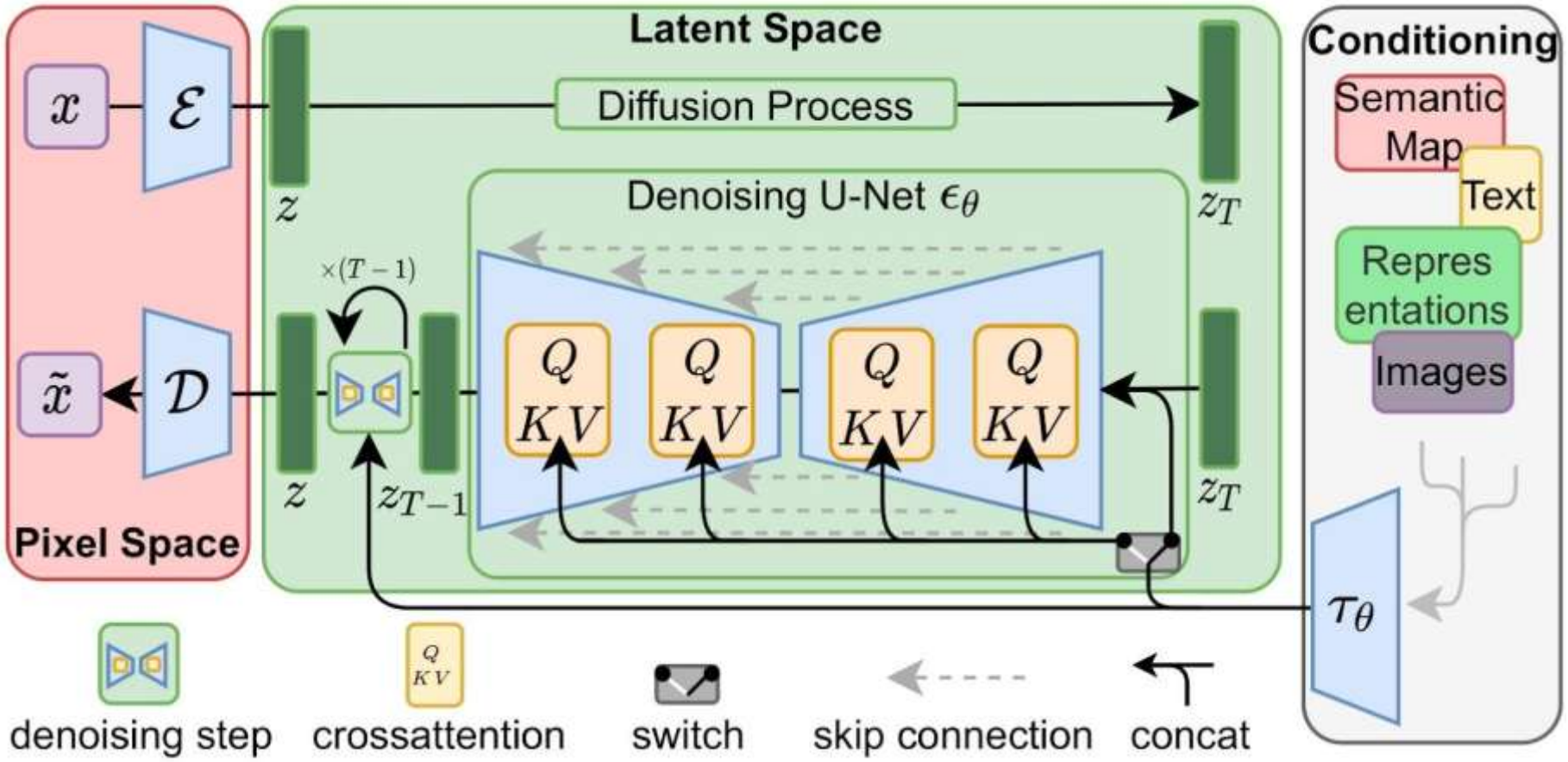
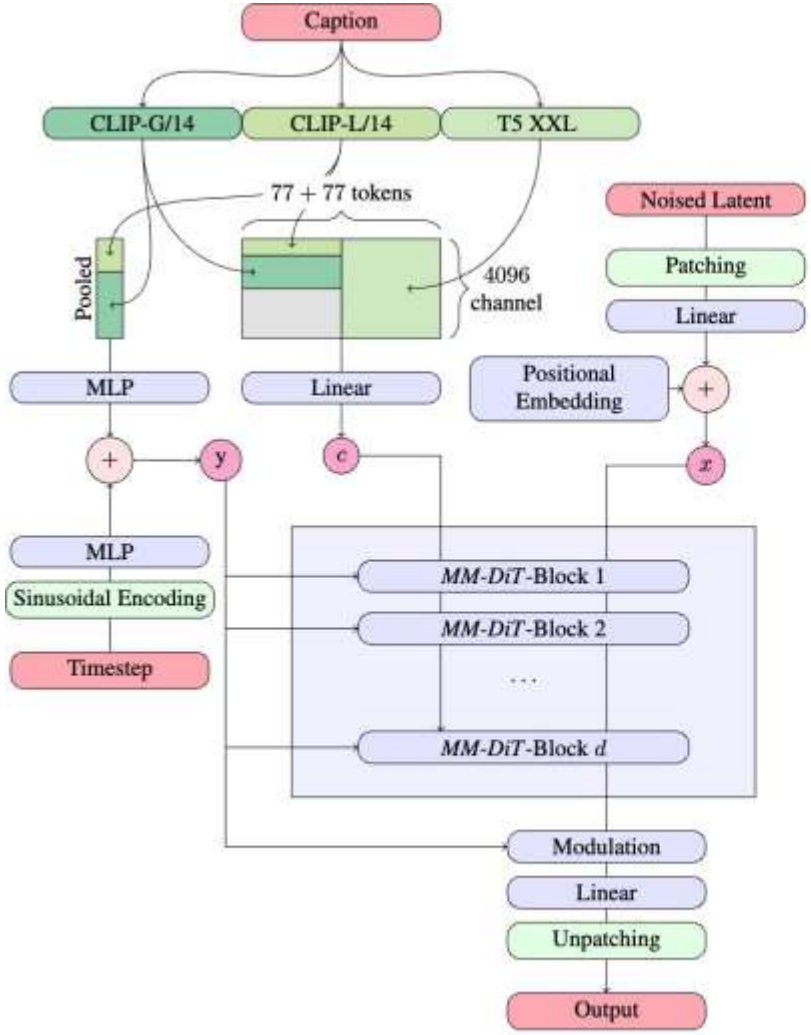
Recap.

Attempt 1: Cross Attention

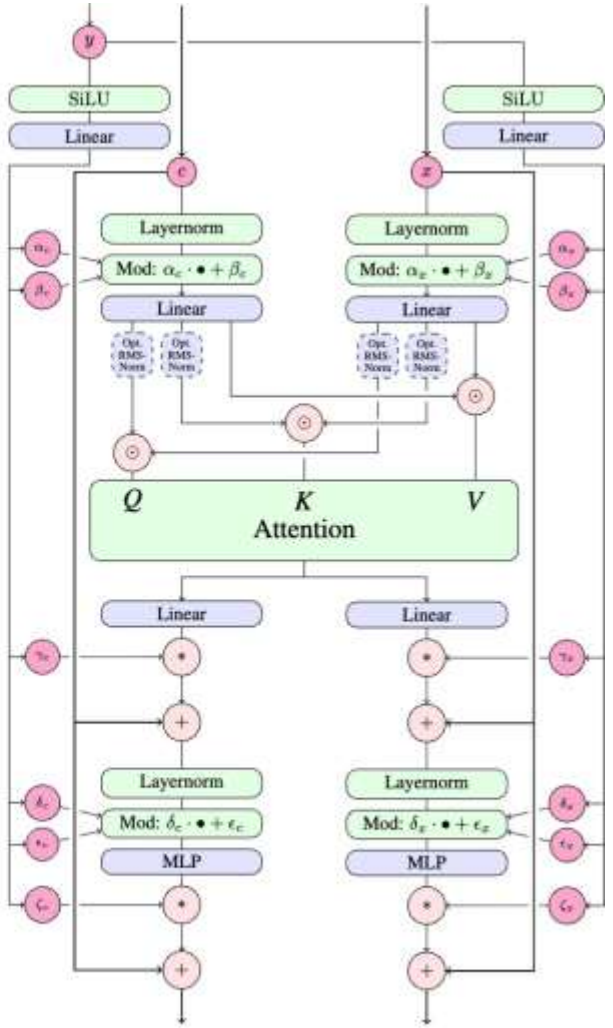


Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". CVPR 2022. https://arxiv.org/pdf/2112.10752

Attempt 2: Double Stream Multimodal-DiT



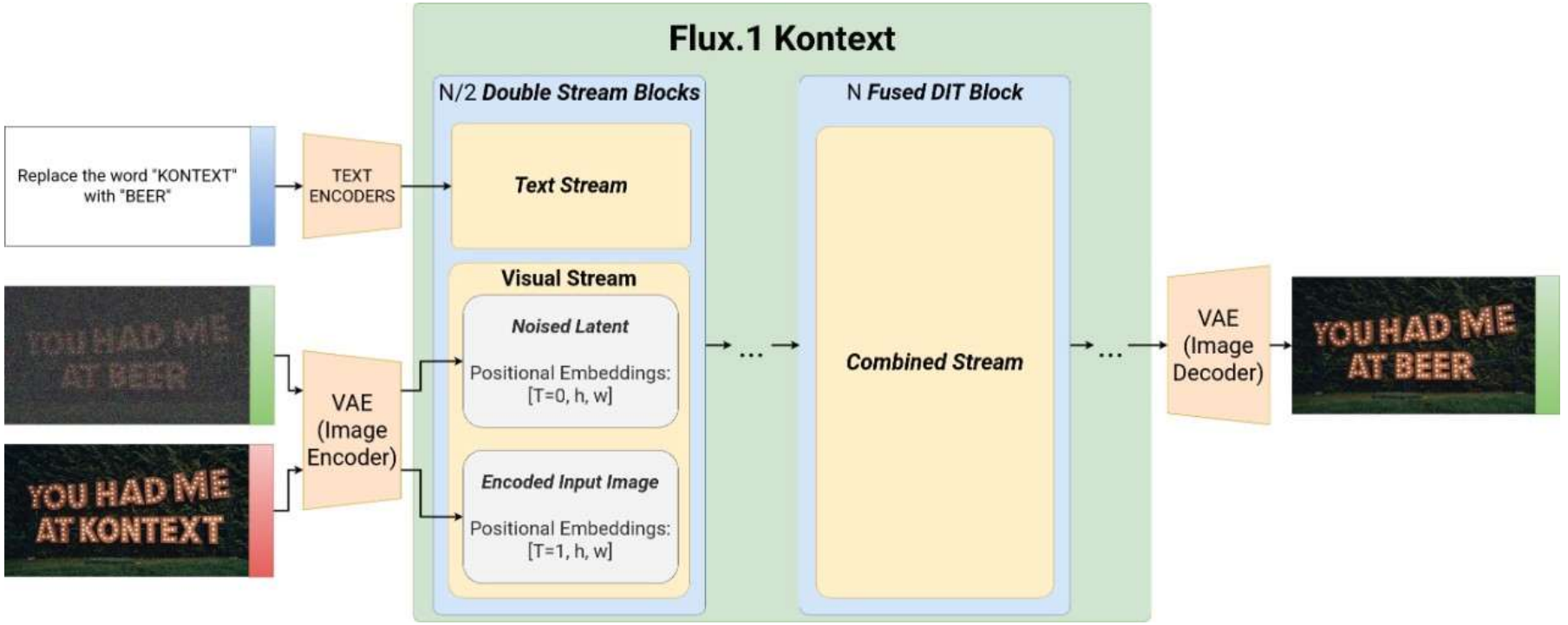
(a) Overview of all components.



(b) One *MM-DiT* block

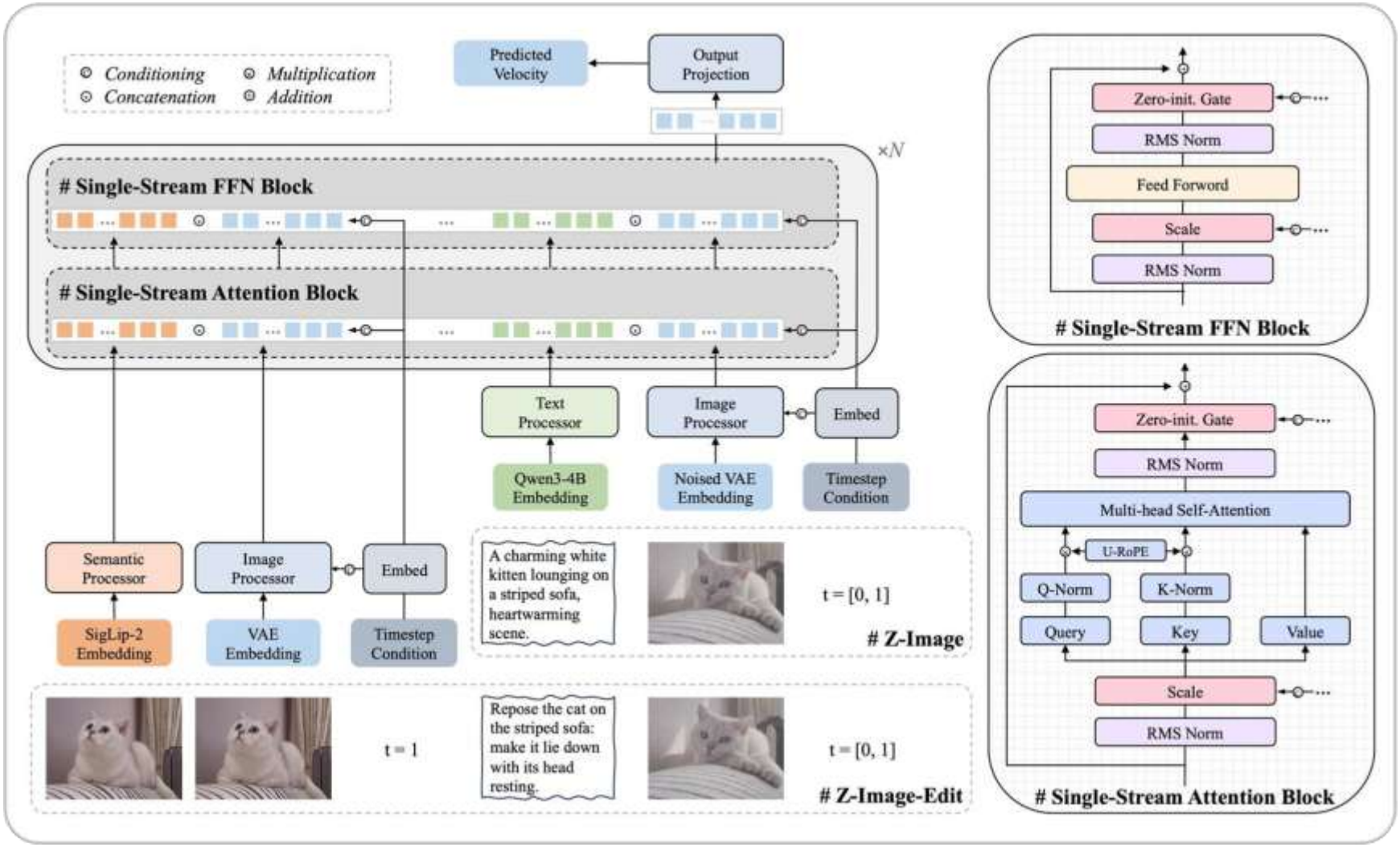
Esser et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". https://arxiv.org/pdf/2403.03206

