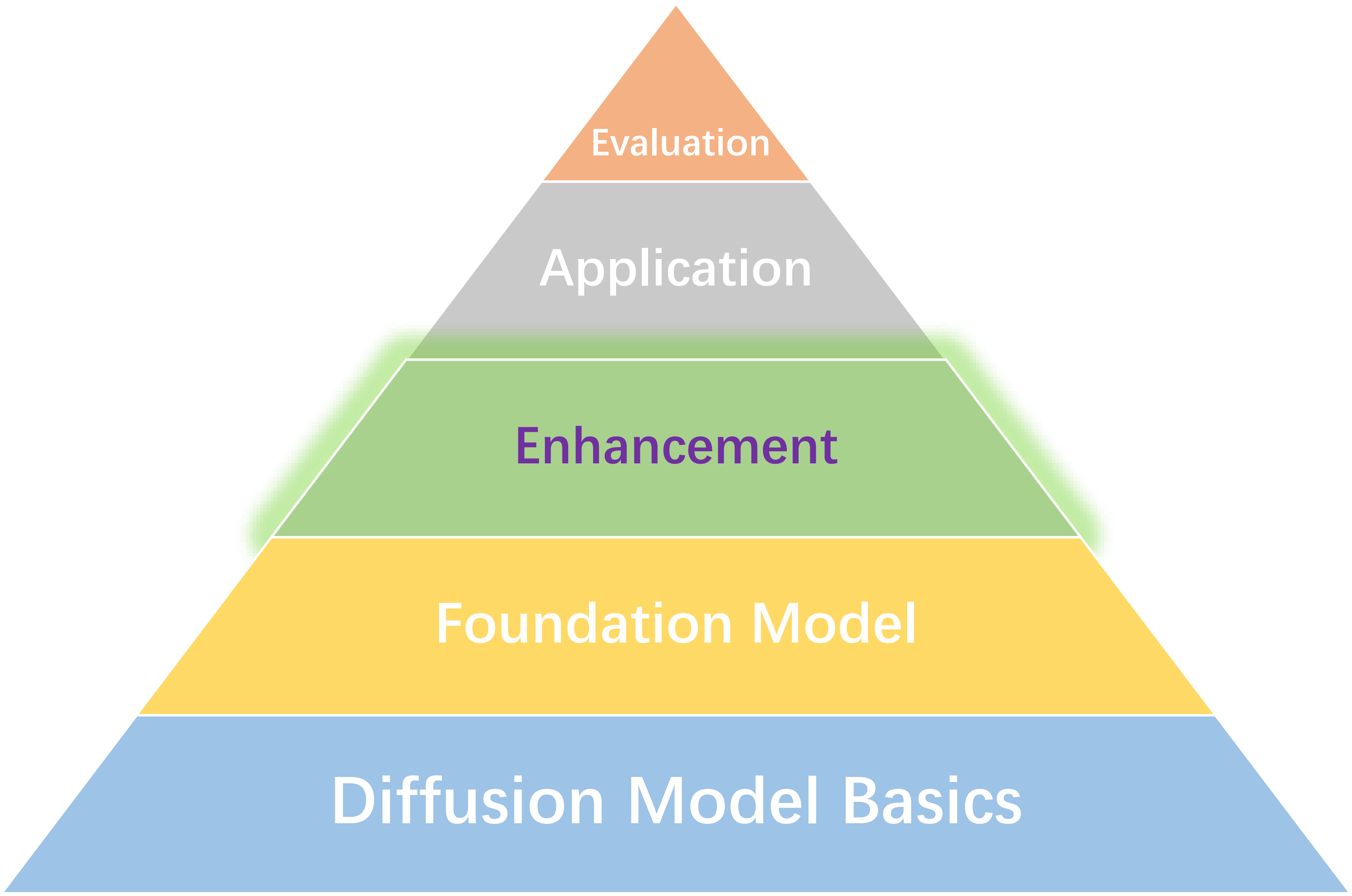
Vchitect Foundation Models

Vchitect Foundation Models

Video Generation



Video Generation



VEnhancer

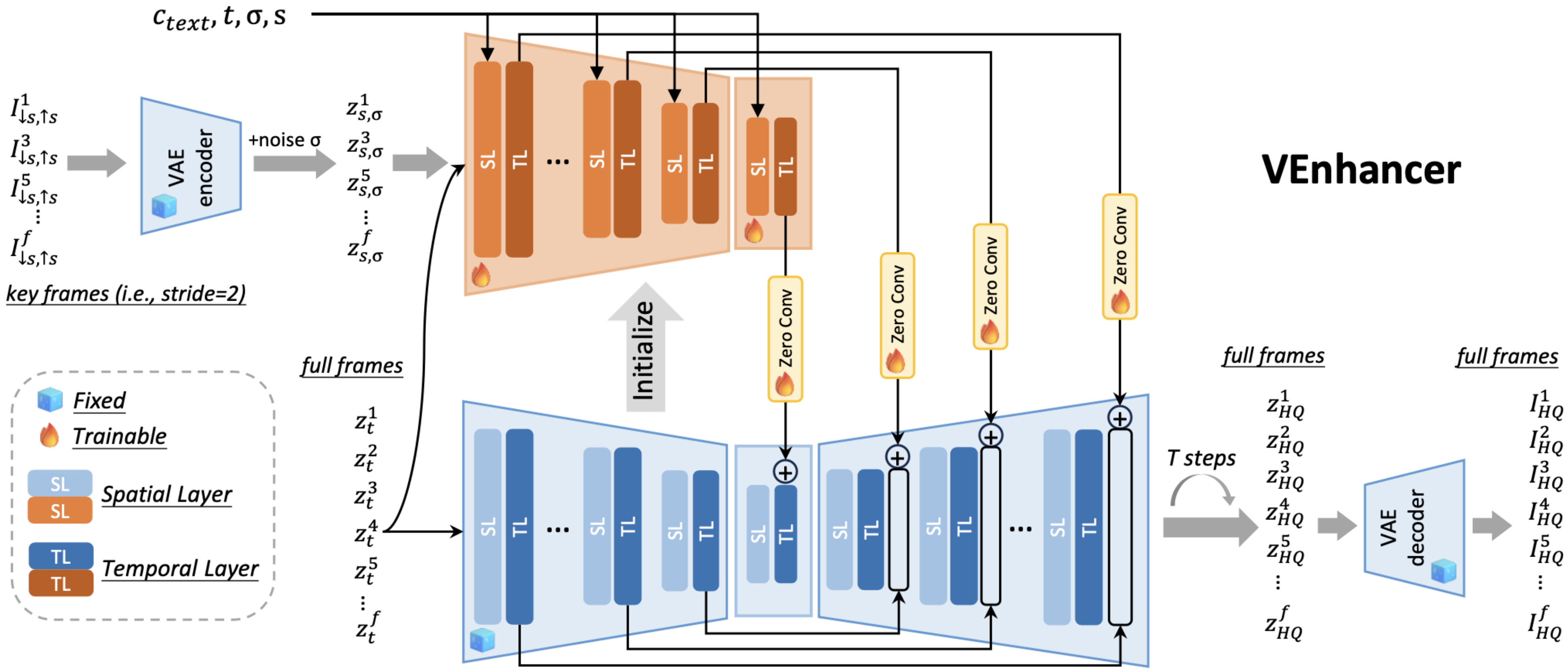
Generative Space-Time Enhancement for Video Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Clown fish swimming through the coral reef.

- A Unified model for generative spatial super-resolution (S-SR), temporal super-resolution (T-SR), and video refinement.
- Support arbitrary upsampling factors for S-SR and T-SR, as well as flexible control to modify refinement strength.

VEnhancer – Architecture



- Base model: Pretrained Video diffusion model (blue part), fixed.
- Condition network: Video ControlNet (orange part), finetuned.

VEnhancer – Results

Iron Man flying in the sky.

VEnhancer outperforms state-of-the-art video super-resolution methods and space-time super-resolution methods in enhancing AI-generated videos.