Attempt 2: Double stream -> merged stream MM-DiT



Black Forest Labs. "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space". https://arxiv.org/pdf/2506.15742

Attempt 3: Single stream MM-DiT



Z-Image Team, Alibaba Group. "Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer". https://arxiv.org/pdf/2511.22699

Attempt 3.5: “Native multimodal model”

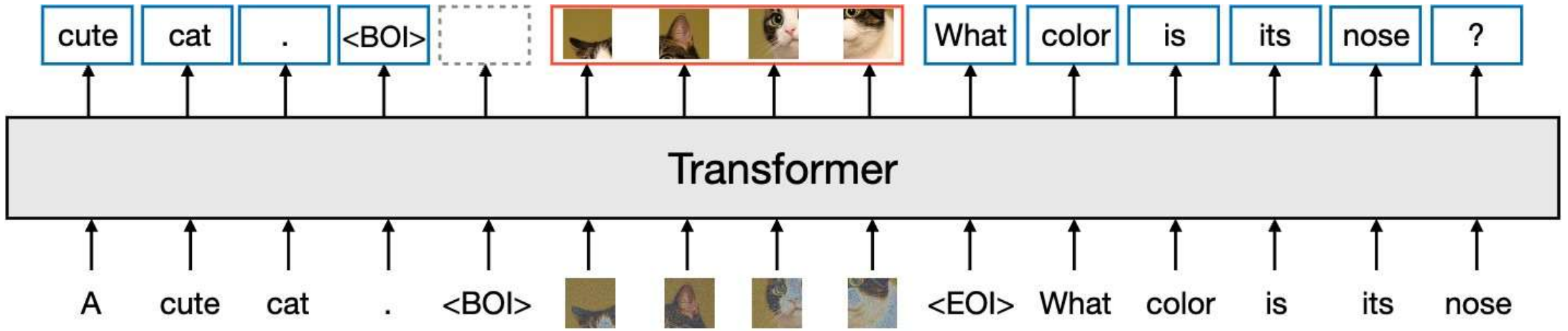
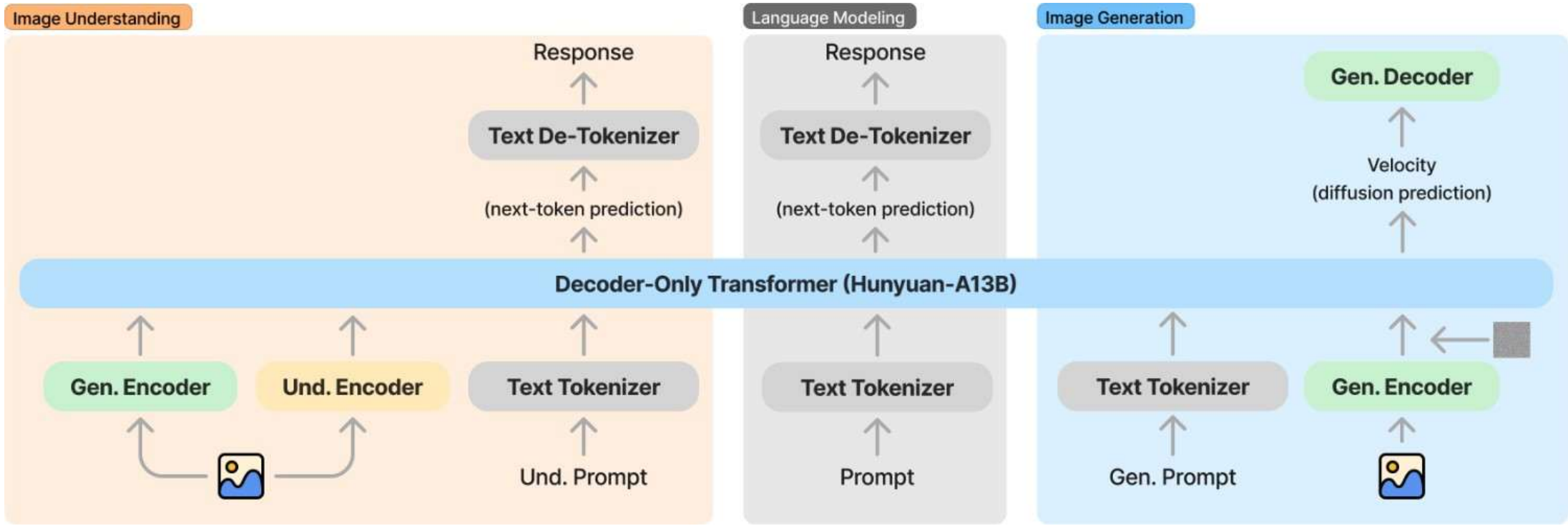


Figure 1: A high-level illustration of Transfusion. A single transformer perceives, processes, and produces data of every modality. Discrete (text) tokens are processed autoregressively and trained on the next token prediction objective. Continuous (image) vectors are processed together in parallel and trained on the diffusion objective. Marker BOI and EOI tokens separate the modalities.

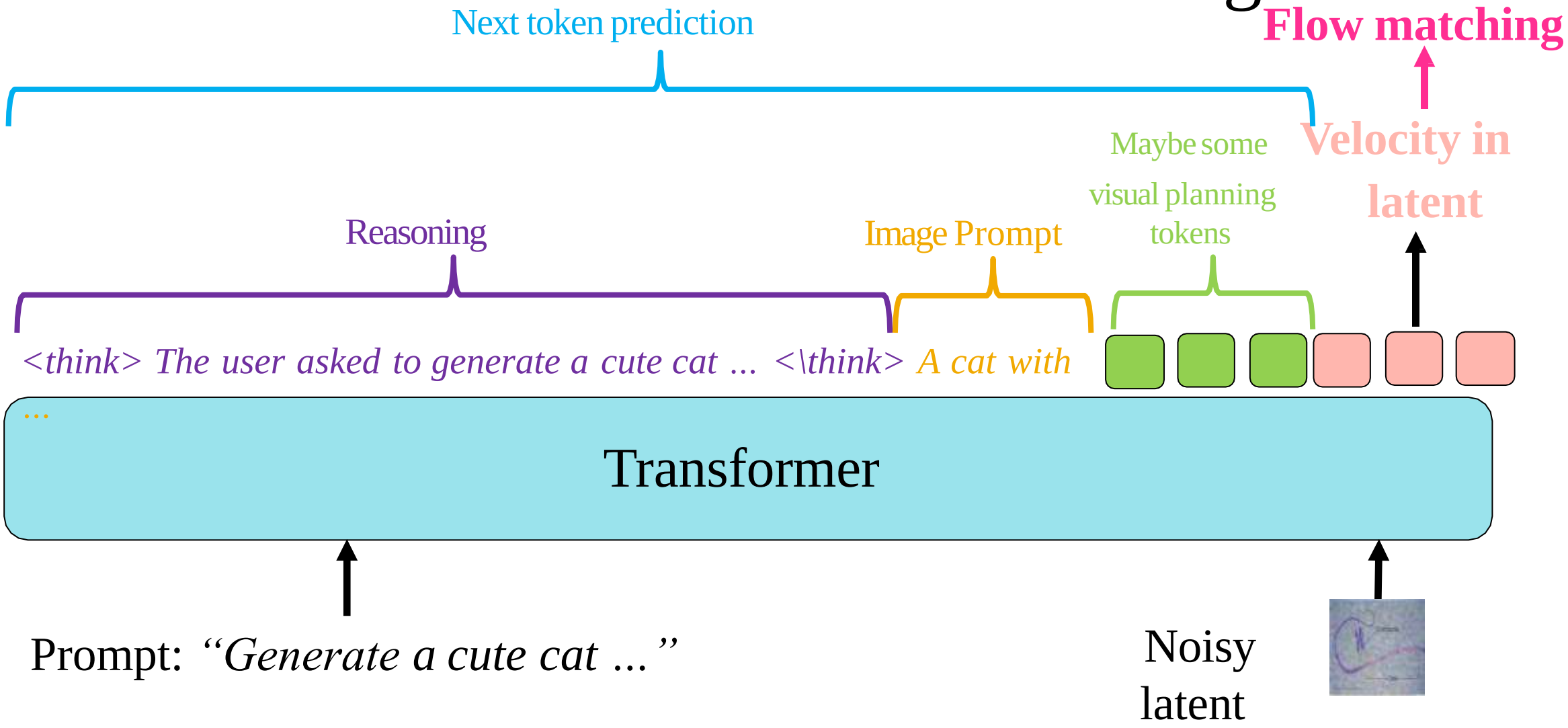
Zhou et al. “Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model”. https://arxiv.org/pdf/2408.11039

Attempt 3.5: “Native multimodal model”

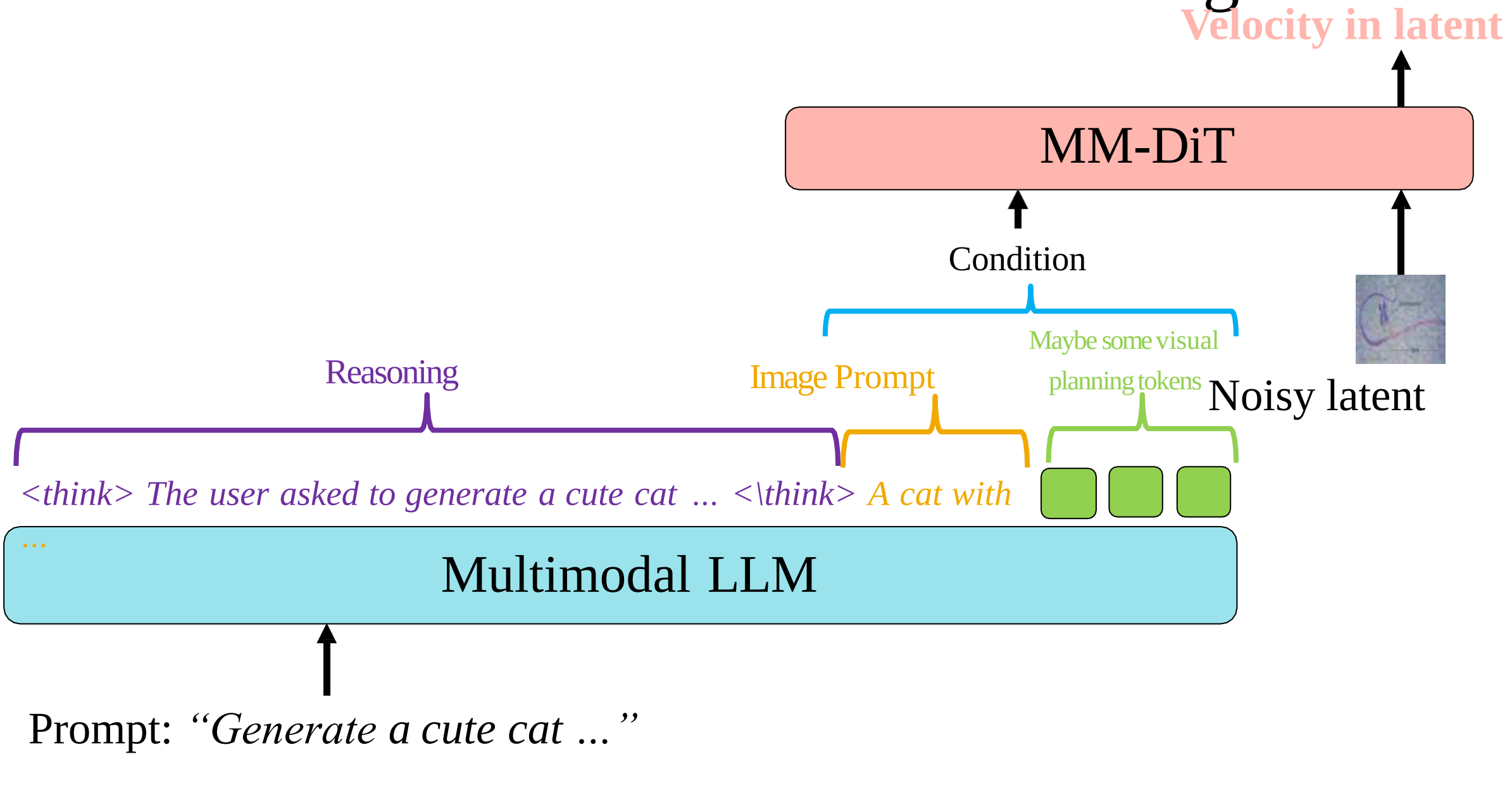


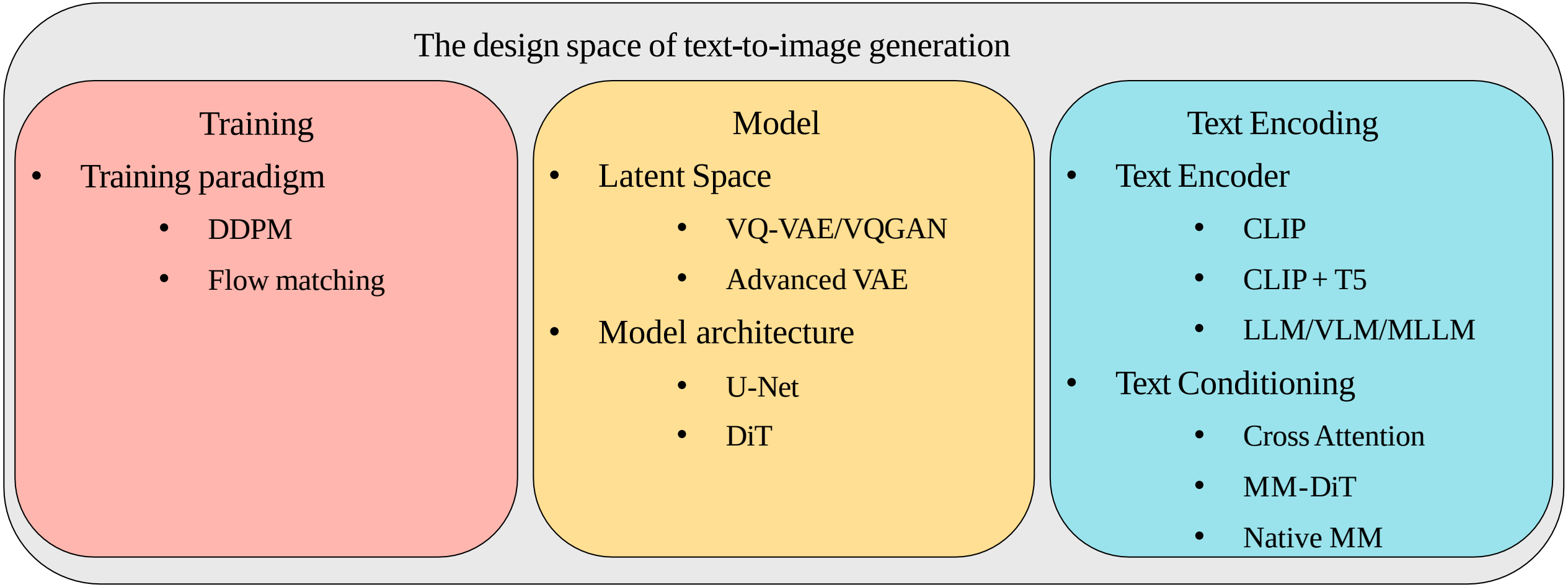
Tencent Hunyuan Foundation Model Team. “HunyuanImage 3.0 Technical Report”. https://arxiv.org/pdf/2408.11039

Guess on how Nano Banana & GPT-4o Image work



Guess on how Nano Banana & GPT-4o Image work




The design space of text-to-image generation
Training
Training paradigm
DDPM
Flow matching
Model
Latent Space
VQ-VAE/VQGAN
Advanced VAE
Model architecture
U-Net
DiT
Text Encoding
Text Encoder
CLIP
CLIP + T5
LLM/VLM/MLLM
Text Conditioning
Cross Attention
MM-DiT
Native MM

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

This is Stable Diffusion 1 & 2!

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

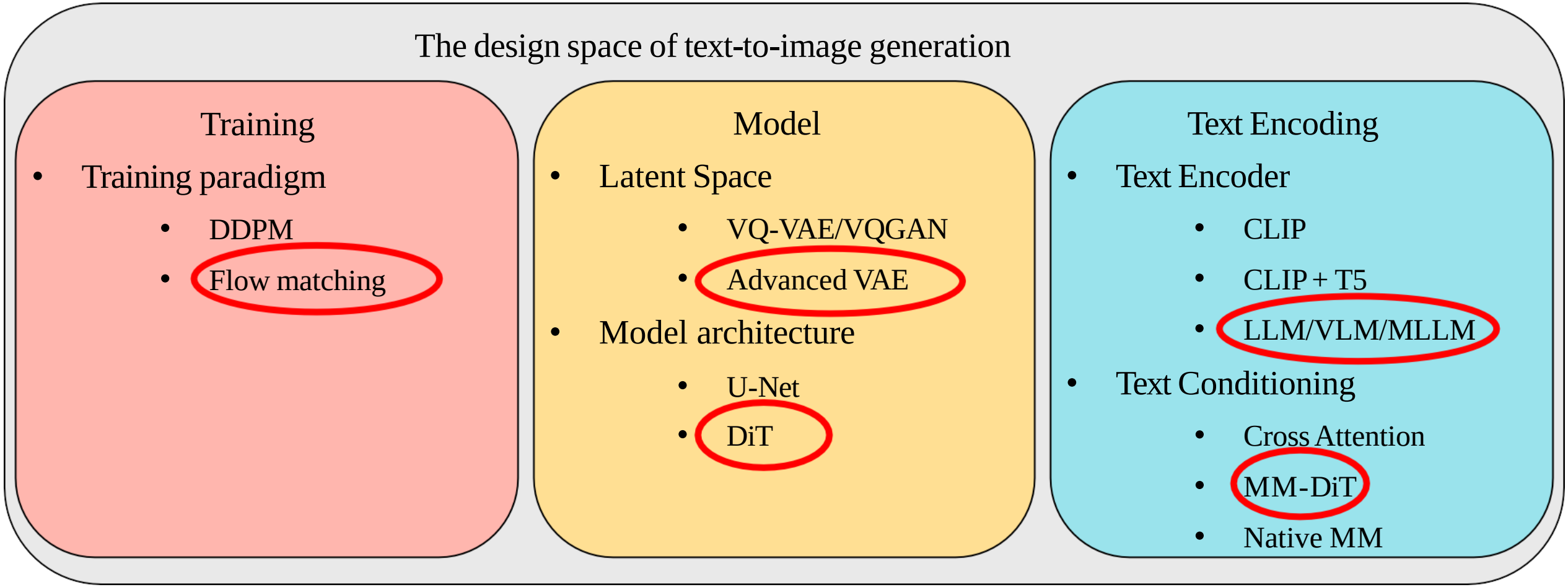
Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

This is Stable Diffusion 3 & Flux 1!



This is Flux 2, Z-Image, Qwen-Image, etc!

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

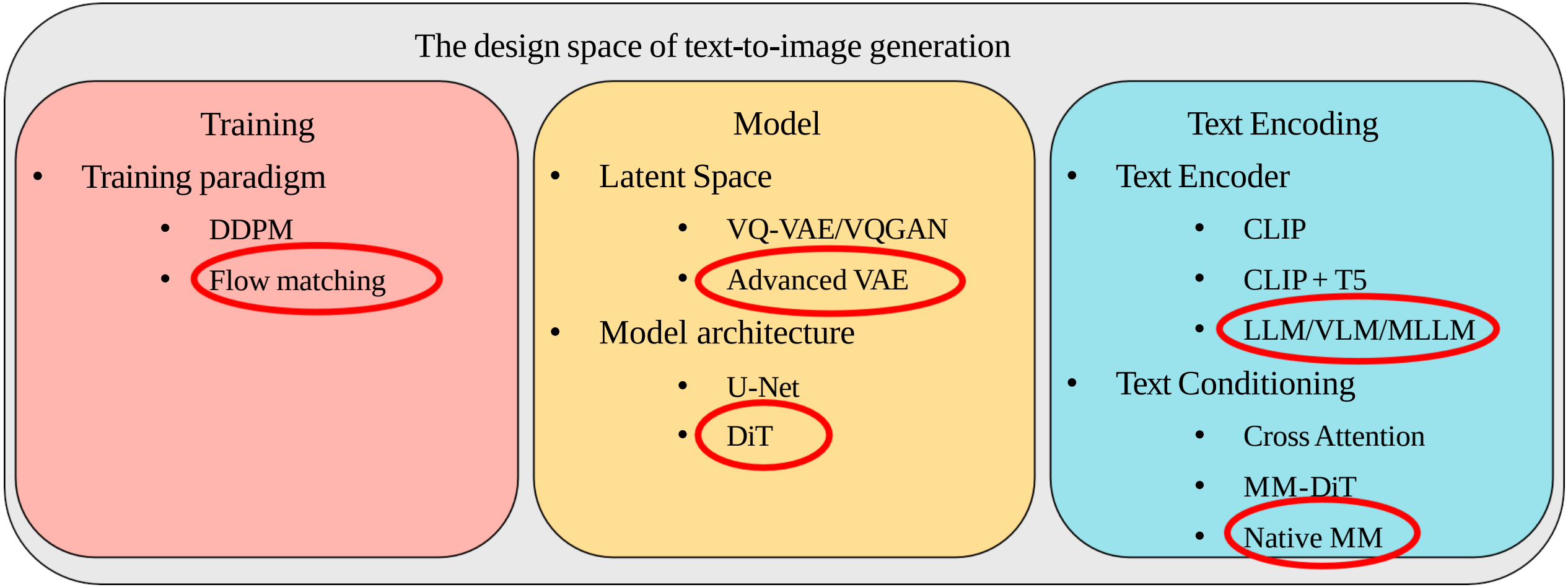
Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

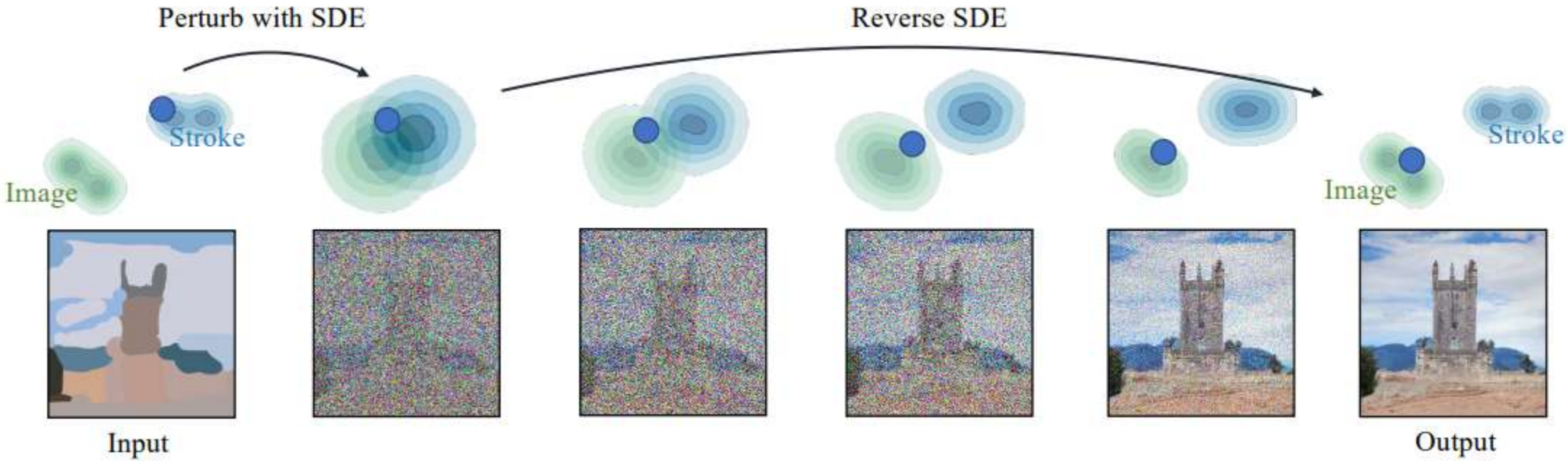
- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

This is Transfusion, Hunyuan 3.0, etc!



(Probably also Nano Banana & GPT-4o Image)

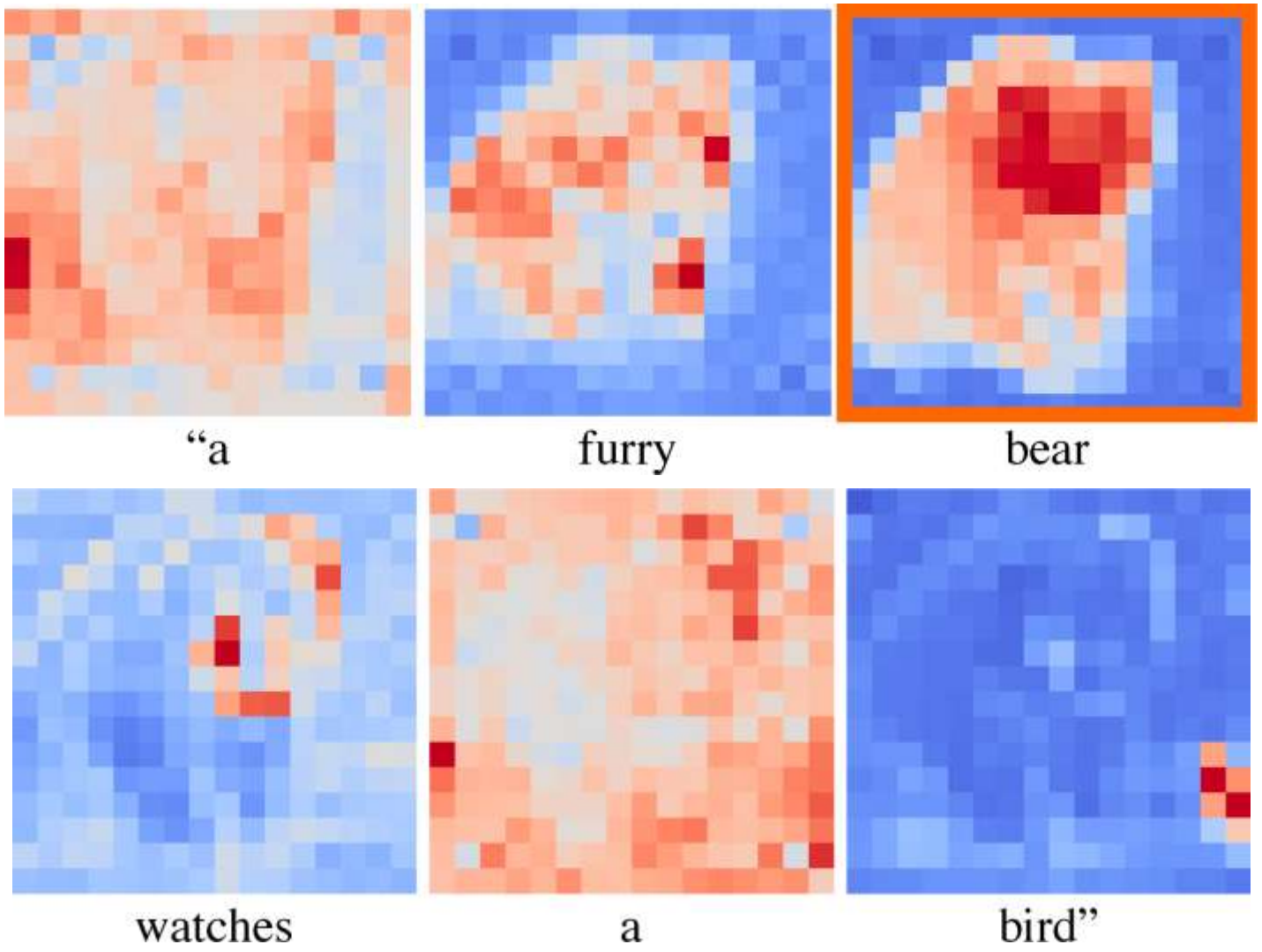
SDEdit: Guided Image Synthesis with Diffusion



[Meng et al., ICLR 2022]

Attention visualization

“a furry bear watches a bird”



Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

Prompt-to-Prompt

"Photo of a cat riding on a bicycle."