VEnhancer – Results

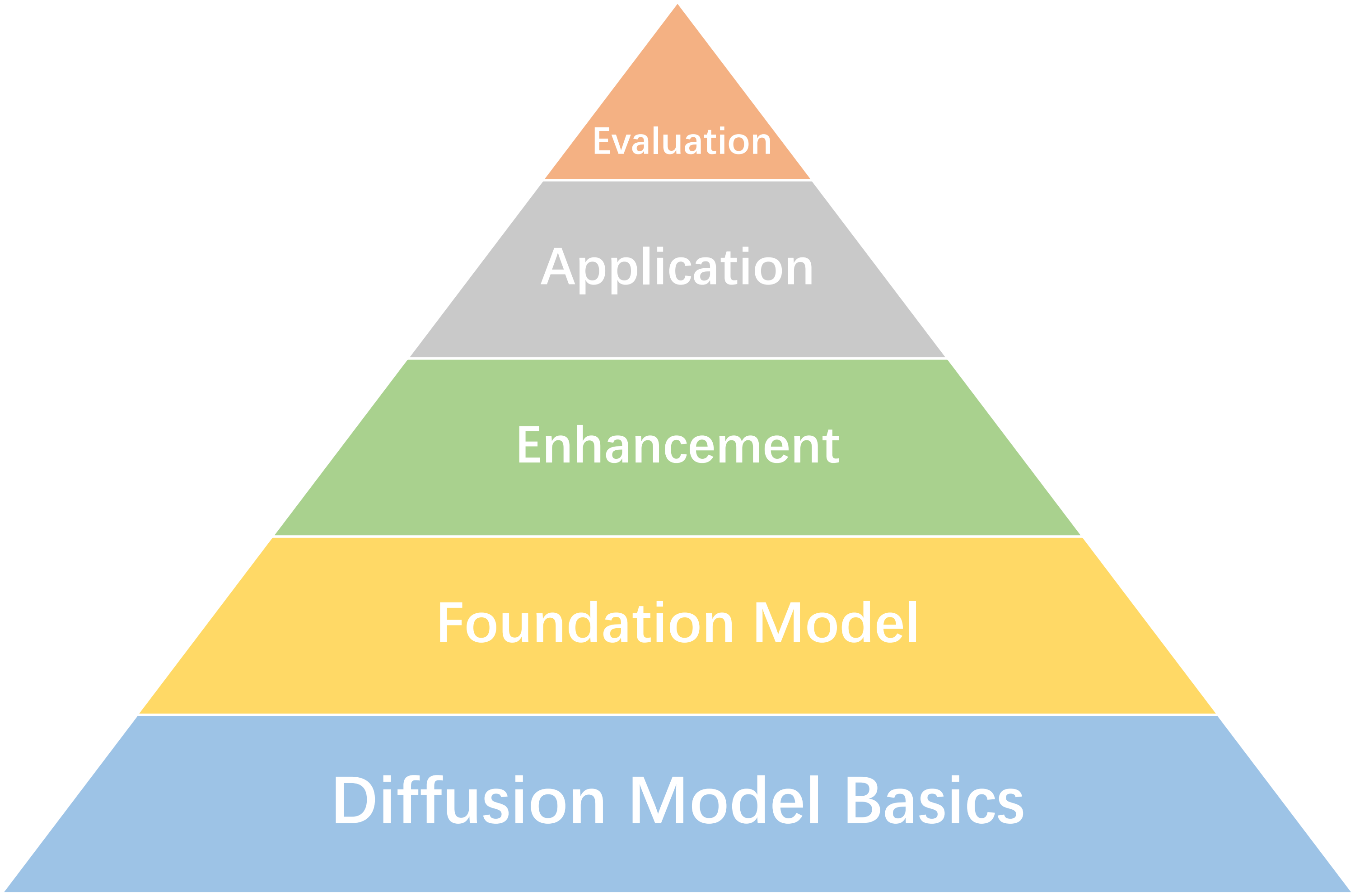
	Dimensions	Show-1 [46]	LaVie [40]	Open-Sora	Pika	Gen-2	VC-2 [9]	VC-2+VEnhancer
Quality	**Subject Consistency**	95.53%	91.41%	92.09%	96.76%	97.61%	96.85%	97.17%
	Background Consistency	98.02%	97.47%	97.39%	98.95%	97.61%	98.22%	98.54%
	Temporal Flickering	99.12%	98.30%	98.41%	99.77%	99.56%	98.41%	98.46%
	Motion Smoothness	98.24%	96.38%	95.61%	99.51%	99.58%	97.73%	97.75%
	Aesthetic Quality	57.35%	54.94%	57.76%	63.15%	66.96%	63.13%	65.89%
	Dynamic Degree	44.44%	49.72%	48.61%	37.22%	18.89%	42.50%	42.50%
	Imaging Quality	58.66%	61.90%	61.51%	62.33%	67.42%	67.22%	70.45%
Semantic	**Object Class**	93.07%	91.82%	74.98%	87.45%	90.92%	92.55%	93.39%
	Multiple Objects	45.47%	33.32%	33.64%	46.69%	55.47%	40.66%	49.83%
	Human Action	95.60%	96.80%	85.00%	88.00%	89.20%	95.00%	95.00%
	Color	86.35%	86.39%	78.15%	85.31%	89.49%	92.92%	94.41%
	Spatial Relationship	53.50%	34.09%	43.95%	65.65%	66.91%	35.86%	64.88%
	Scene	47.03%	52.69%	37.33%	44.80%	48.91%	55.29%	51.82%
	Appearance Style	23.06%	23.56%	21.58%	21.89%	19.34%	25.13%	24.32%
	Temporal Style	25.28%	25.93%	25.46%	24.44%	24.12%	25.84%	25.17%
	Overall Consistency	27.46%	26.41%	26.18%	25.47%	26.17%	28.23%	27.57%
Overall	**Quality**	80.42%	78.78%	78.82%	82.68%	82.46%	82.20%	**83.28%**
	Semantic	72.98%	70.31%	64.28%	71.26%	73.03%	73.42%	**76.73%**

With VEnhancer, VideoCrafter-2 [1] achieves the top one in VBench in both *semantic* and *quality*, outperforming professional video generation products, Gen-2 and Pika.

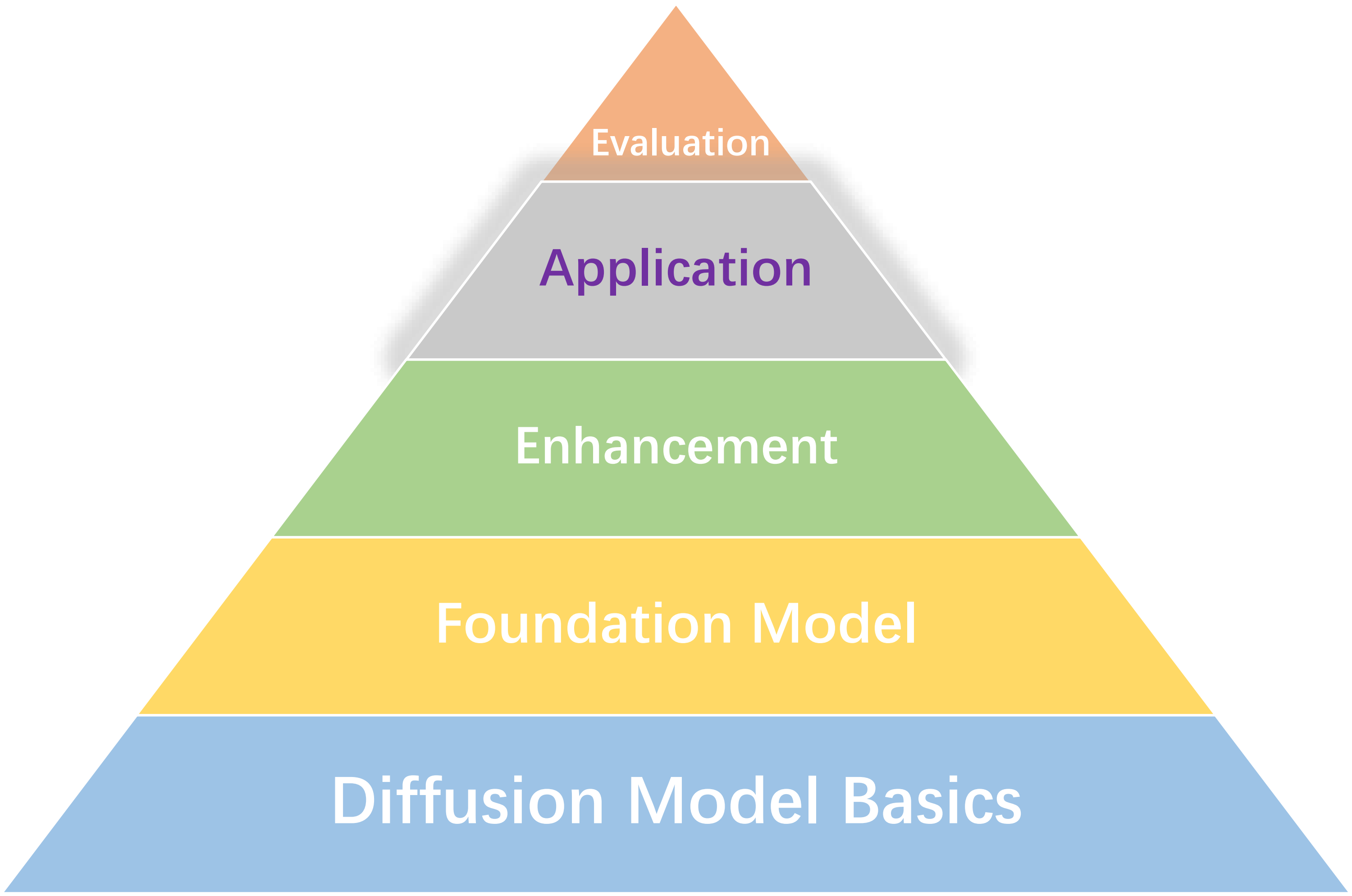
[1] Chen et.al. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

An astronaut is riding a horse in the space in a photorealistic style.