source image

cat → dog

cat → chicken

cat → squirrel

cat → elephant

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Final Method

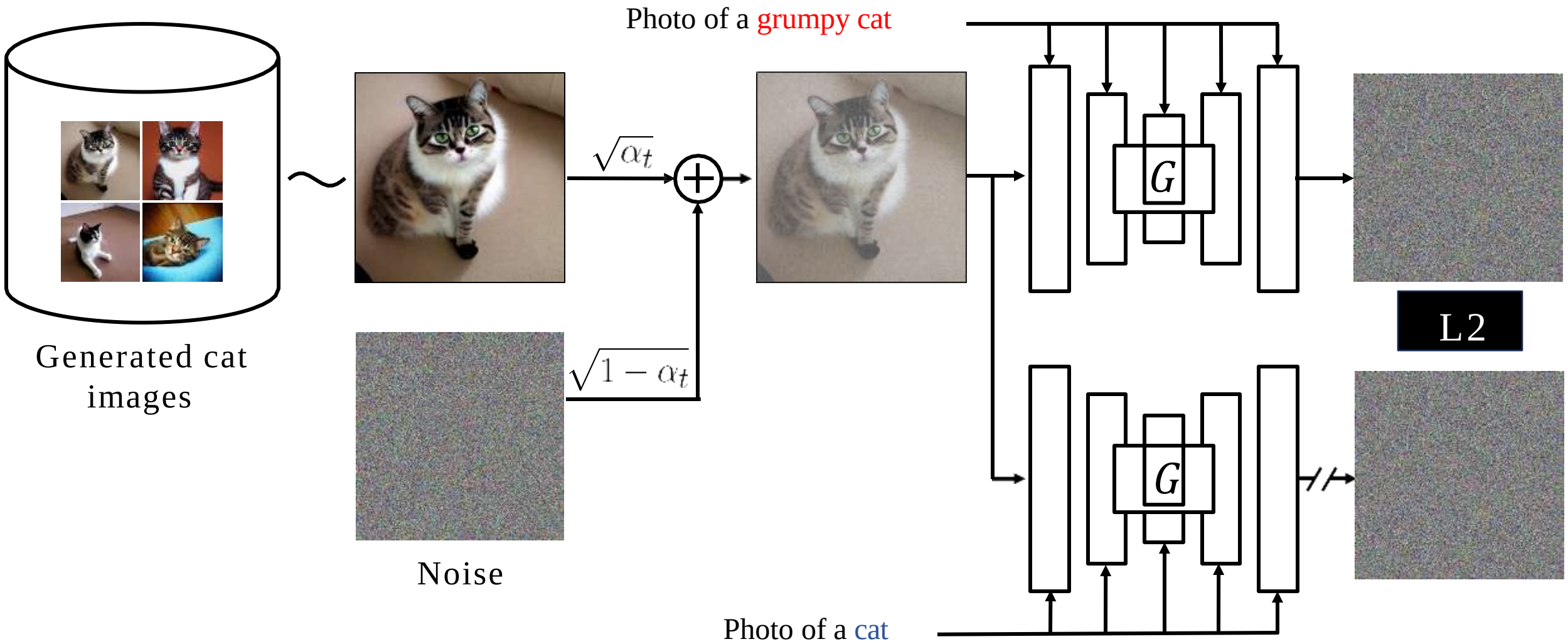


Photo of a grumpy cat

Target concept

Photo of a british shorthair cat

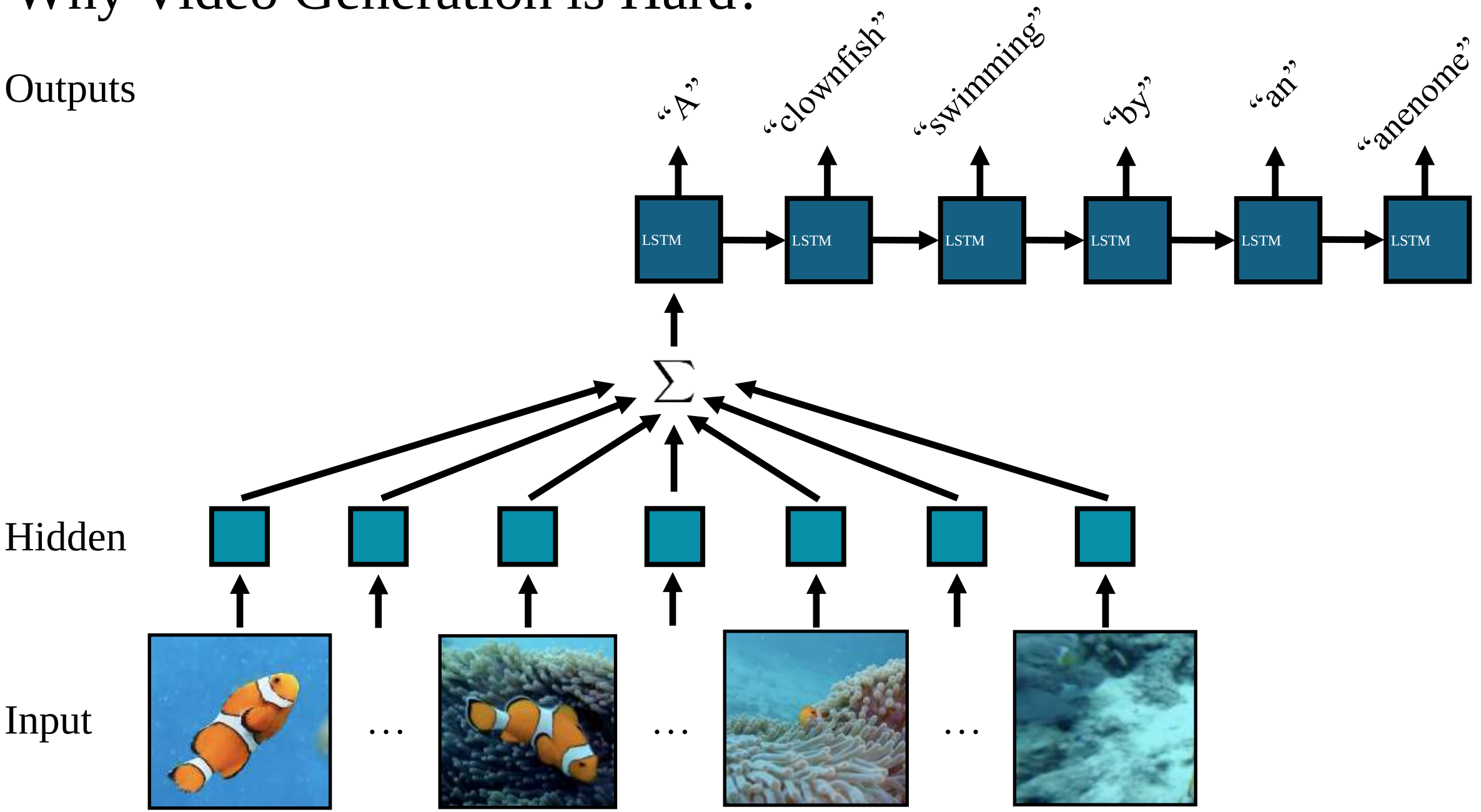
Nearby concept

Video Generation

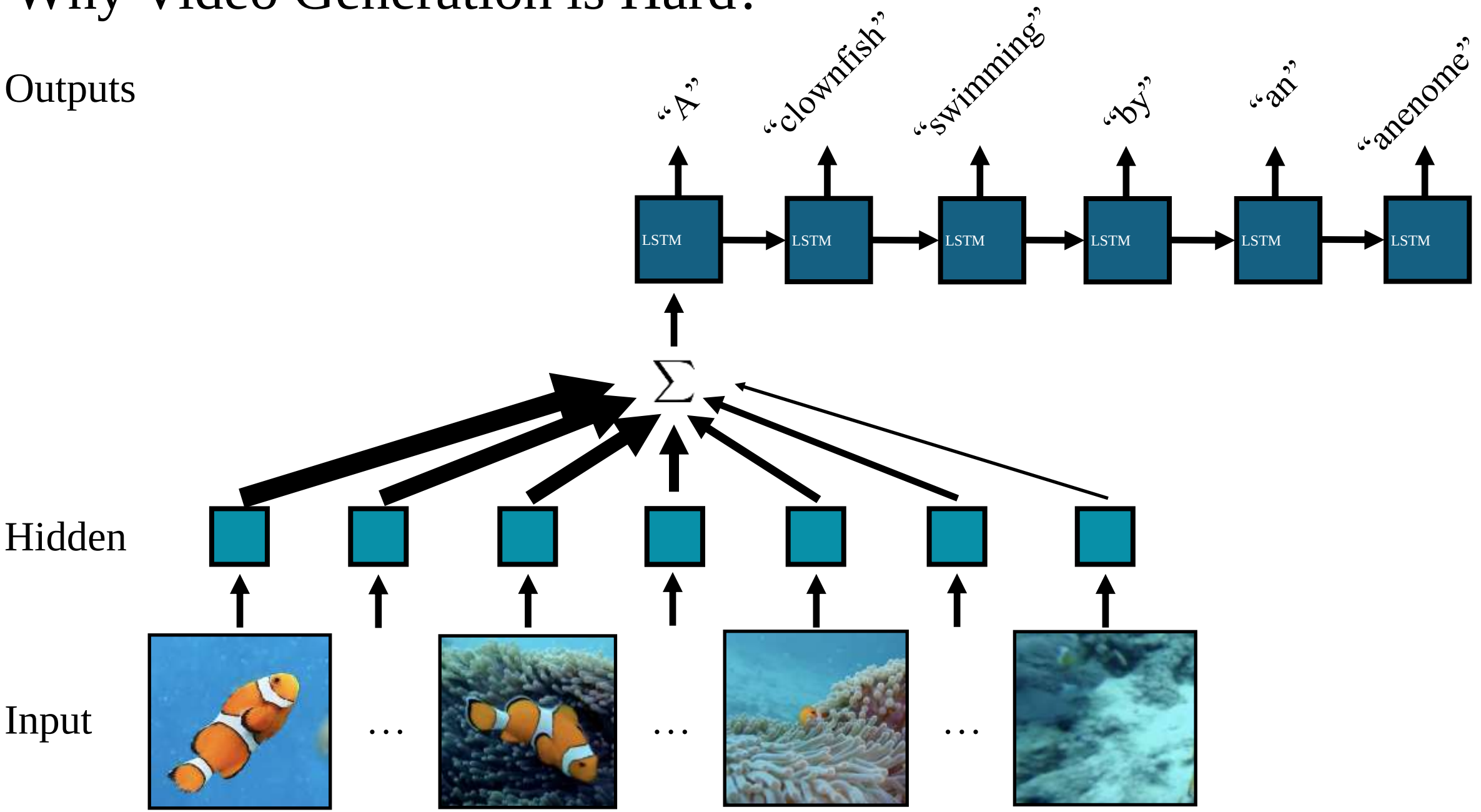
Sora

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

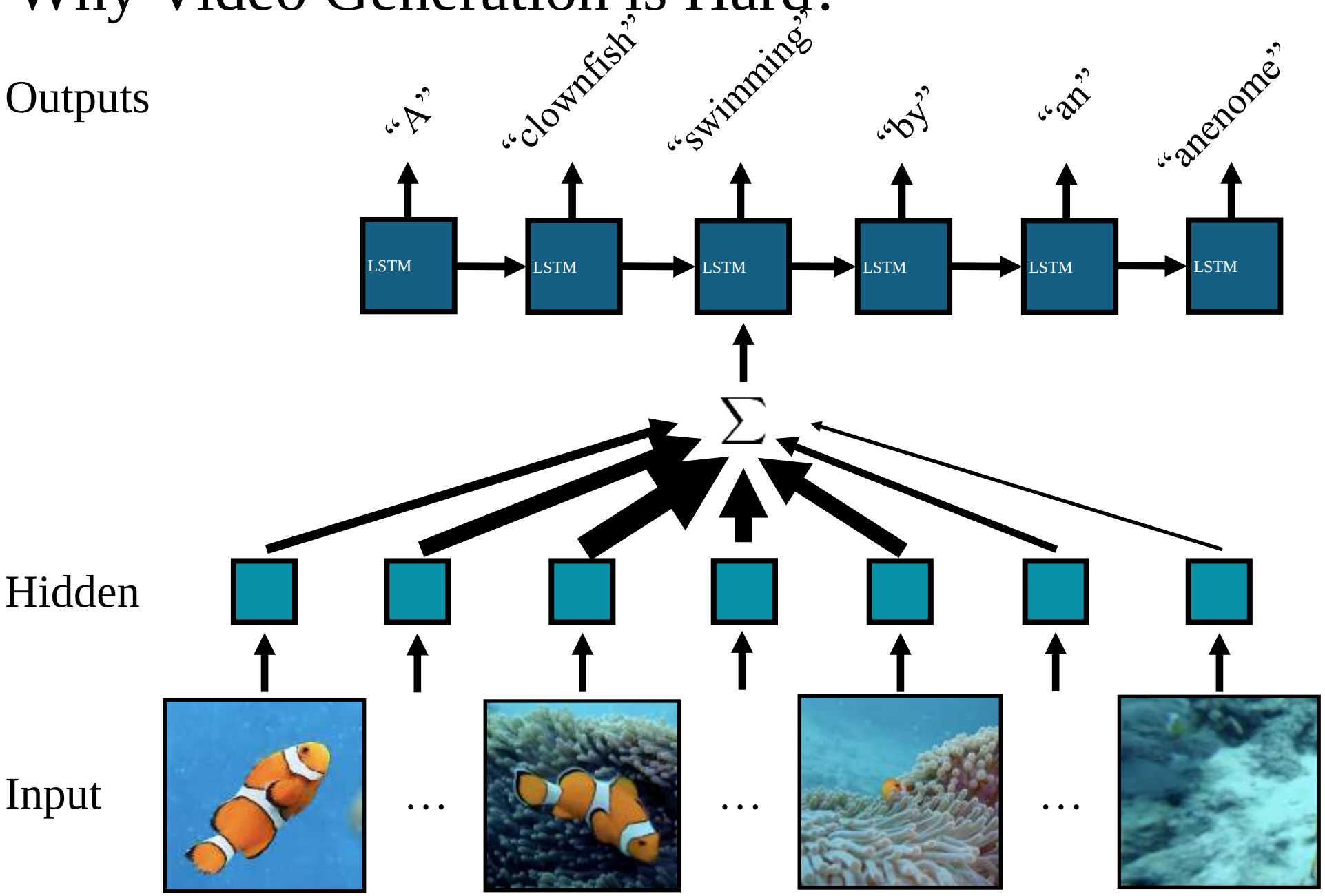
Why Video Generation is Hard?



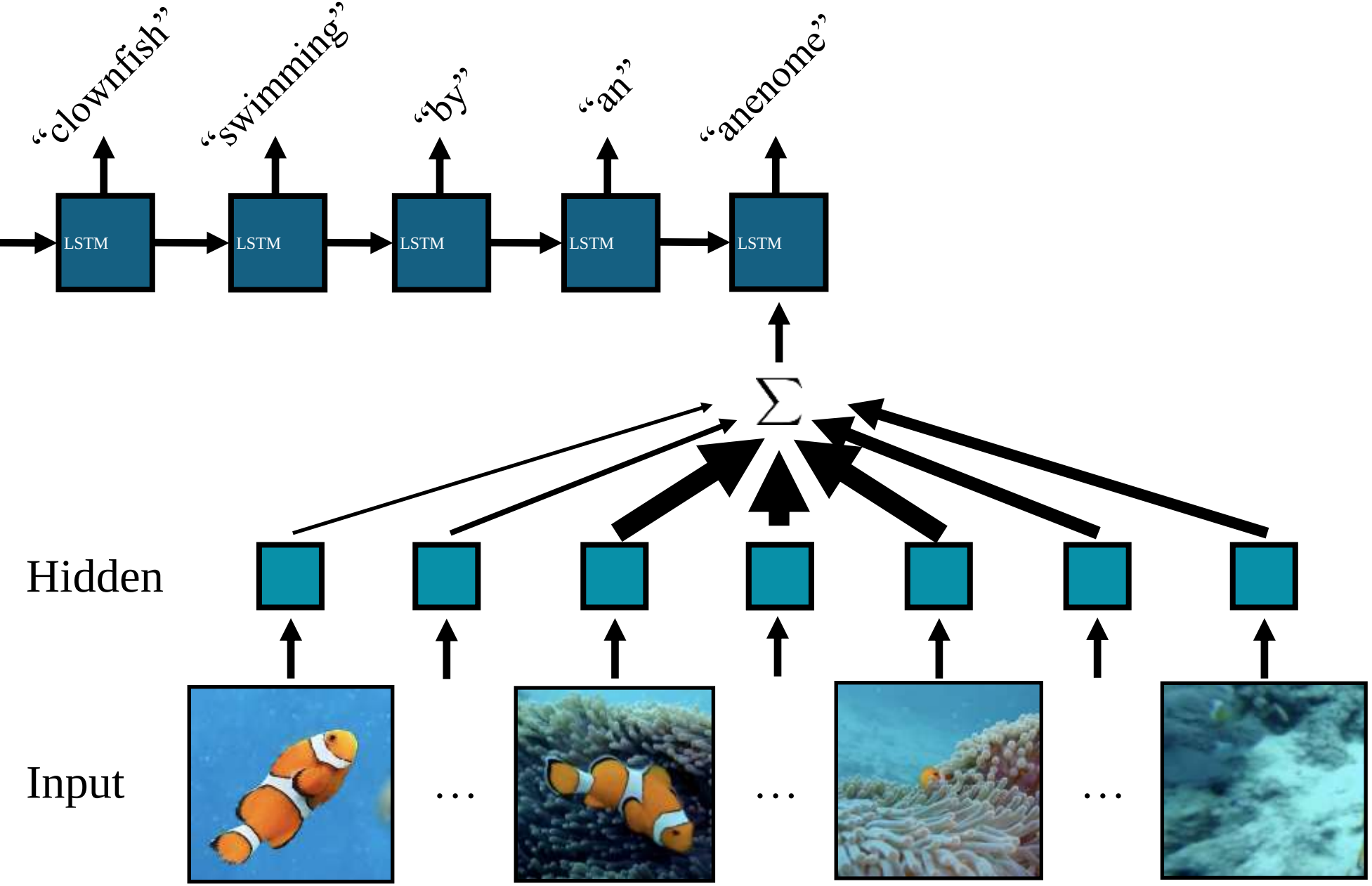
Why Video Generation is Hard?



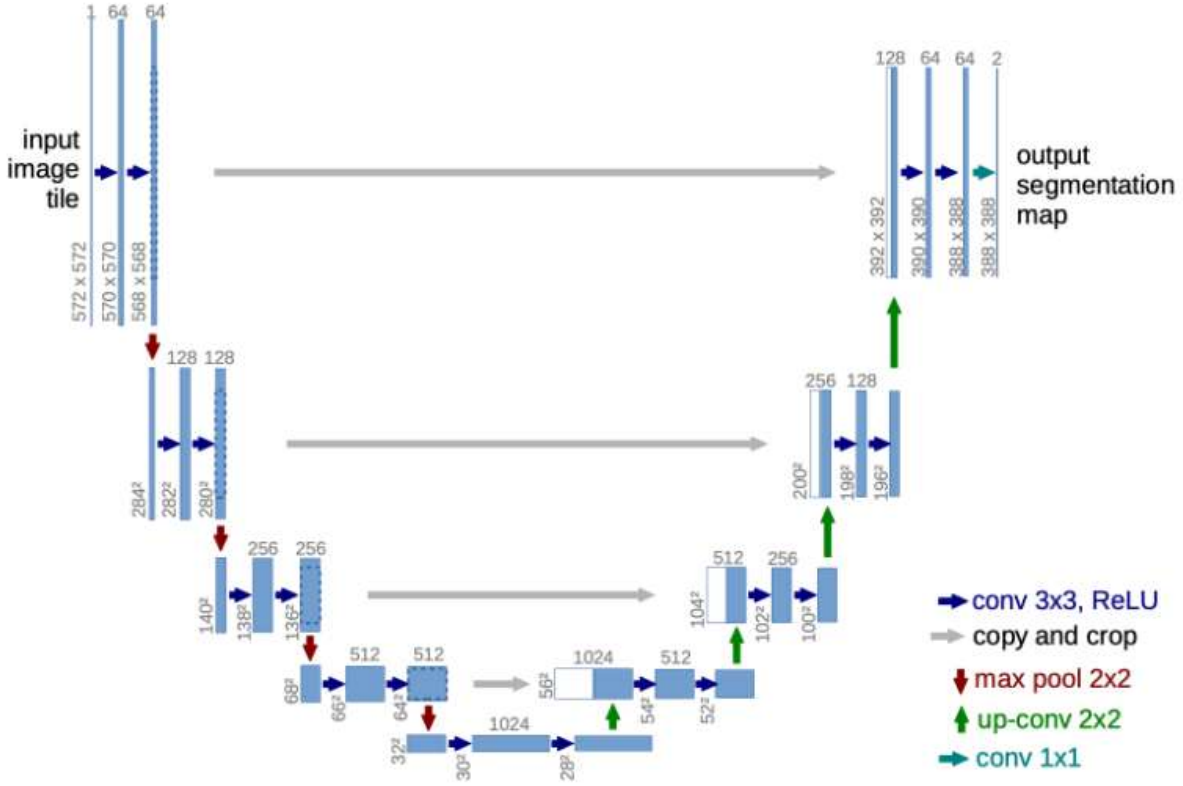
Why Video Generation is Hard?



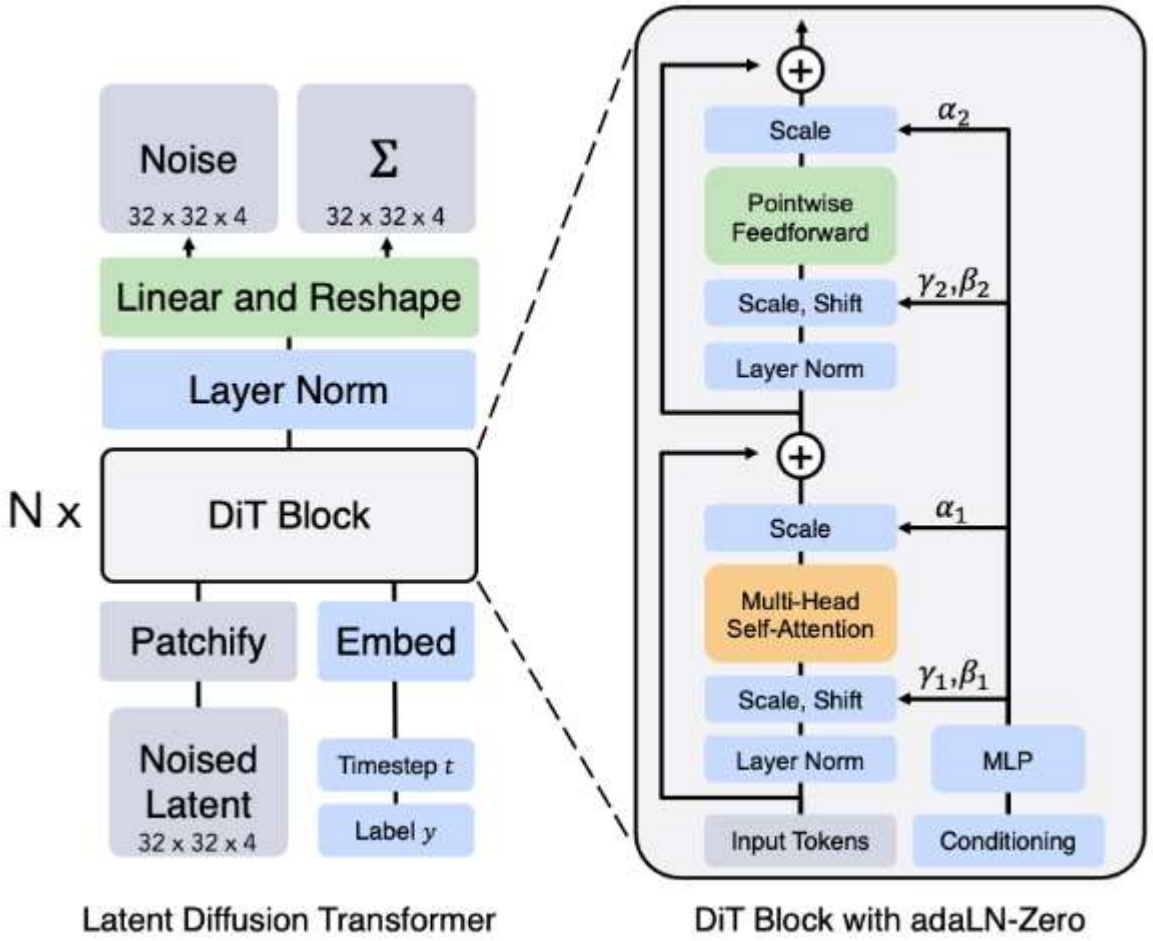
Why Video Generation is Hard?



Model Architecture

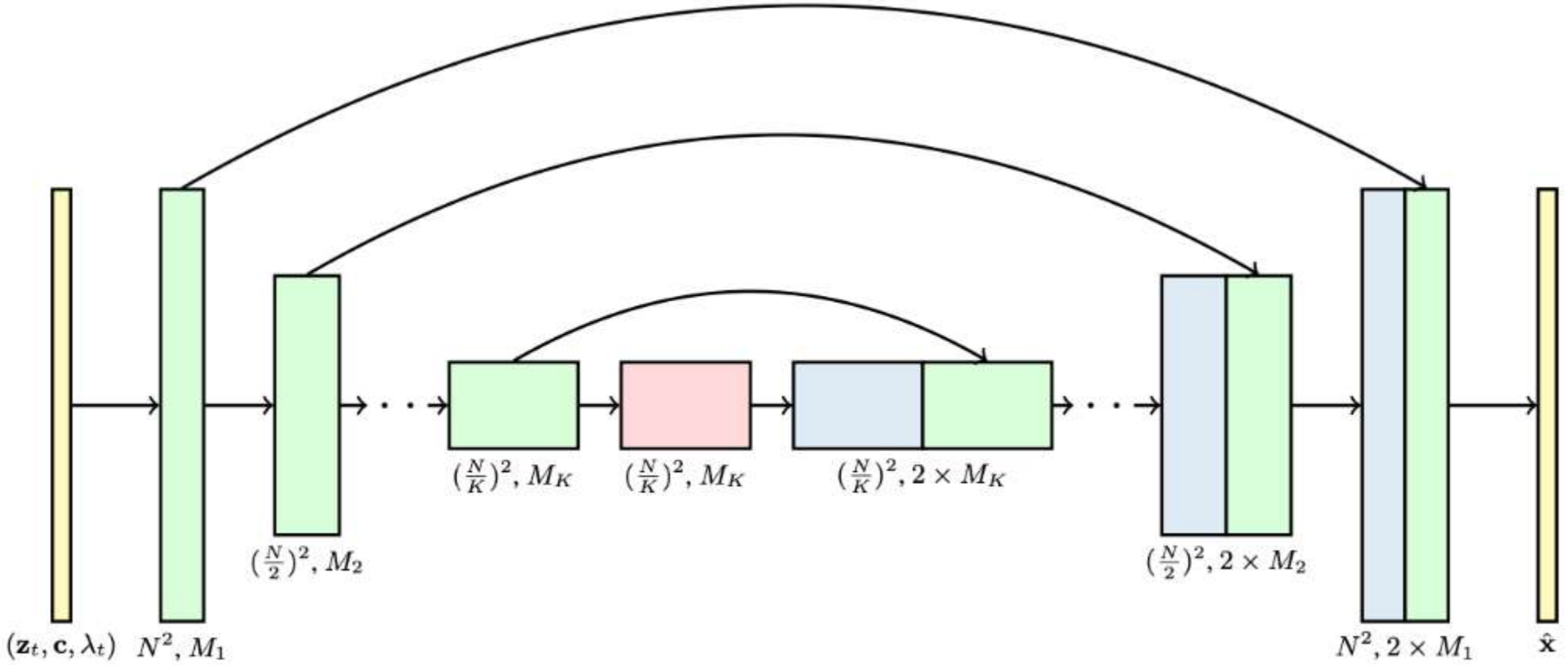


U-Net: https://arxiv.org/abs/1505.04597



DiT: https://arxiv.org/abs/2212.09748

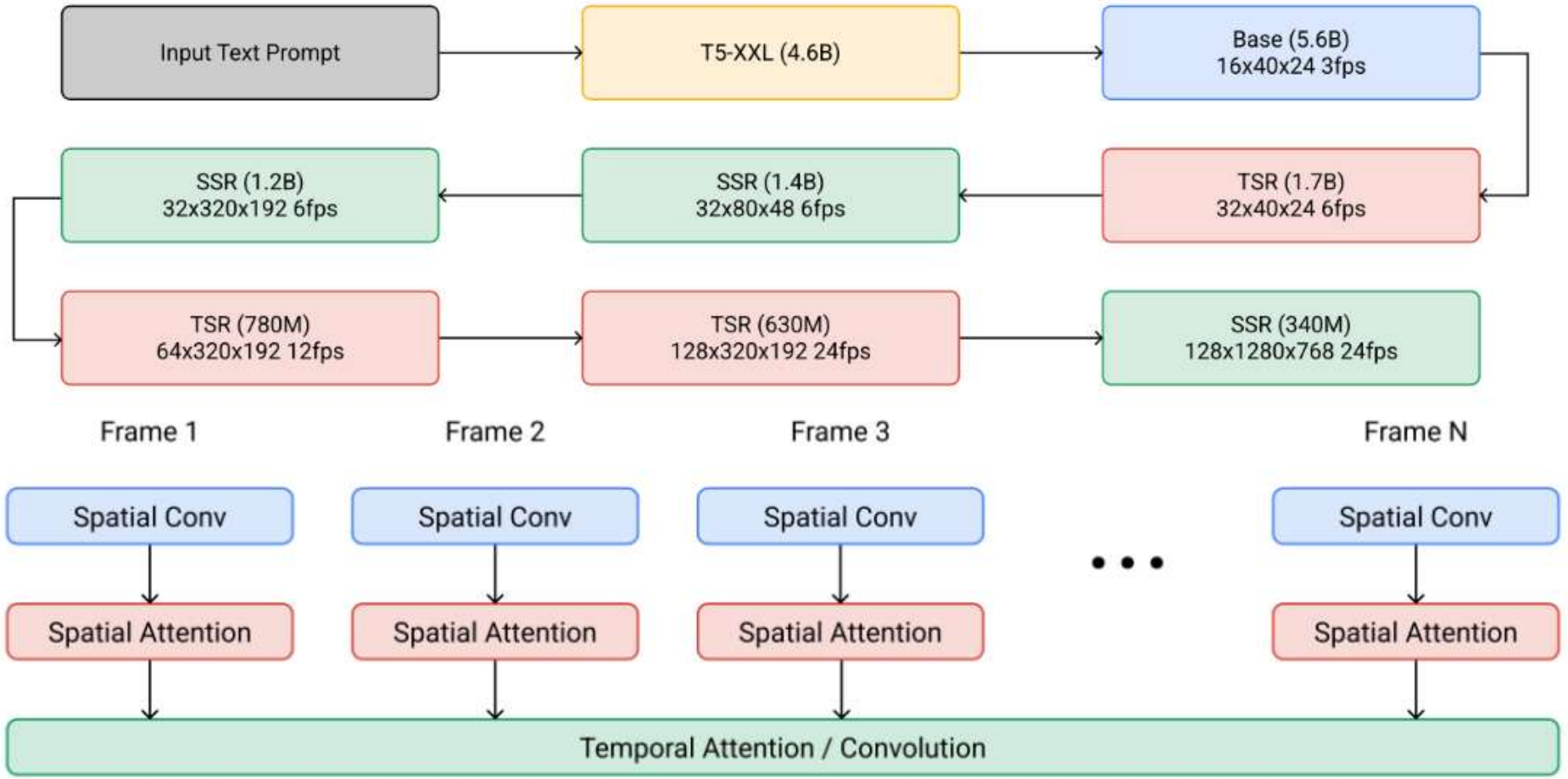
Video Diffusion Models



The 3D U-net architecture. The noisy video z_t , conditioning information c and the log signal-to-noise ratio (log-SNR) λ_t are inputs to the network. The channel multipliers $M_1, \ldots, M_K$ represent the channel counts across layers.

https://arxiv.org/abs/2204.03458

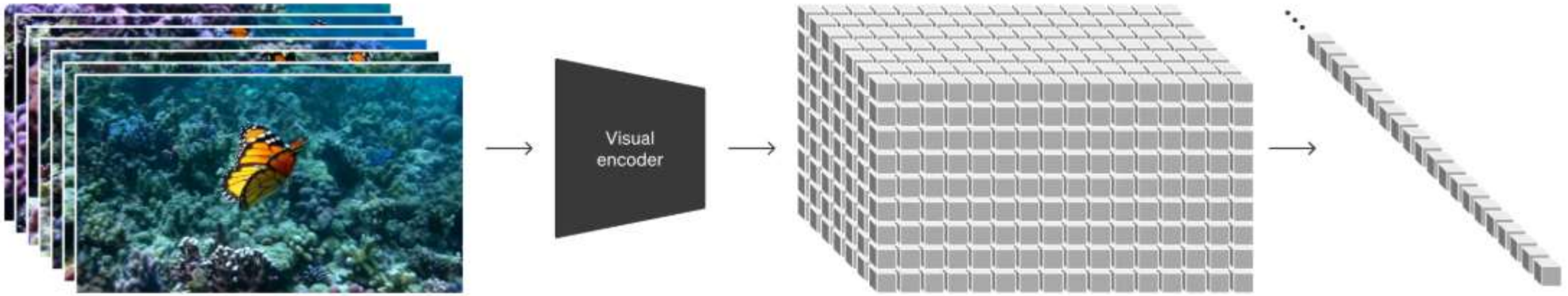
Imagen Video



Upper: The cascaded sampling pipeline in Imagen Video. In practice, the text embeddings are injected into all components, not just the base model.
Below: The architecture of one space-time separable block in the Imagen Video diffusion model.