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

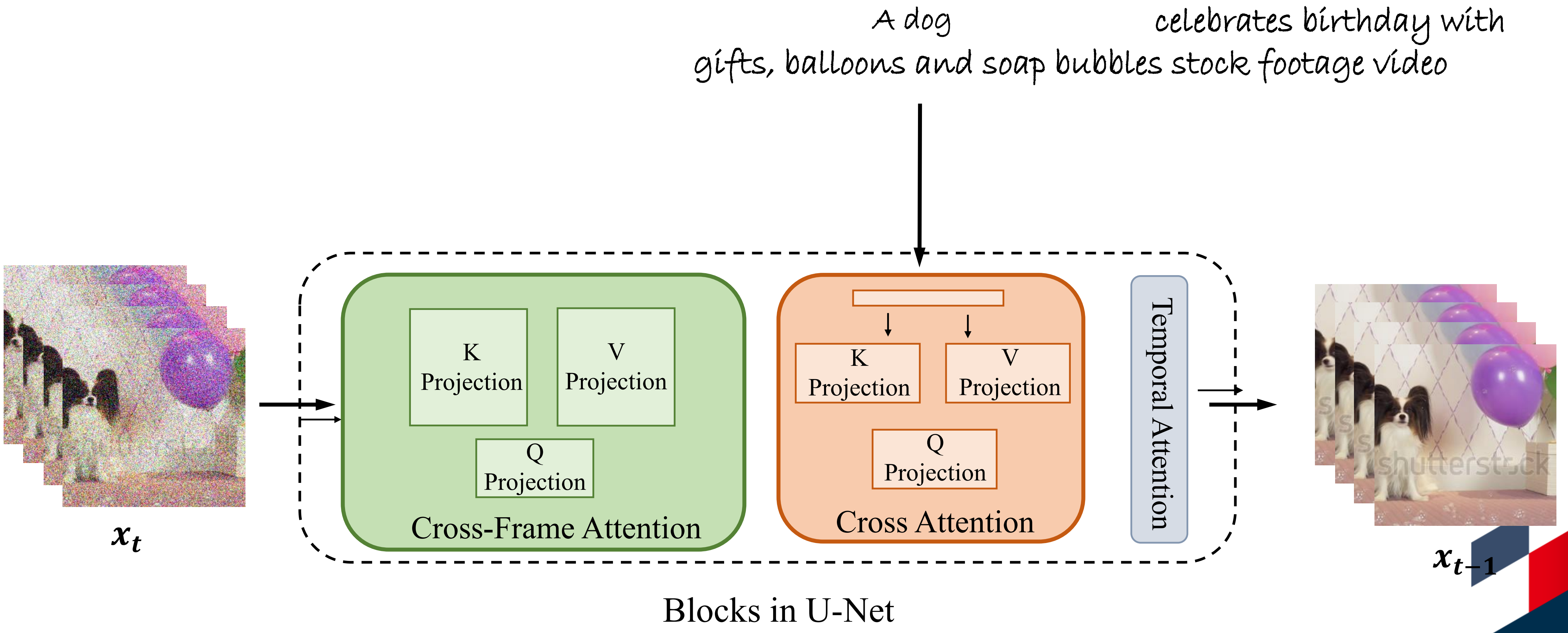
Dog drinking from bowl of water

Dog swimming in lake happily

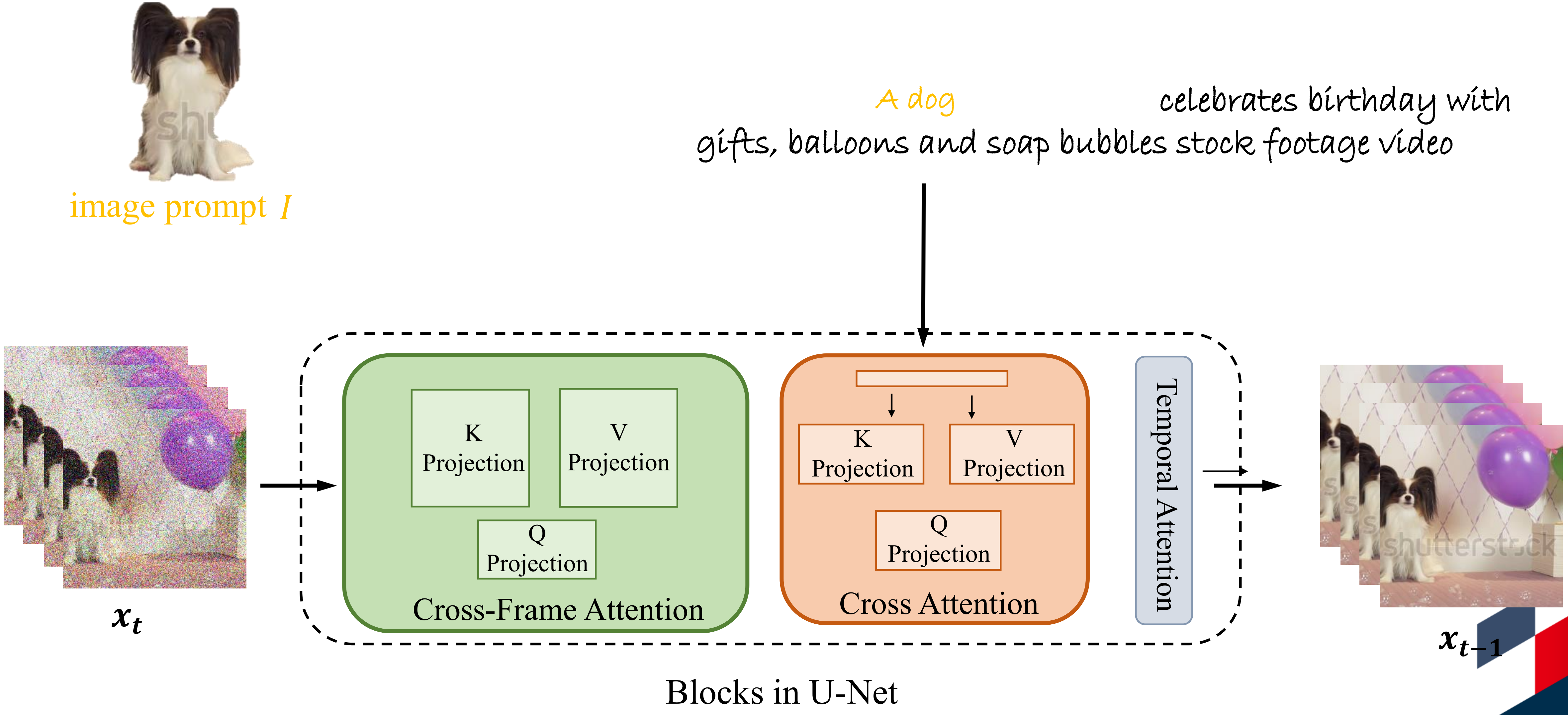
Dog in park

Portrait of a dog, looks out the car window

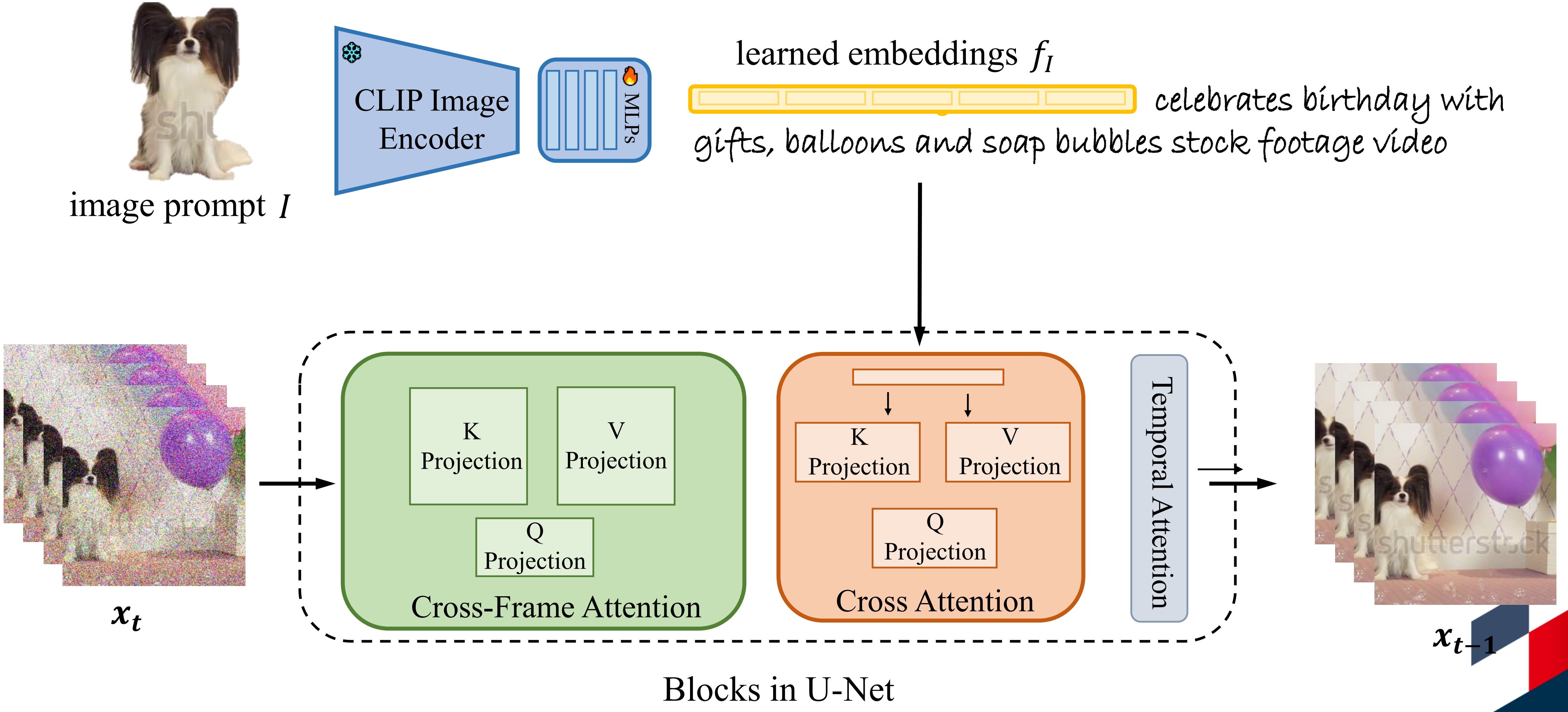
VideoBooth - Method



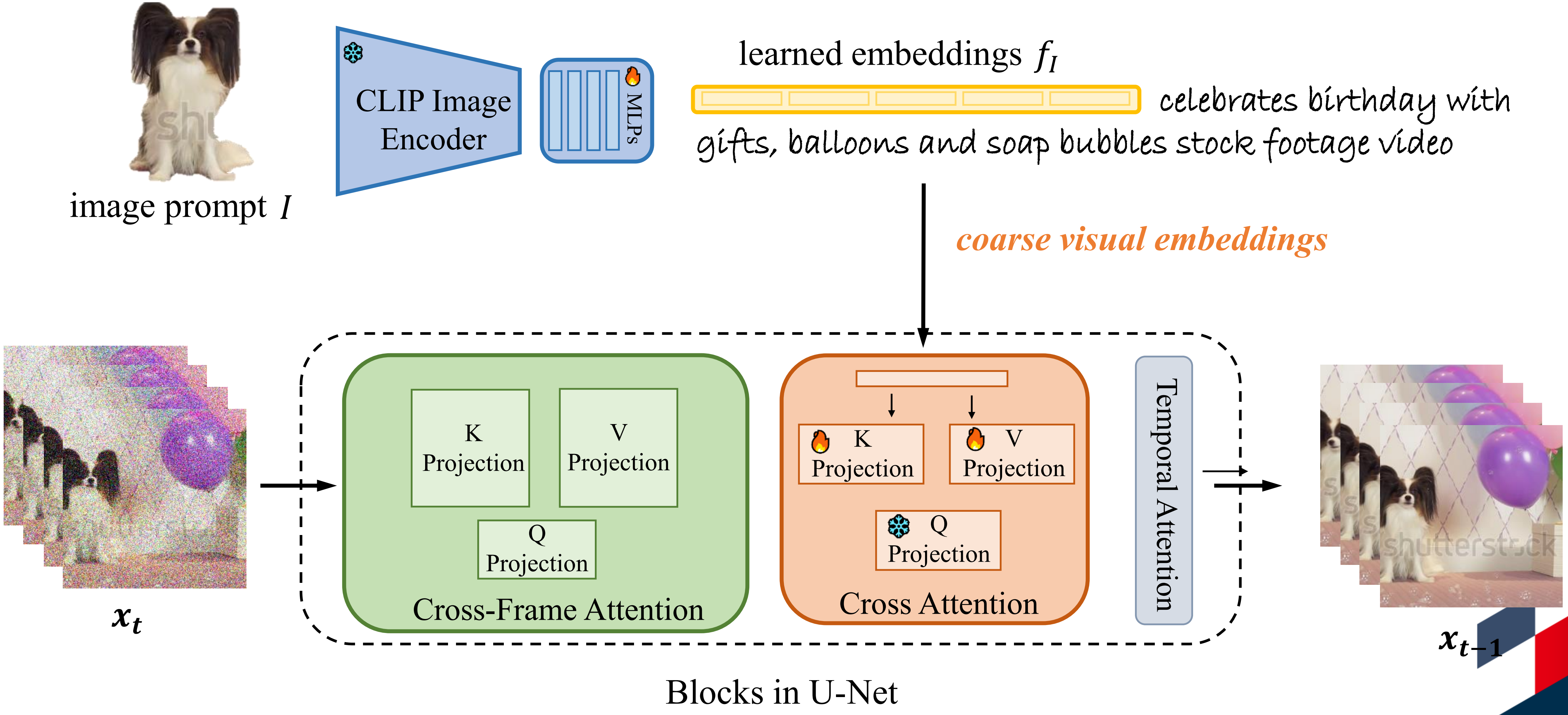
VideoBooth - Method



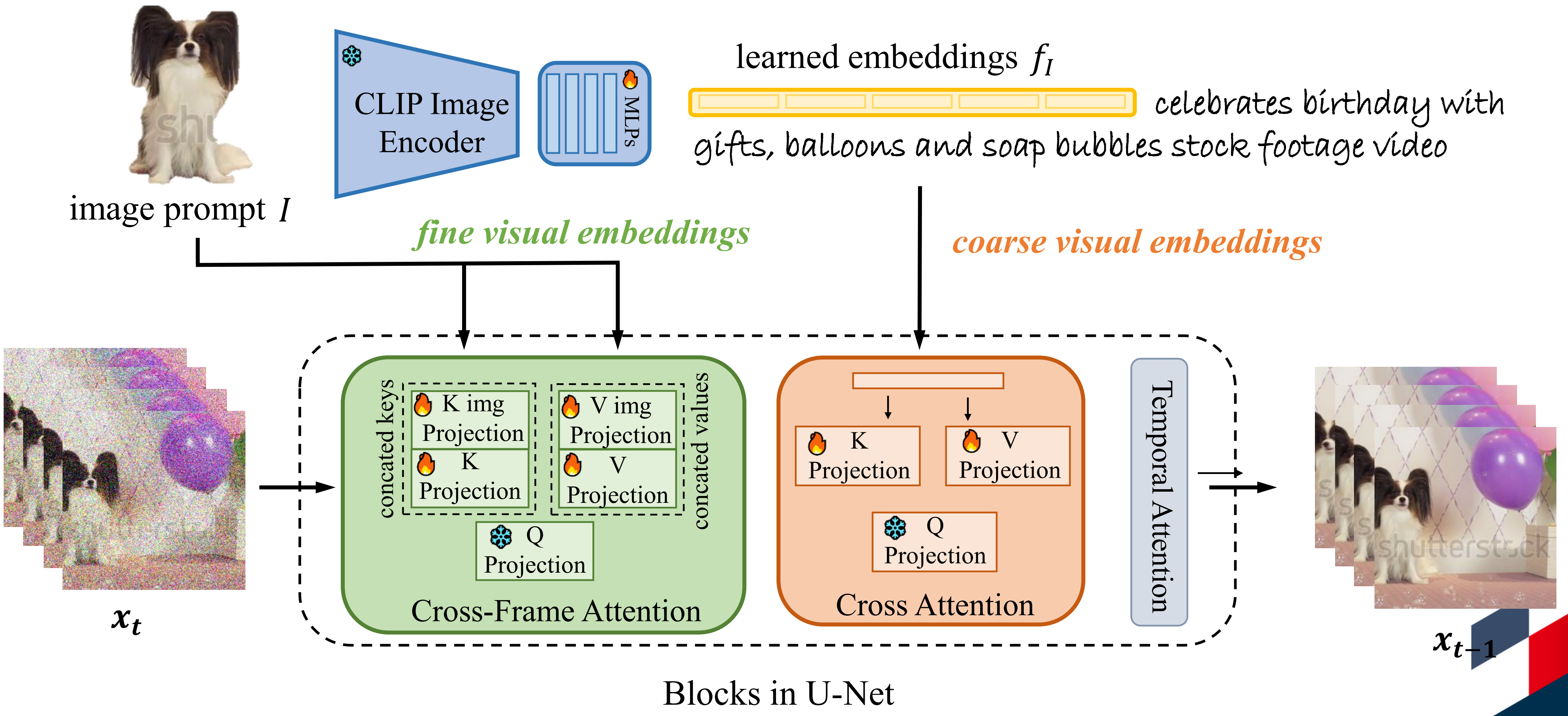
VideoBooth - Method



VideoBooth - Method



VideoBooth - Method



VideoBooth - Results

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Image Prompt

Text Prompt

dog laying on ground

Textual Inversion

DreamBooth

ELITE

VideoBooth (Ours)

VideoBooth - Results

Image Prompt

Text Prompt

close up of cat on top of a vintage chair

Textual Inversion

DreamBooth

ELITE

VideoBooth (Ours)

VideoBooth - Results

Image Prompt

Textual Inversion

DreamBooth

Text Prompt

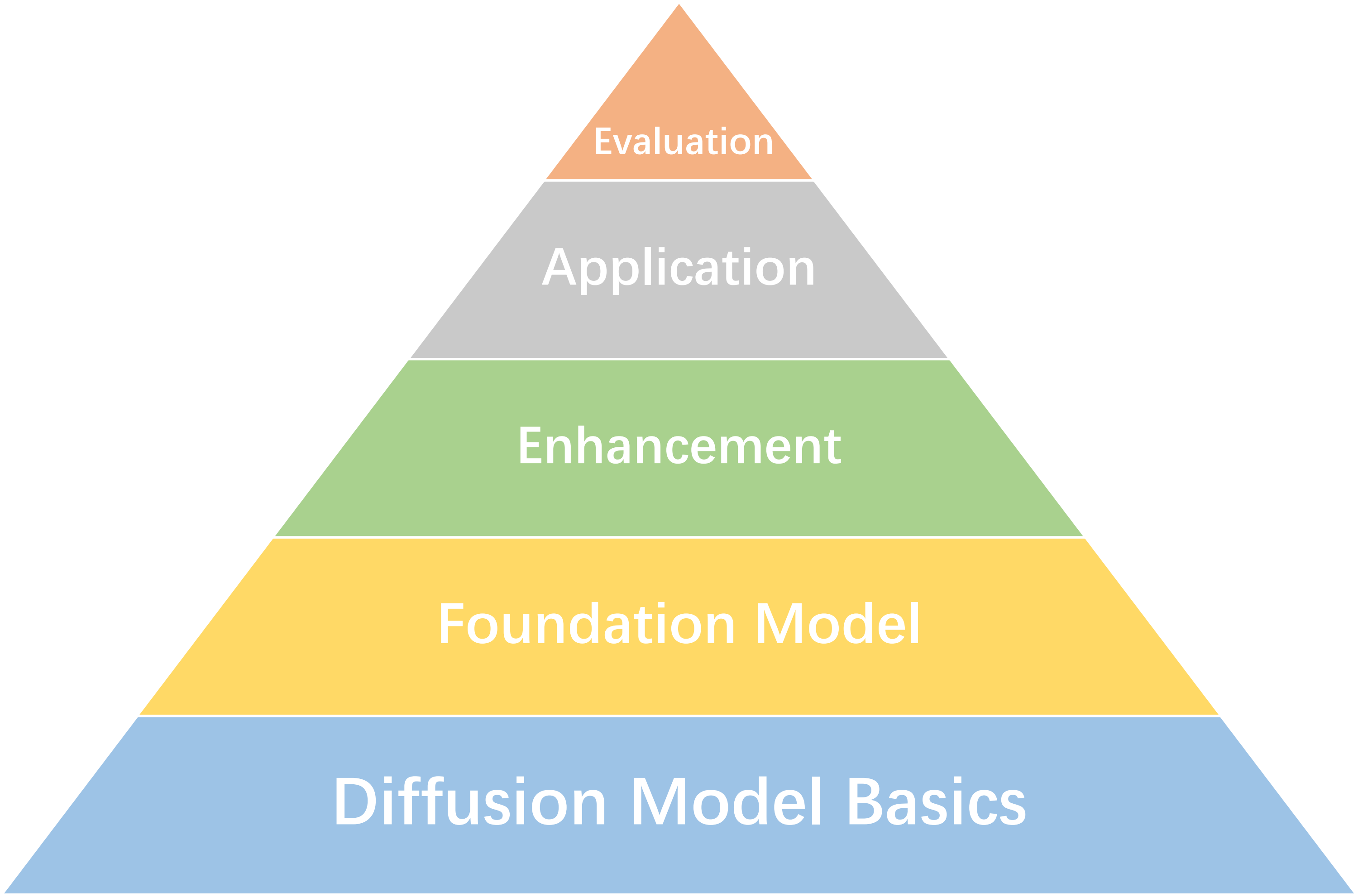
car in the bush

ELITE

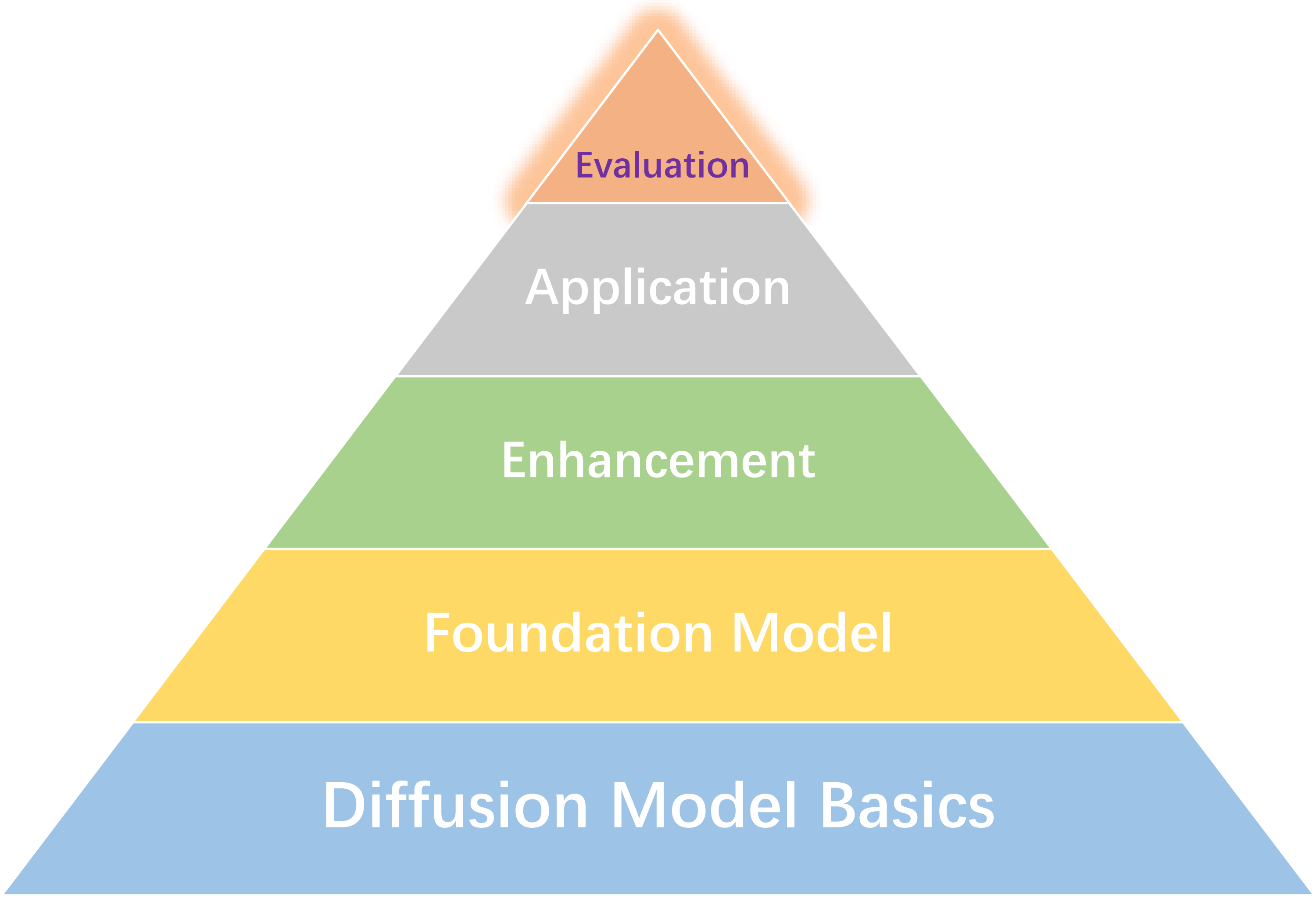
VideoBooth (Ours)