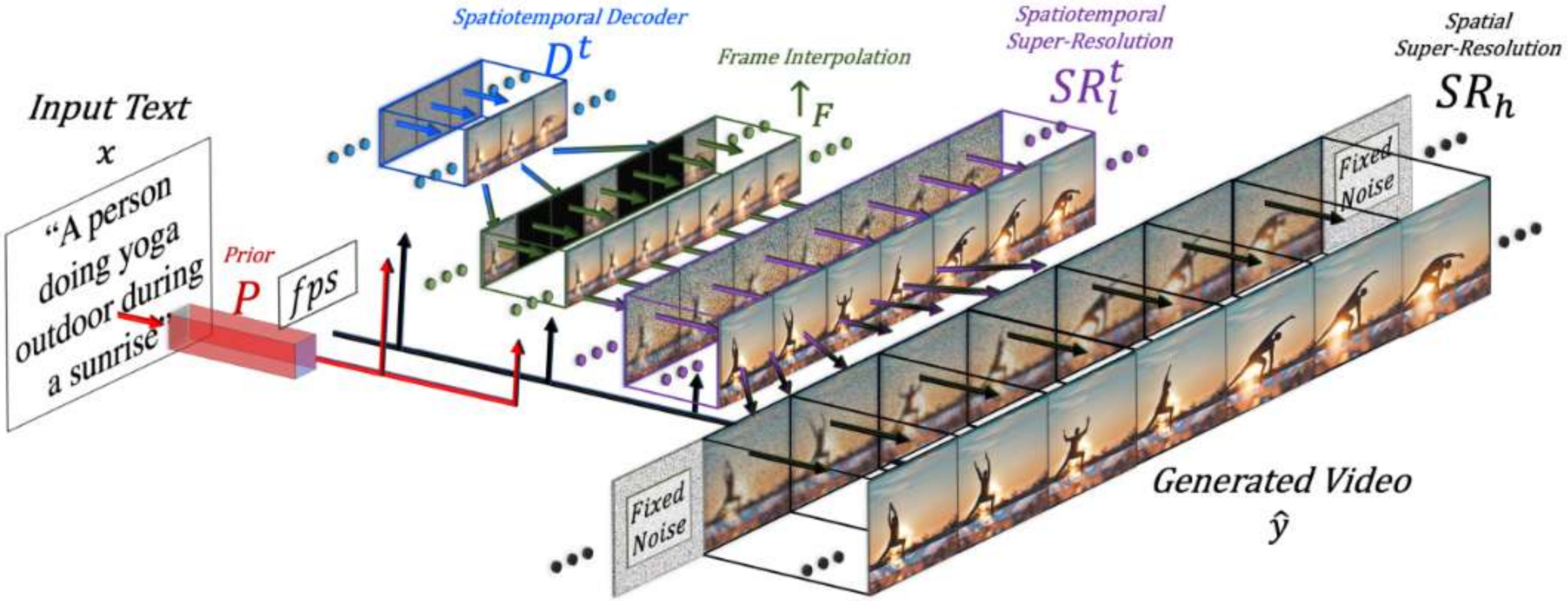
https://arxiv.org/abs/2210.02303

Sora: DiT for Video Generation



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

Adapting Image Models to Generate Videos

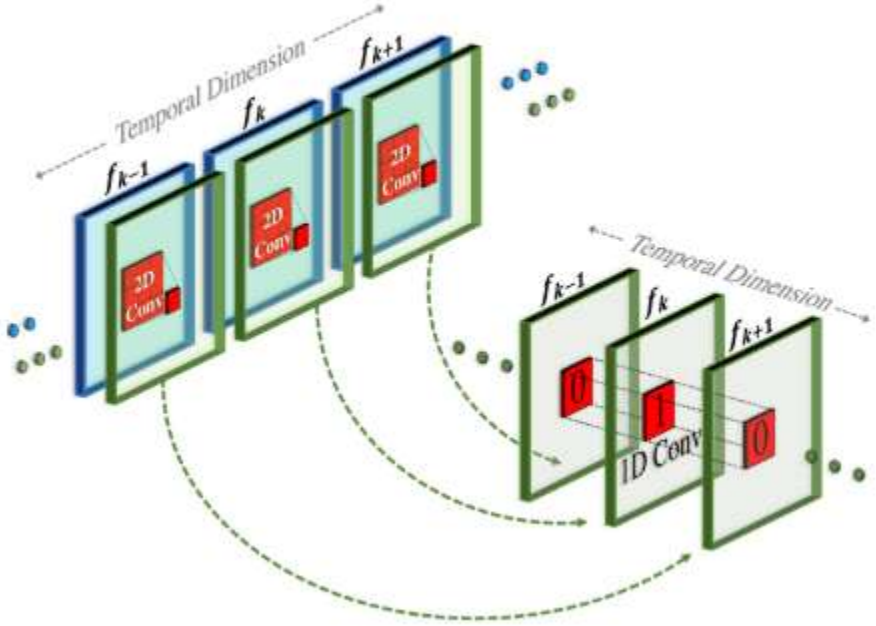


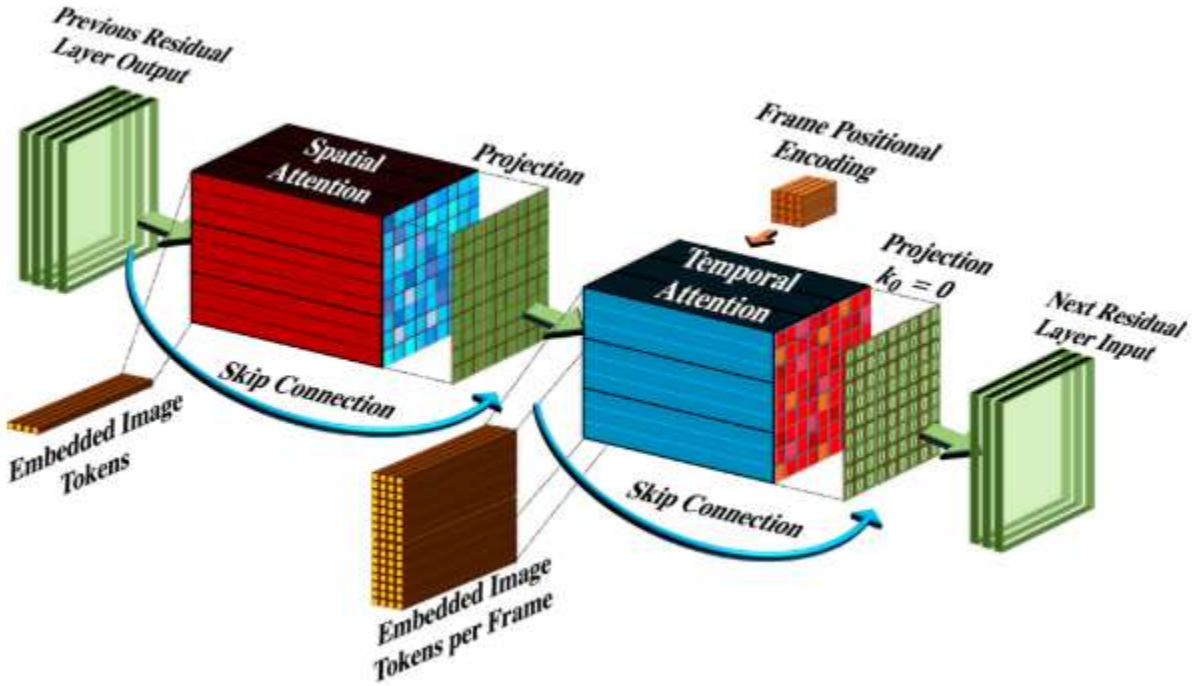
$$\hat{\mathbf{y}}_t = \mathrm{SR}_h \circ \mathrm{SR}_l^t \circ \uparrow_F \circ D^t \circ P \circ (\hat{\mathbf{x}}, \mathrm{CLIP}_{\mathrm{text}}(\mathbf{x}))$$

Make-A-Video https://arxiv.org/abs/2209.14792

Adapting Image Models to Generate Videos

pseudo-3D convo layers and pseudo-3D attention layers





$$\mathrm{Conv}_{\mathrm{P3D}} = \mathrm{Conv}_{\mathrm{1D}}(\mathrm{Conv}_{\mathrm{2D}}(\mathbf{h}) \circ T) \circ T$$

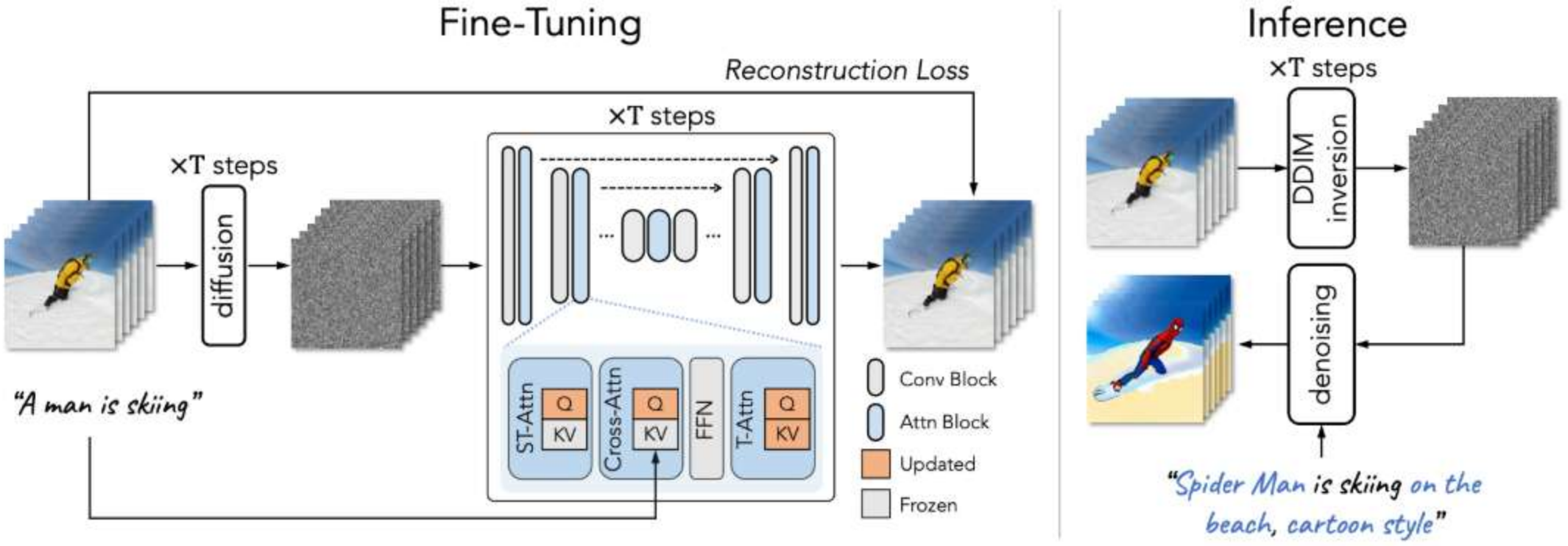
$$\mathrm{Attn}_{\mathrm{P3D}} = \mathrm{flatten}^{-1}(\mathrm{Attn}_{\mathrm{1D}}(\mathrm{Attn}_{\mathrm{2D}}(\mathrm{flatten}(\mathbf{h})) \circ T) \circ T)$$

Make-A-Video https://arxiv.org/abs/2209.14792

Adapting Image Models to Generate Videos

https://ai.meta.com/blog/generative-ai-text-to-video/

Adapting Image Models to Generate Videos



$$\mathbf{Q} = \mathbf{W}^Q \mathbf{z}_{v_i}, \quad \mathbf{K} = \mathbf{W}^K [\mathbf{z}_{v_1}, \mathbf{z}_{v_{i-1}}], \quad \mathbf{V} = \mathbf{W}^V [\mathbf{z}_{v_1}, \mathbf{z}_{v_{i-1}}]$$

$$\mathbf{O} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \cdot \mathbf{V}$$

Tune-A-Video https://arxiv.org/abs/2212.11565

Adapting Image Models to Generate Videos

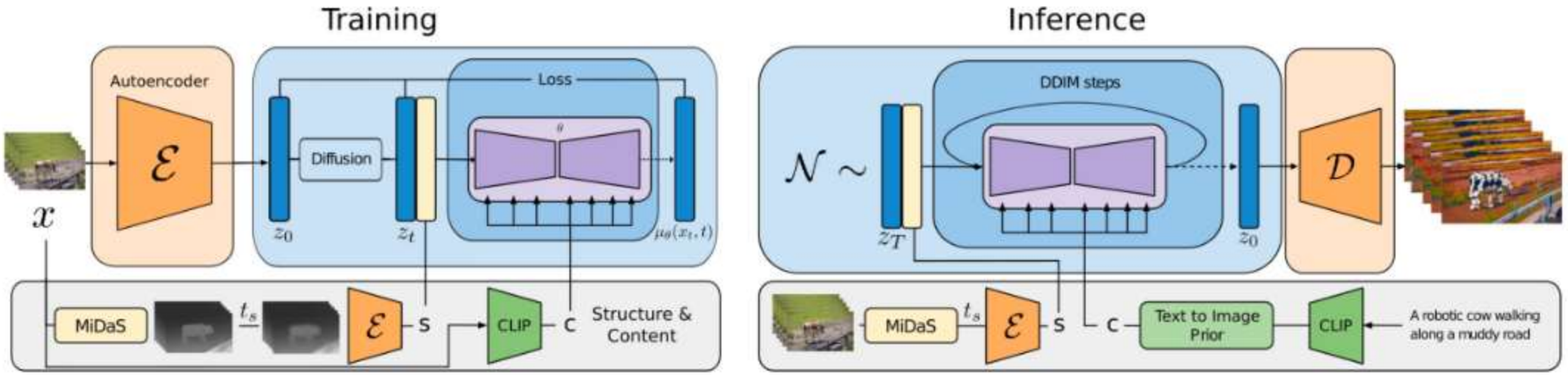
"A man is skiing"

"Spider Man is skiing on the beach, cartoon style”

"A man, wearing pink clothes, is skiing at sunset"

Tune-A-Video https://arxiv.org/abs/2212.11565

Adapting Image Models to Generate Videos



Gen-1 https://arxiv.org/abs/2302.03011

Adapting Image Models to Generate Videos

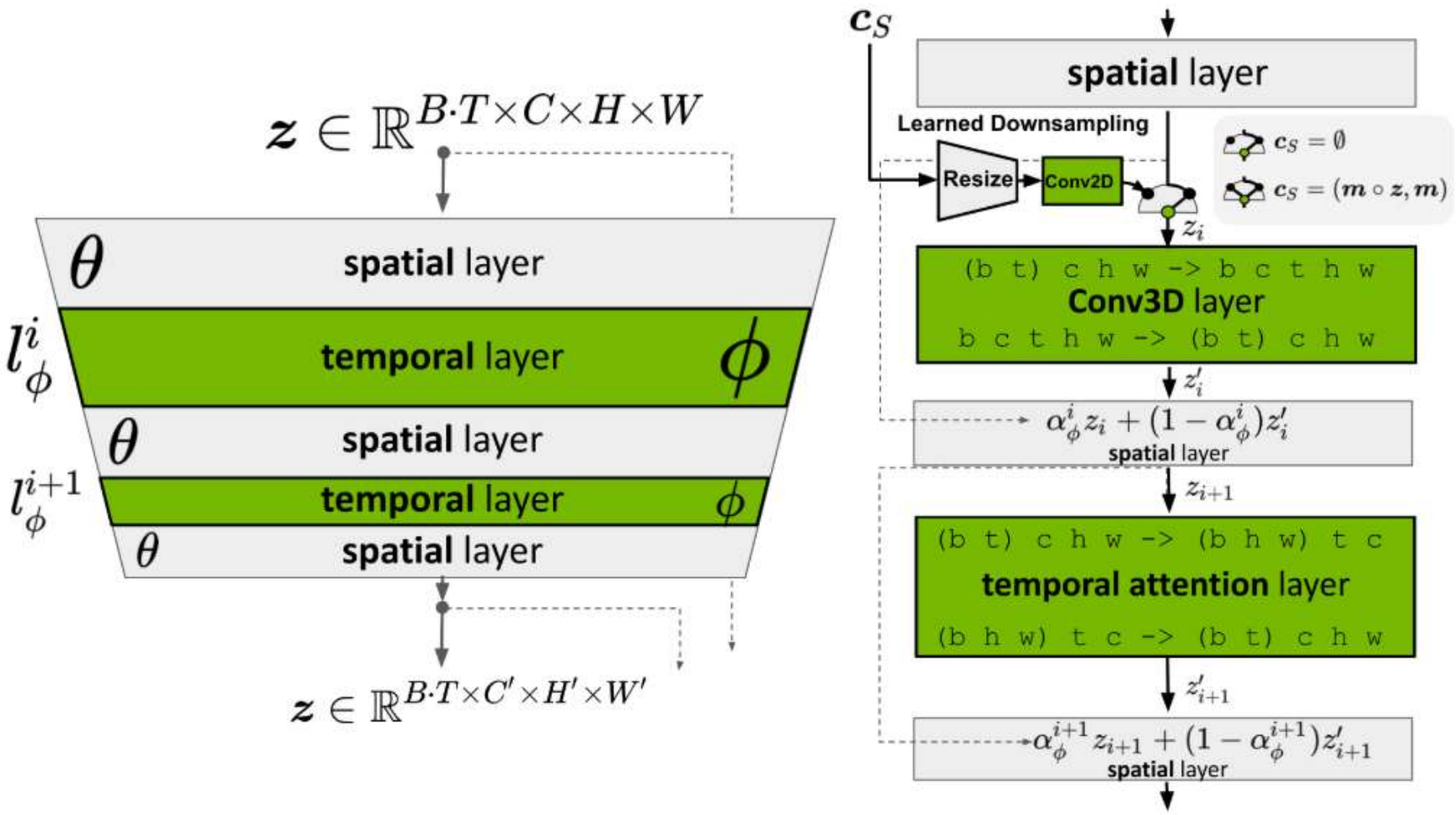
Gen-1 https://arxiv.org/abs/2302.03011

Recap.