Video Generation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

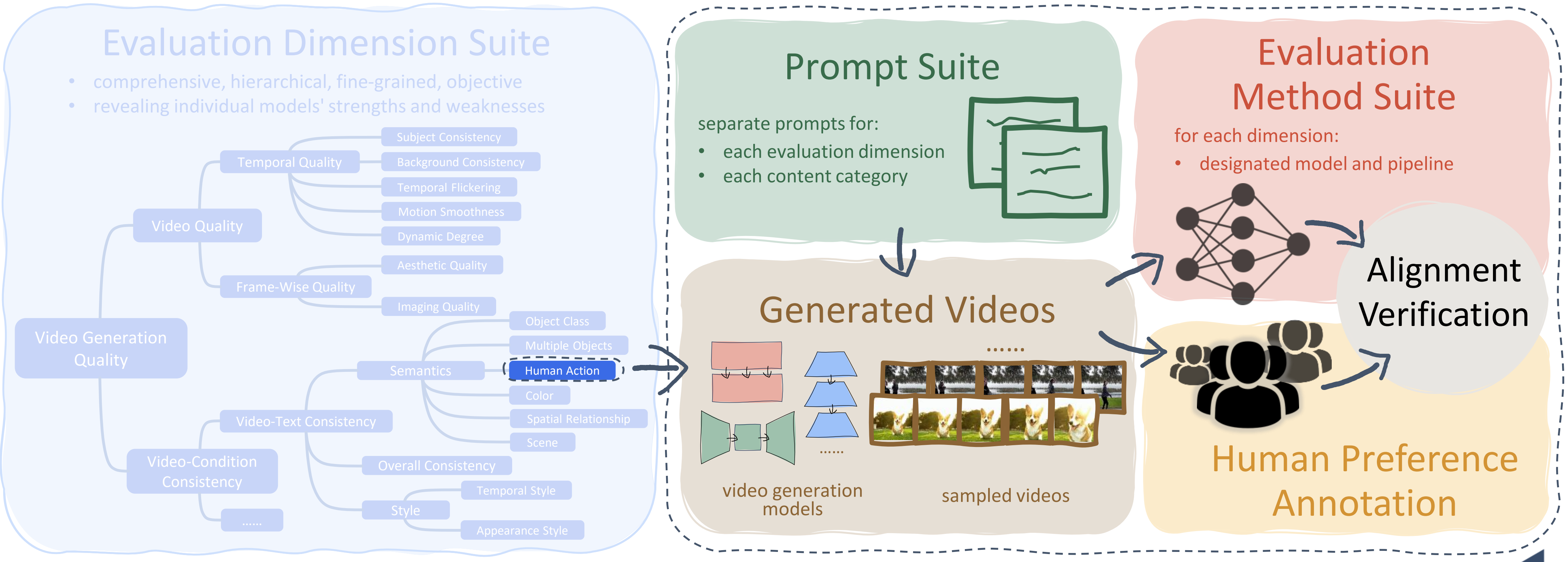


Video Generation

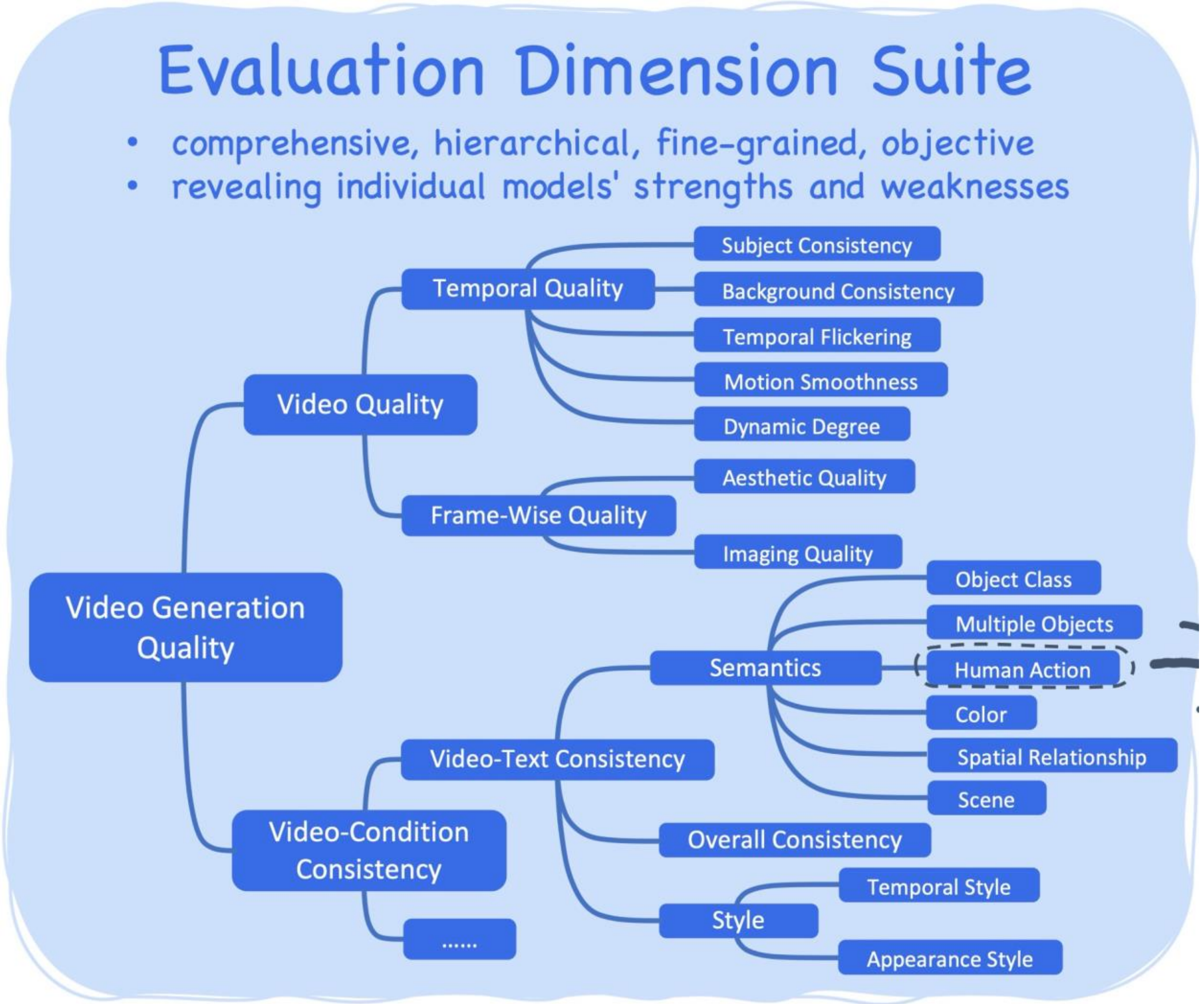


VBench

Comprehensive Benchmark Suite for Video Generative Models

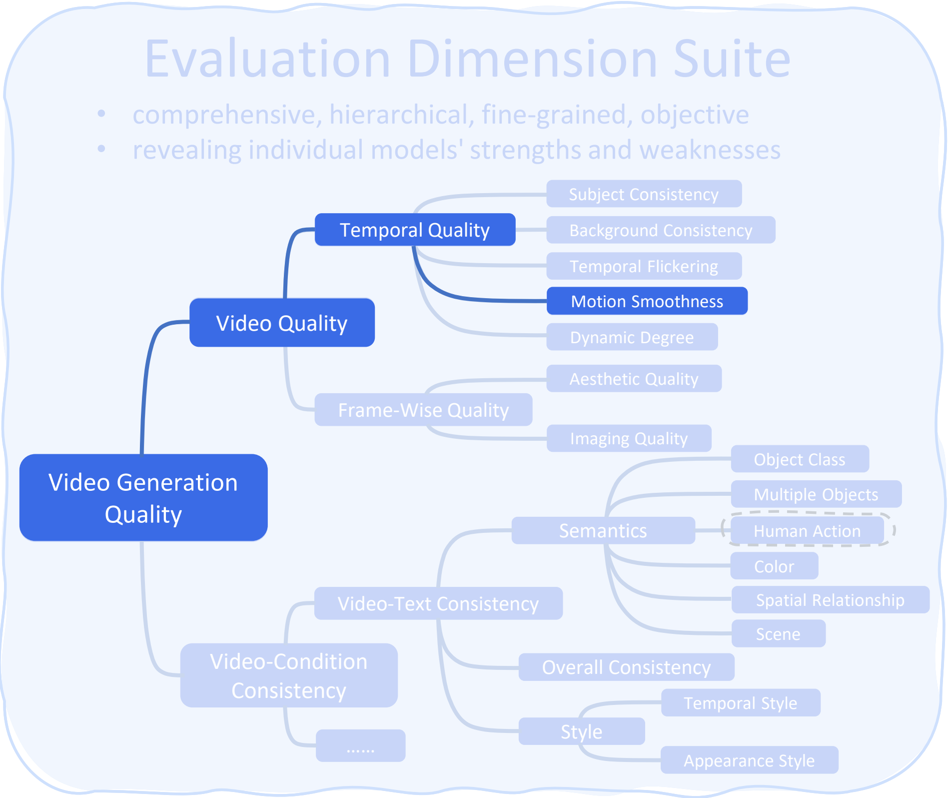


Dimension Suite

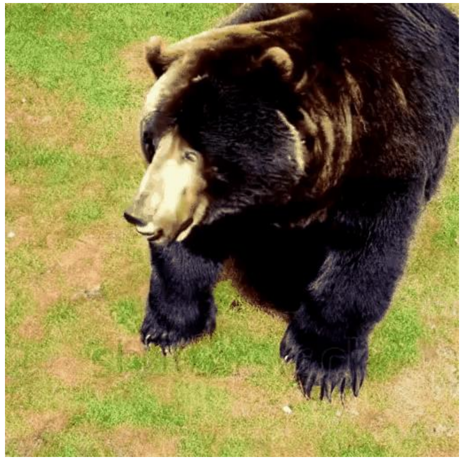


- 16 ability dimensions, hierarchical and disentangled
- Each dimension assesses one aspect of video generation quality
- Why Multiple Dimensions?
  - Reveal individual model’s strengths and weaknesses
  - Different people prioritize each ability dimension differently

Evaluation Dimension: *Motion Smoothness*



score 96.04% *(better)*

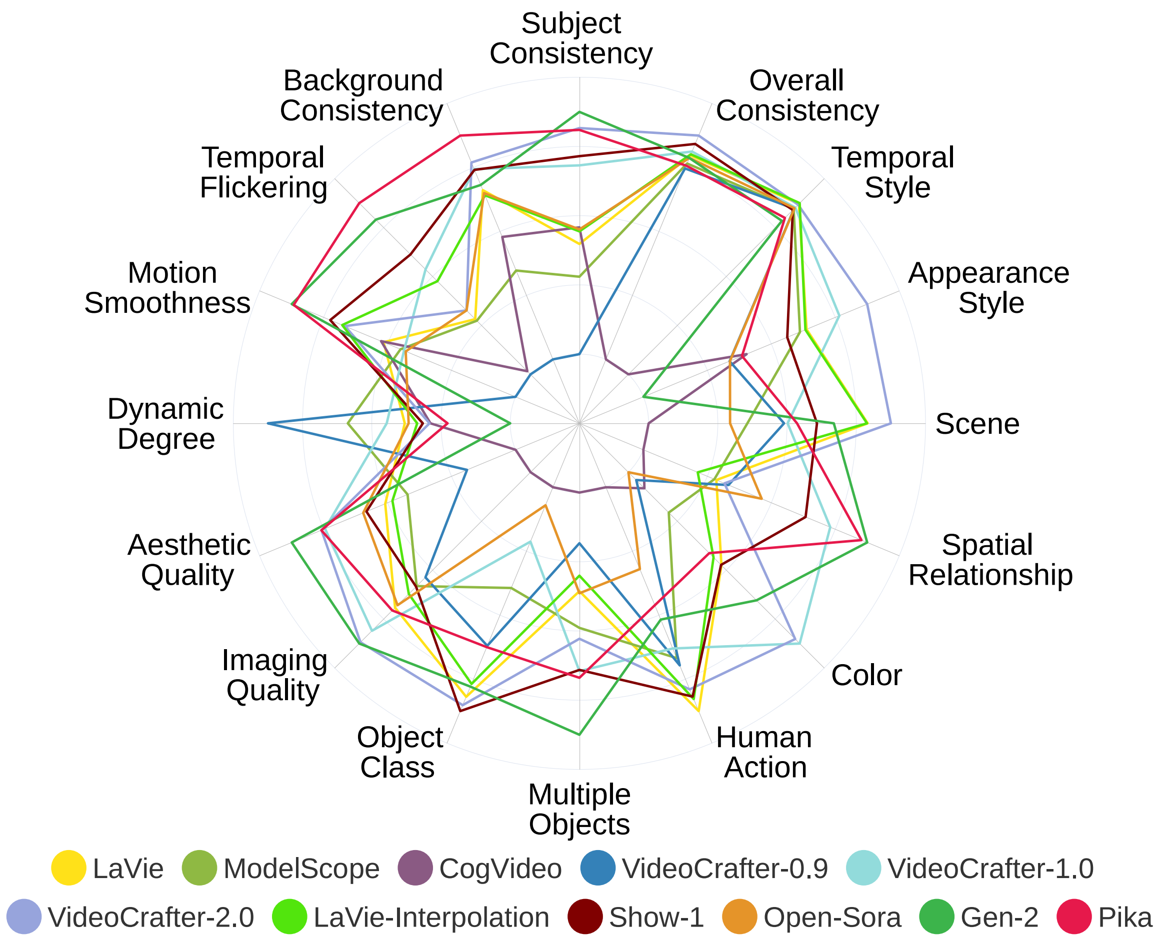


score 88.47%

whether the motion in the generated video is smooth

Evaluation Results

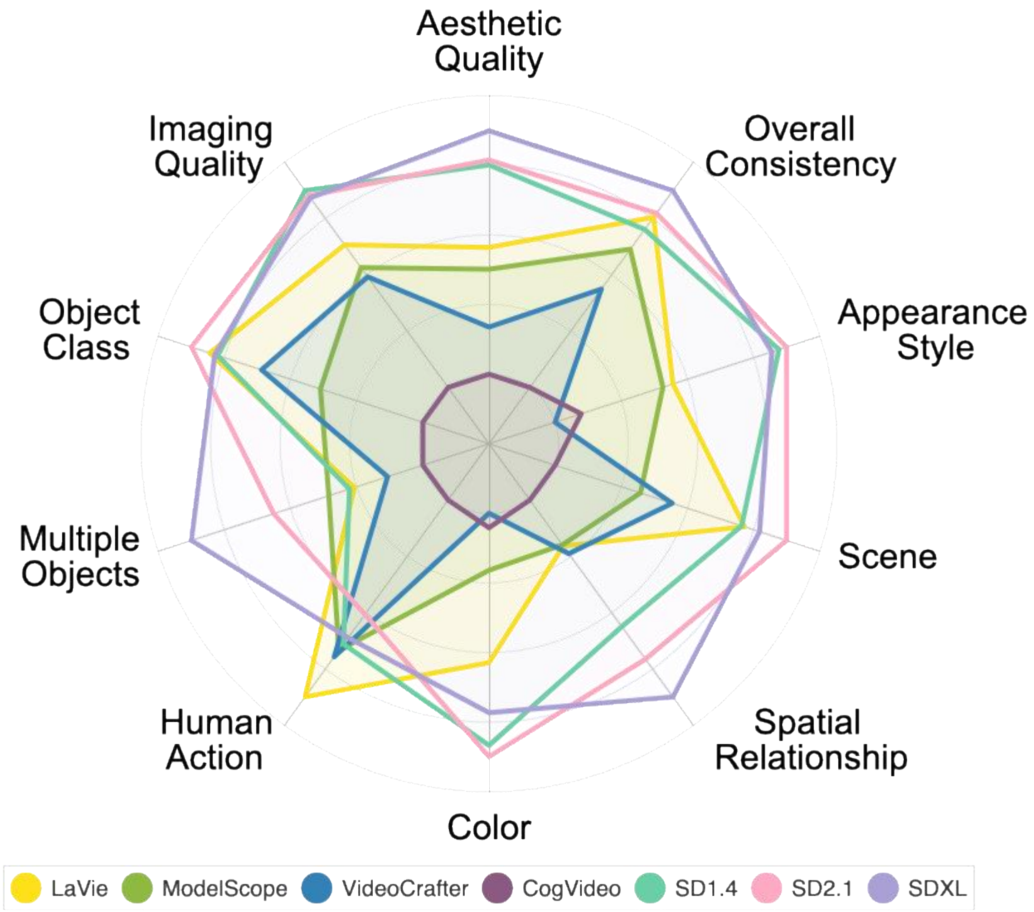
Video Generative Models



- **Trade-off across dimensions:**
  - e.g., temporal consistency vs. dynamic degree

Evaluation Results

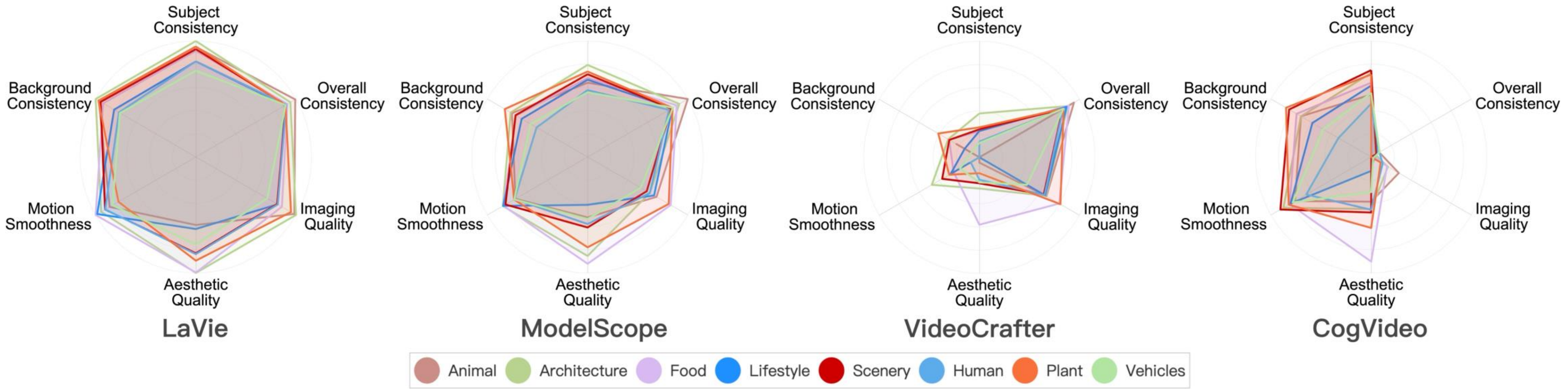
Video vs. Image Generative Models



- **Gap with T2I in compositionality**
  - e.g., multiple objects,
  - e.g., spatial relations

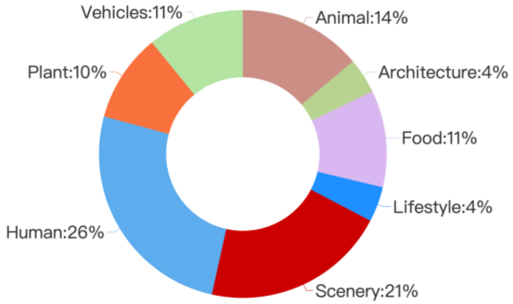
Evaluation Results

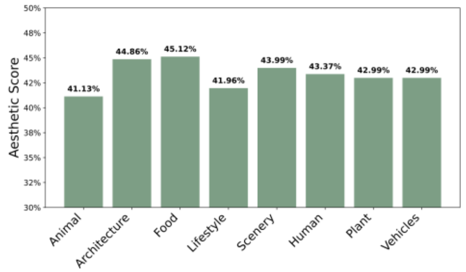
Content Categories

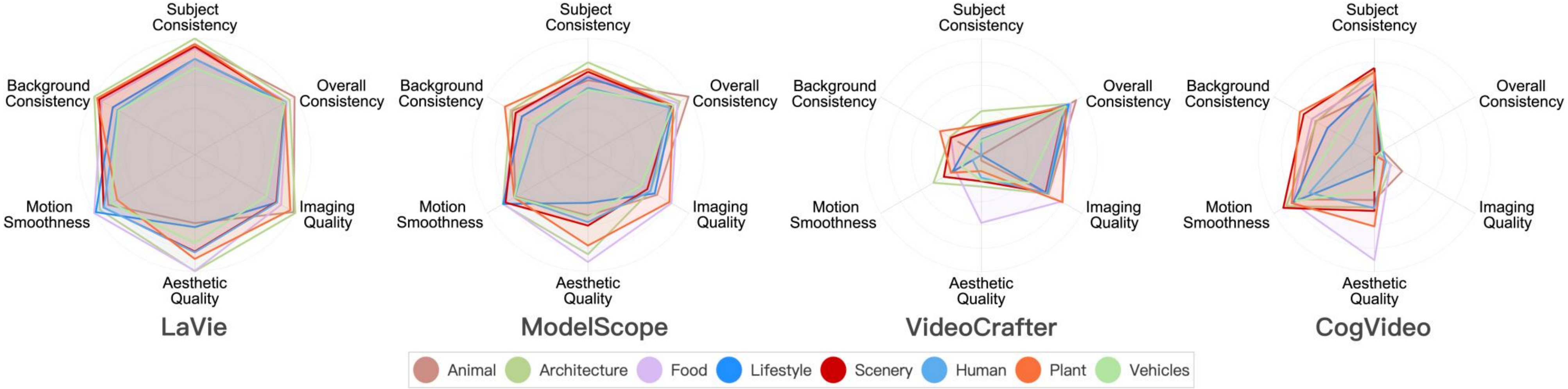


- **Uncovering hidden potential of models in specific content categories**
  - *e.g.,* CogVideo has strong aesthetics in Food category.
  - CogVideo's potential in aesthetics by improving such ability in other content types.
  - We recommend *evaluating video generation models not just based on ability dimensions but also considering specific content categories to uncover their hidden potential.*