Bilibili BV1Lg4y1t797 with ControlNet

https://arxiv.org/abs/2302.05543

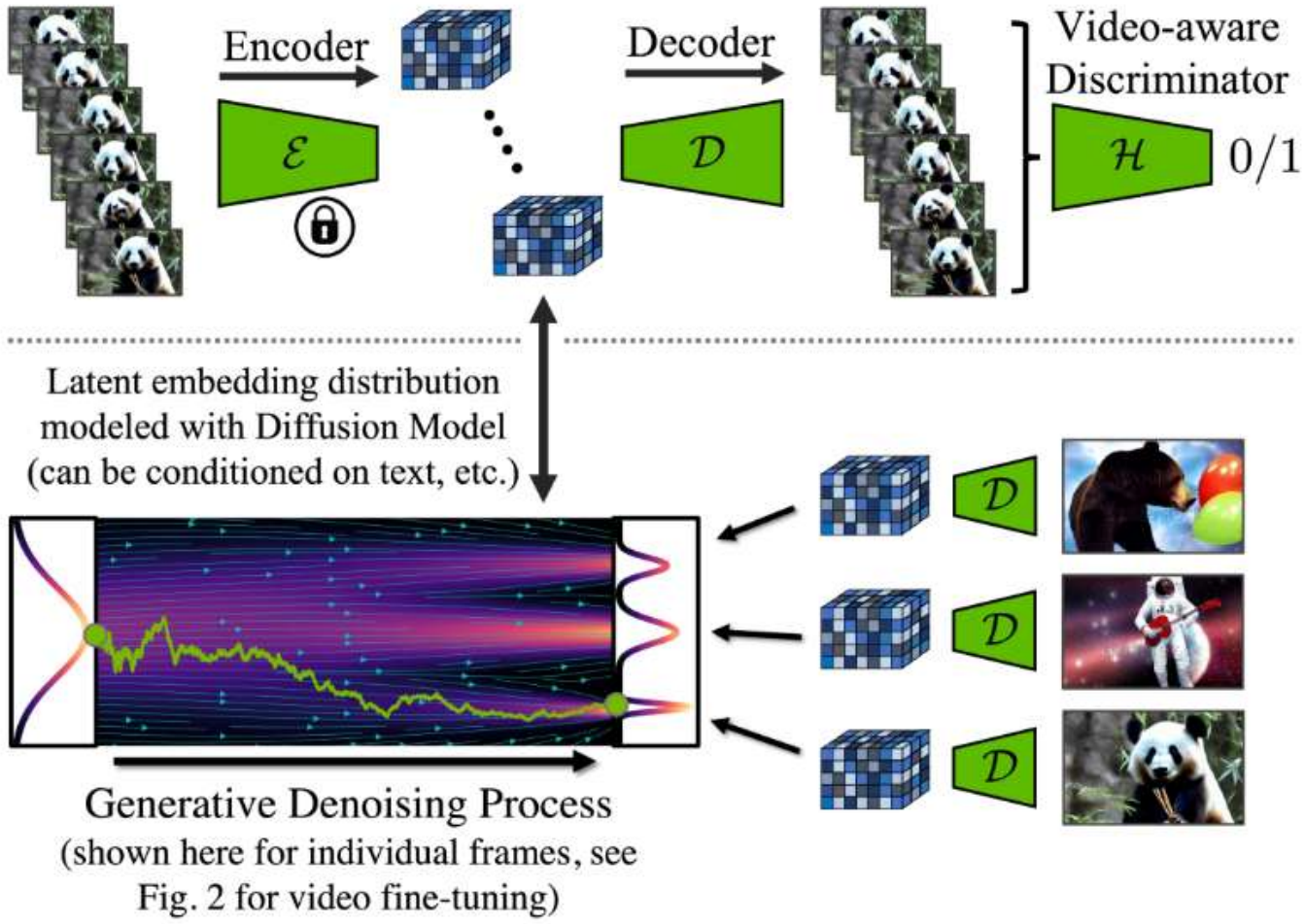
Adapting Image Models to Generate Videos



Base image model θ

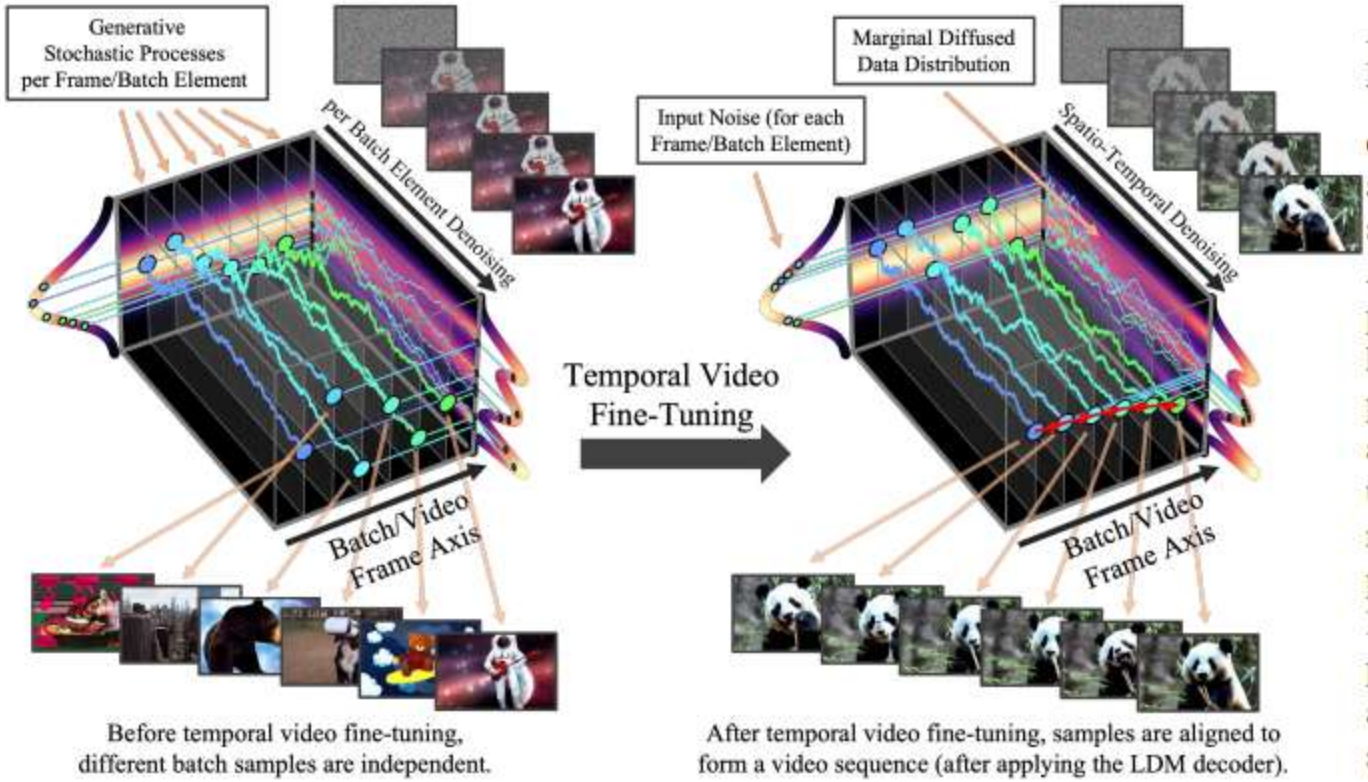
Video LDM https://arxiv.org/abs/2304.08818

Adapting Image Models to Generate Videos



Video LDM https://arxiv.org/abs/2304.08818

Adapting Image Models to Generate Videos



Before temporal video fine-tuning, different batch samples are independent.

After temporal video fine-tuning, samples are aligned to form a video sequence (after applying the LDM decoder).

Figure 2. **Temporal Video Fine-Tuning.** We turn pre-trained image diffusion models into temporally consistent video generators. Initially, different samples of a batch synthesized by the model are independent. After temporal video fine-tuning, the samples are temporally aligned and form coherent videos. The stochastic generation process before and after fine-tuning is visualised for a diffusion model of a one-dim. toy distribution. For clarity, the figure corresponds to alignment in pixel space. In practice, we perform alignment in LDM's latent space and obtain videos after applying LDM's decoder (see Fig. 3). We also video fine-tune diffusion model upsamplers in pixel or latent space (Sec. 3.4).

Video LDM https://arxiv.org/abs/2304.08818

Adapting Image Models to Generate Videos

Video LDM https://arxiv.org/abs/2304.08818

Adapting Image Models to Generate Videos

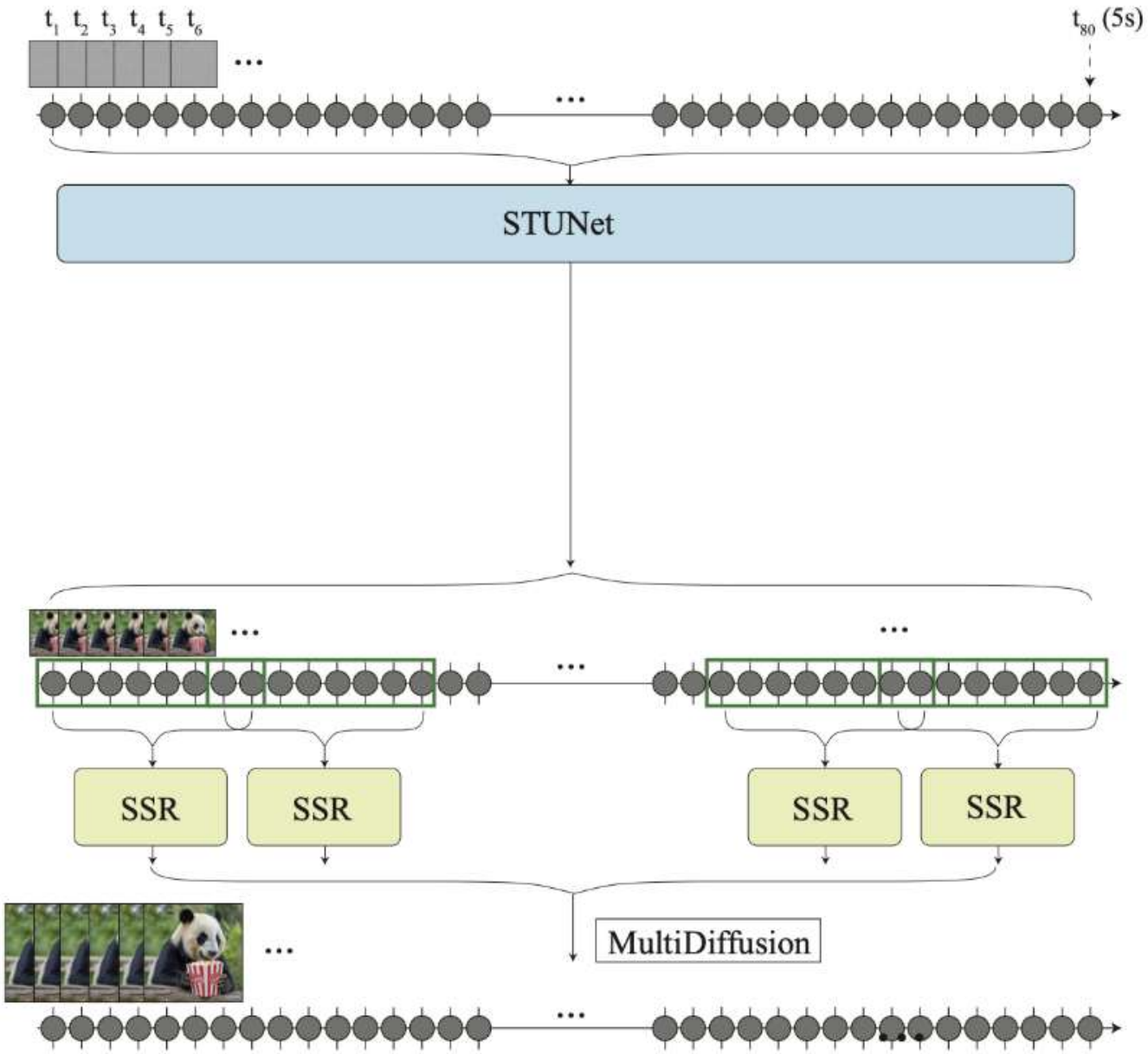
"A robot dj is playing the turntables, in heavy raining futuristic tokyo, rooftop, sci-fi, fantasy"

"An exploding cheese house"

"A fat rabbit wearing a purple robe walking through a fantasy landscape"

SVD https://arxiv.org/abs/2311.15127

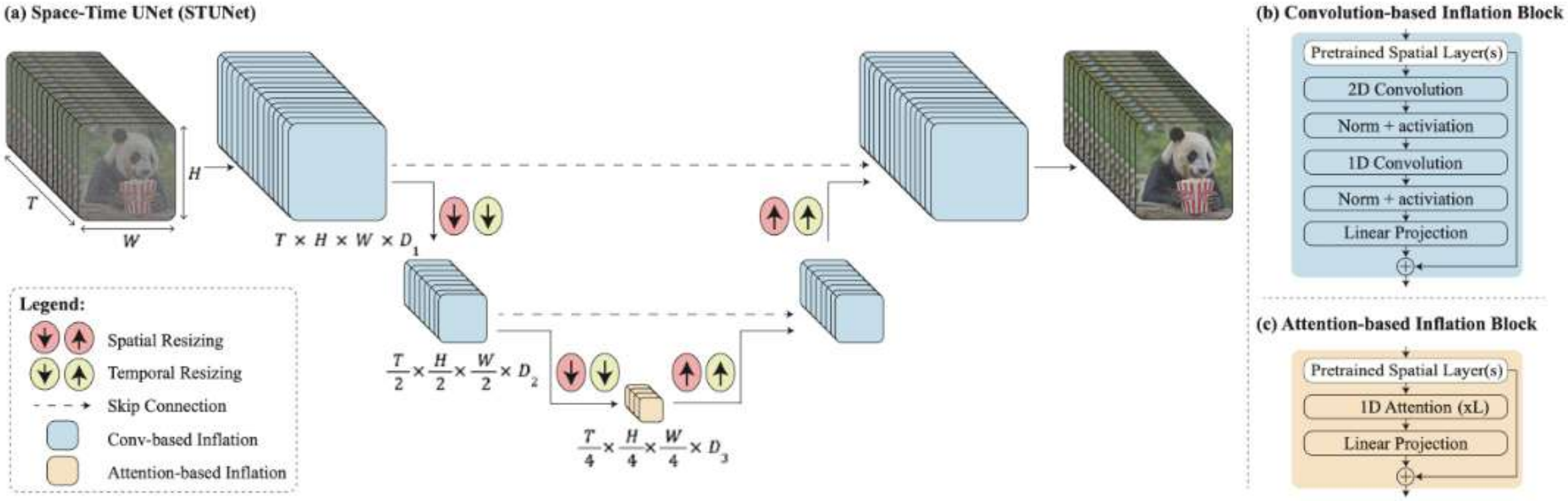
Adapting Image Models to Generate Videos



Lumiere removes TSR (temporal super-resolution) models. The inflated SSR network can operate only on short segments of the video due to memory constraints and thus SSR models operate on a set of shorter but overlapped video snippets.