Evaluation Results

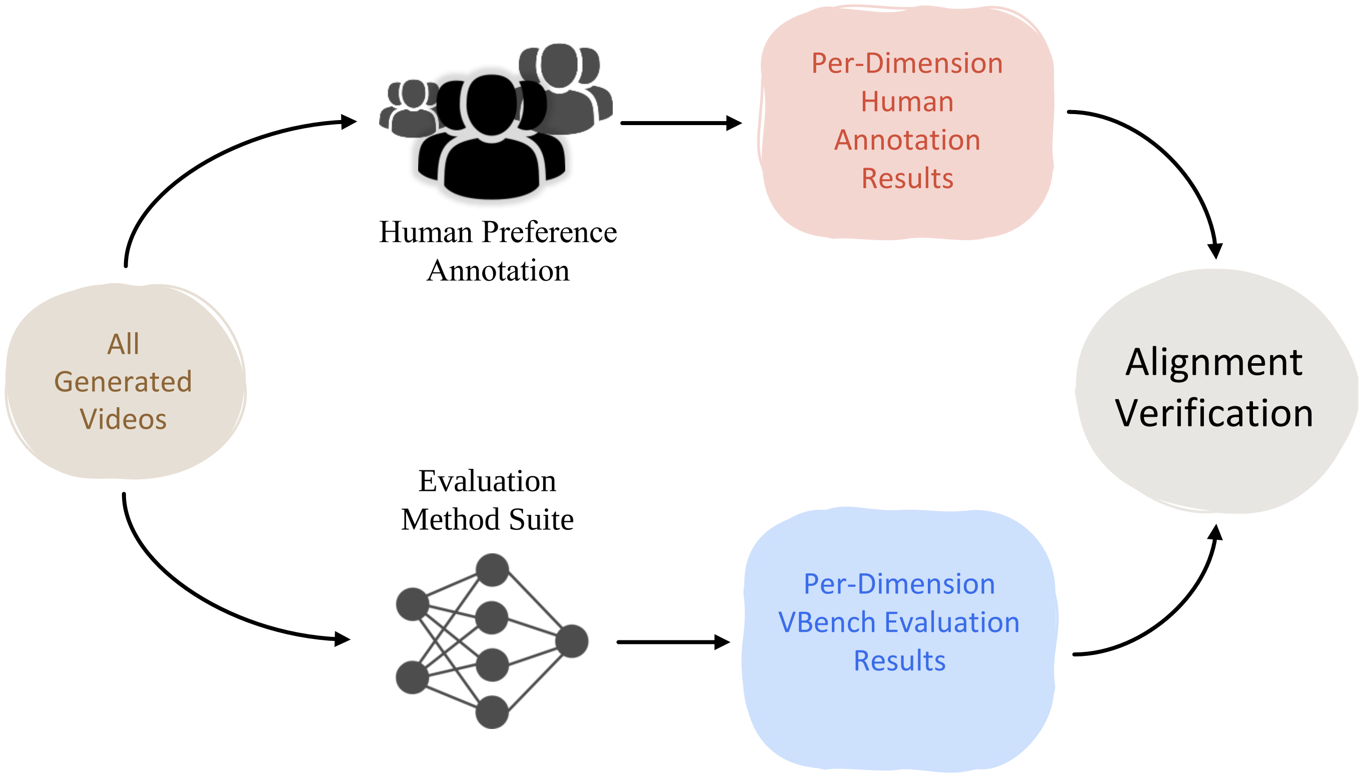


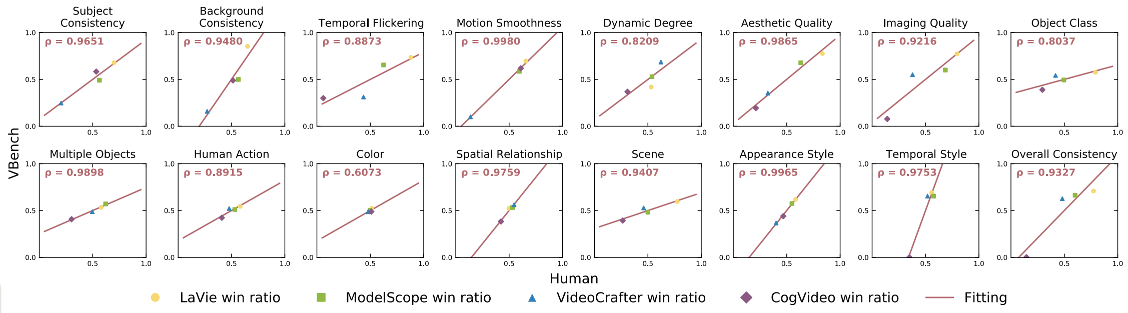




- ## **Data quality over data quantity**
  - *Despite constituting only 11% of the WebVid-10M dataset, the "Food" category consistently achieves the highest aesthetic quality scores. Further analysis reveals it maintains the highest aesthetic ratings within WebVid-10M. This underscores the importance of enhancing data quality rather than expanding data volume.*

Human Alignment of VBench





VBench evaluations across all dimensions closely match human perceptions.

VBench Leaderboard

VBench

Comprehensive Benchmark Suite for Video Generative Models

Select Quality Dimensions

Select Semantic Dimensions

Deselect All

Evaluation Dimension

- [x] subject consistency
- [x] background consistency
- [x] temporal flickering
- [x] motion smoothness
- [x] dynamic degree
- [x] aesthetic quality
- [x] imaging quality
- [x] object class
- [x] multiple objects
- [x] human action
- [x] color
- [x] spatial relationship
- [x] scene
- [x] appearance style
- [x] temporal style
- [x] overall consistency

Model Name (clickable)	Source	Total Score	Quality Score	Semantic Score	Selected Score	subject consistency	background consistency	t
T2V-Turbo (VC2)	T2V-Turbo Team	81.01%	82.57%	74.76%	81.01%	96.28%	97.02%	9
Gen-2 (2023-06)	VBench Team	80.58%	82.47%	73.03%	80.58%	97.61%	97.61%	9
VideoCrafter-2.0	VBench Team	80.44%	82.2%	73.42%	80.44%	96.85%	98.22%	9
Pika (2023-06)	VBench Team	80.4%	82.68%	71.26%	80.4%	96.76%	98.95%	9
AnimateDiff-V2	VBench Team	80.27%	82.9%	69.75%	80.27%	95.3%	97.68%	9
VideoCrafter-1.0	VBench Team	79.72%	81.59%	72.22%	79.72%	95.1%	98.04%	9
Show-1	VBench Team	78.93%	80.42%	72.98%	78.93%	95.53%	98.02%	9
Latte-1	VBench Team	77.29%	79.72%	67.58%	77.29%	88.88%	95.4%	9
LaVie-Interpolation	VBench Team	77.11%	79.06%	69.28%	77.11%	92.0%	97.33%	9
LaVie	VBench Team	77.08%	78.78%	70.31%	77.08%	91.41%	97.47%	9
Open-Sora	VBench Team	75.91%	78.82%	64.28%	75.91%	92.09%	97.39%	9
ModelScope	VBench Team	75.75%	78.05%	66.54%	75.75%	89.87%	95.29%	9
VideoCrafter-0.9	VBench Team	73.02%	74.91%	65.46%	73.02%	86.24%	92.88%	9
CogVideo	VBench Team	67.01%	72.06%	46.83%	67.01%	92.19%	96.2%	9

- 14 T2V models
- 12 I2V models
- ***Join our leaderboard!***

Fully Open-Source

- *Evaluation Method Suite (code)*
- *Prompt Suite (text prompts)*
- *Human Preference Annotations*
- *Generated Videos (mp4)*

  LaVie,ModelScope,CogVideo,Show-1, VideoCrafter-0.9/1/2, Pika,Gen-2, OpenSora (more to be added)

```
pip install vbench
```

GitHub

Serial Works in Progress

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

VBENCH-I2V

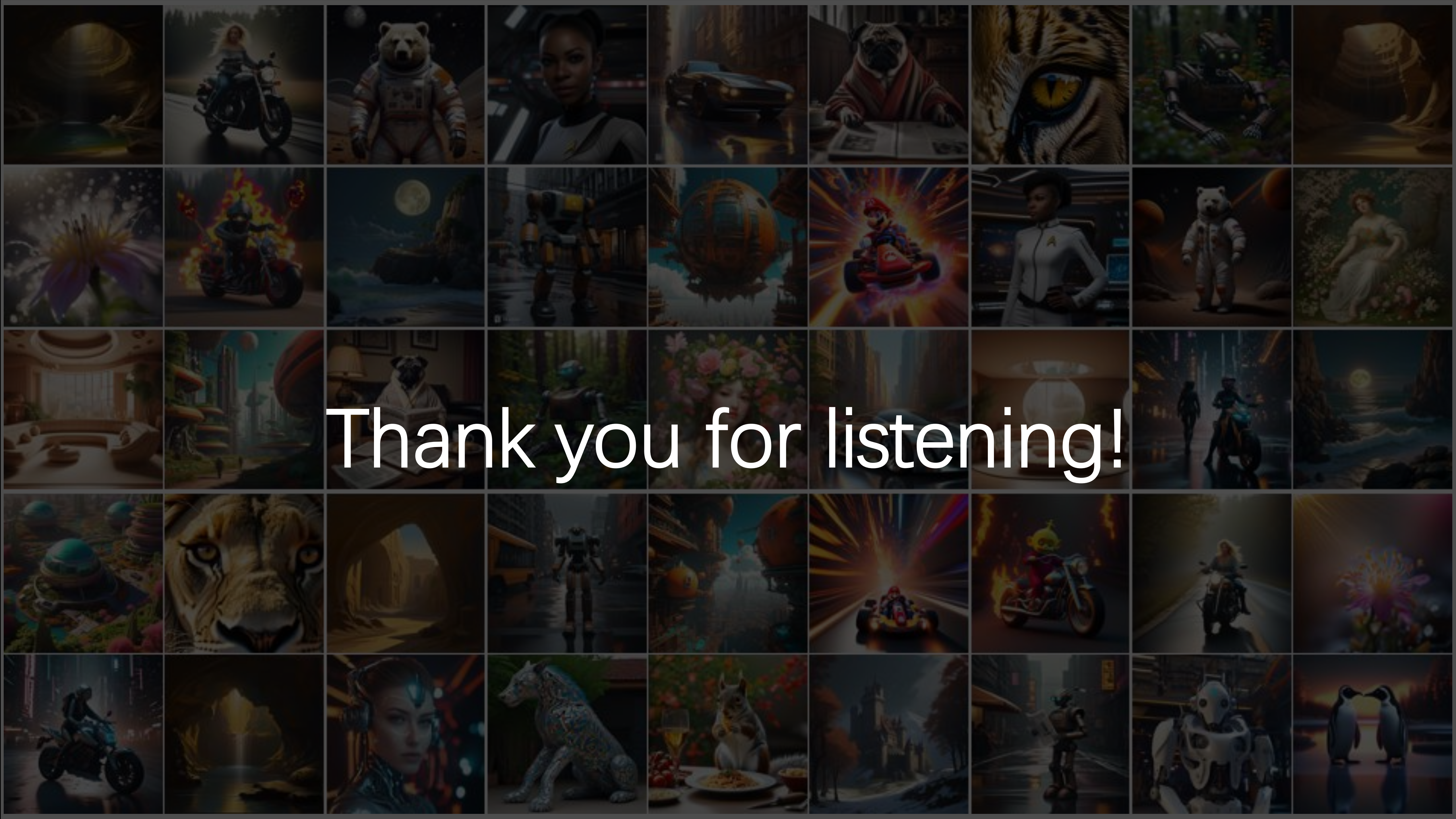
Image-to-Video (I2V): multi-ratio and multi-scale image benchmark, I2V evaluation dimensions

VBENCH-Long

for longer videos (e.g., 10 sec, 20 sec, 1 min)

VBENCH-Trustworthiness

non-technical aspects of video generation model: culture, bias, safety


Thank you for listening!