Lumiere https://arxiv.org/abs/2401.12945

Adapting Image Models to Generate Videos



Lumiere https://arxiv.org/abs/2401.12945

Adapting Image Models to Generate Videos

wearing a gold strapless gown

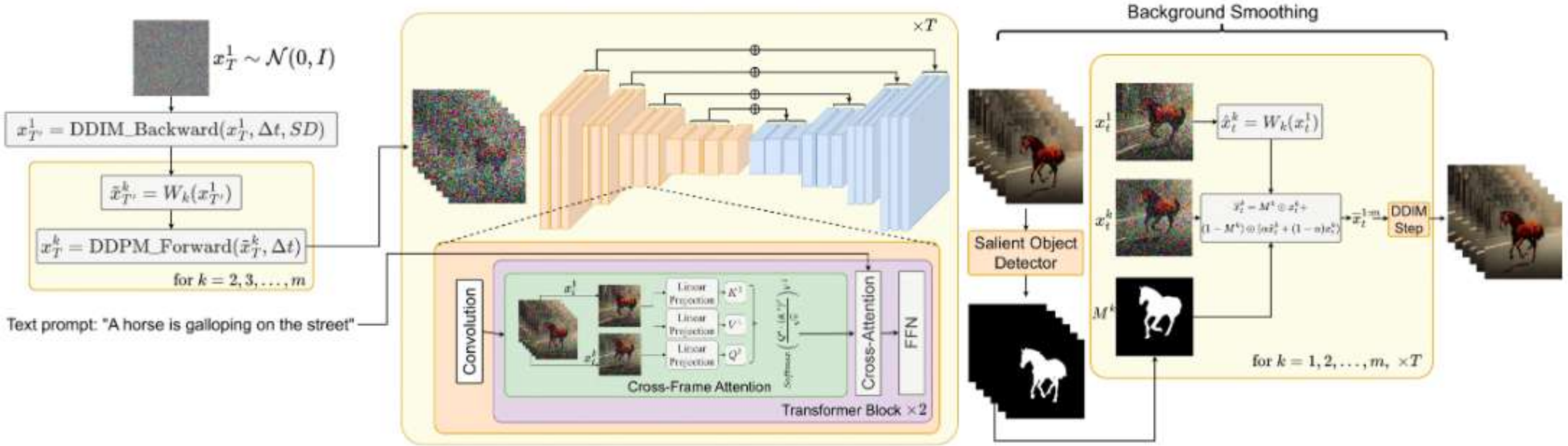
wearing a bath robe

Lumiere https://arxiv.org/abs/2401.12945

Training-Free Adaptation

Text2Video-Zero https://arxiv.org/abs/2303.13439

Training-Free Adaptation



$$\mathbf{x}_{T'}^1 = \text{DDIM-backward}(\mathbf{x}_T^1, \Delta t) \text{ where } T' = T - \Delta t$$

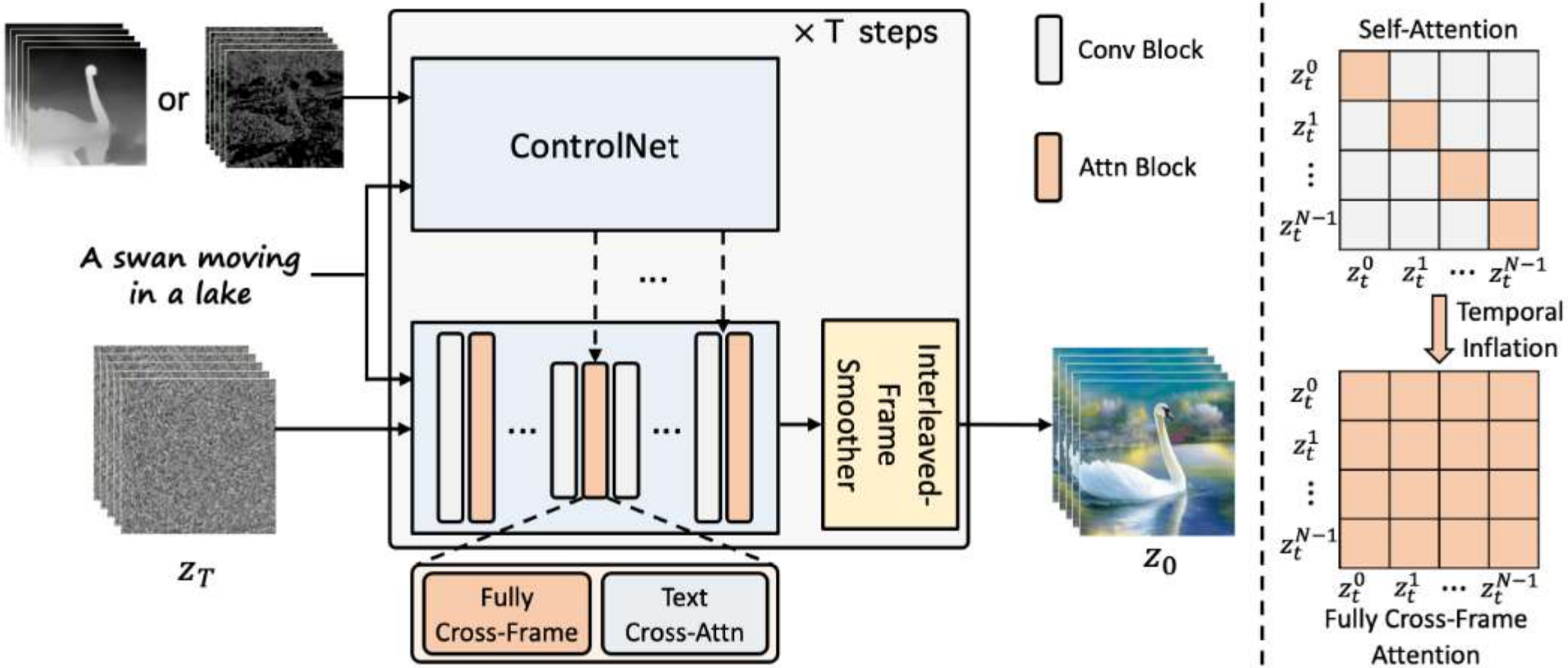
$$W_k \leftarrow \text{a warping operation of } \boldsymbol{\delta}^k = \lambda(k-1)\boldsymbol{\delta}$$

$$\tilde{\mathbf{x}}_{T'}^k = W_k(\mathbf{x}_{T'}^1)$$

$$\mathbf{x}_T^k = \text{DDIM-forward}(\tilde{\mathbf{x}}_{T'}^k, \Delta t) \text{ for } k = 2, \ldots, m$$

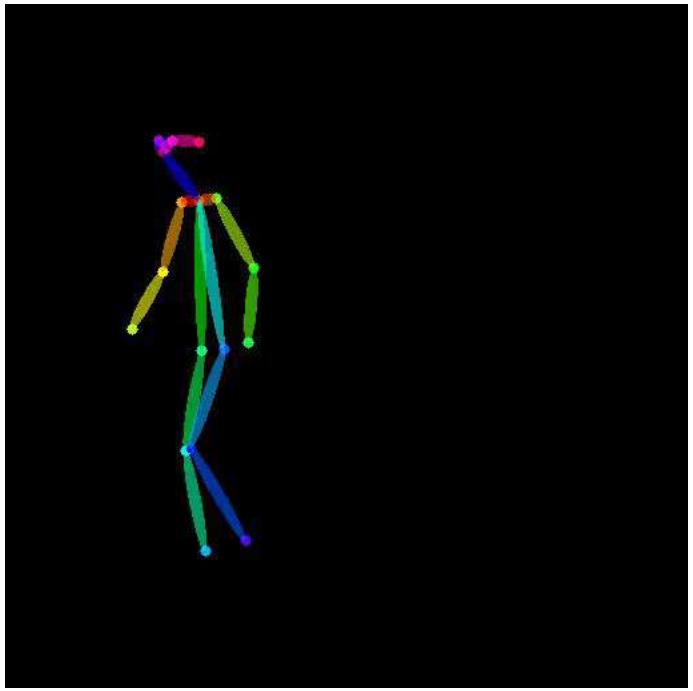
Text2Video-Zero https://arxiv.org/abs/2303.13439

Training-Free Adaptation



ControlVideo https://arxiv.org/pdf/2305.13077

Training-Free Adaptation



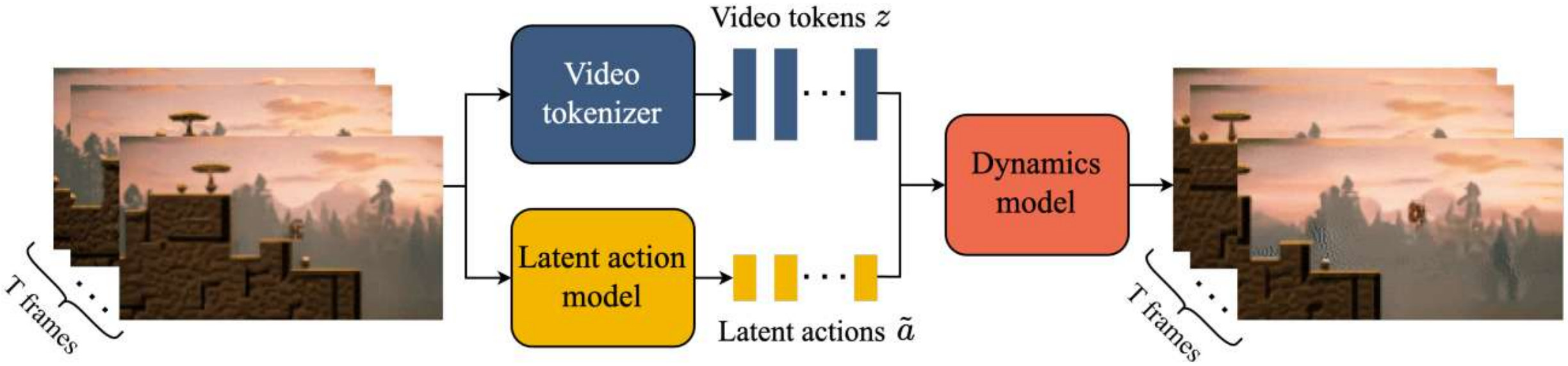
"A man, wearing pink clothes, moonwalk at sunset."

ControlVideo https://arxiv.org/pdf/2305.13077

Genie: Generative Interactive Environments

https://sites.google.com/view/genie-2024/home

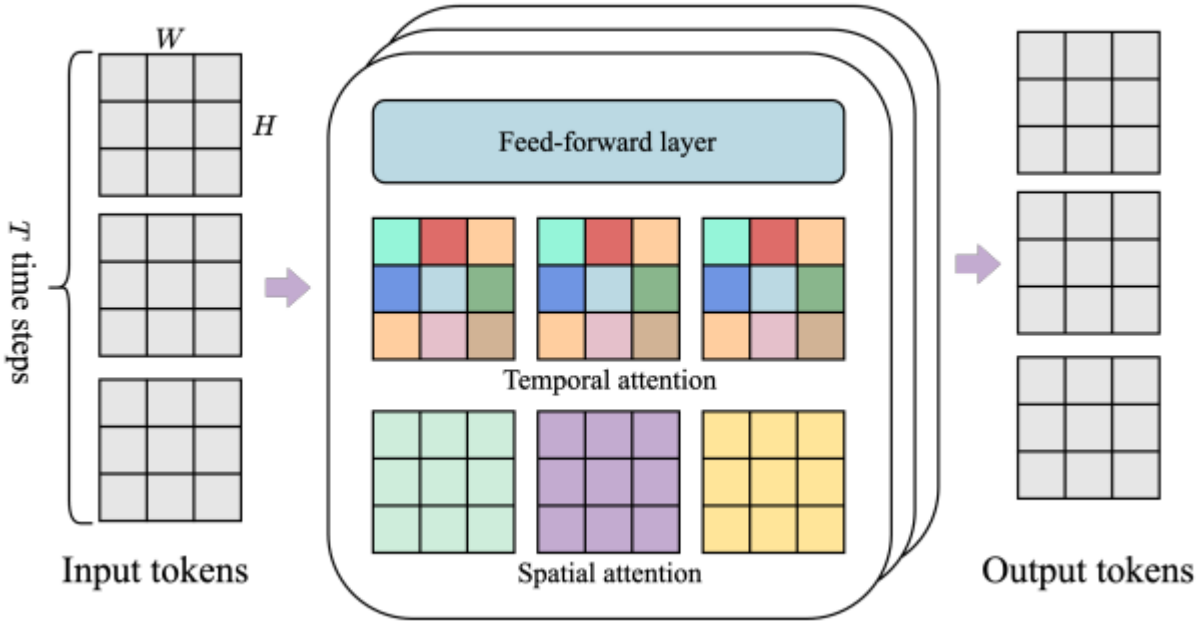
Genie: Generative Interactive Environments



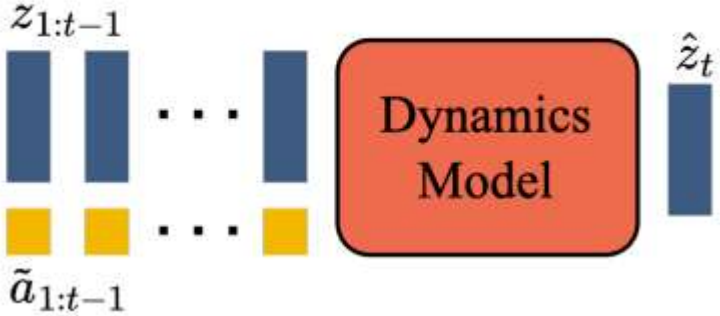
Genie takes in T frames of video as input, tokenizes them into discrete tokens $\boldsymbol{z}$ via the video tokenizer, and infers the latent actions $\boldsymbol{\tilde{a}}$ between each frame with the latent action model. Both are then passed to the dynamics model to generate predictions for the next frames in an iterative manner.

https://sites.google.com/view/genie-2024/home

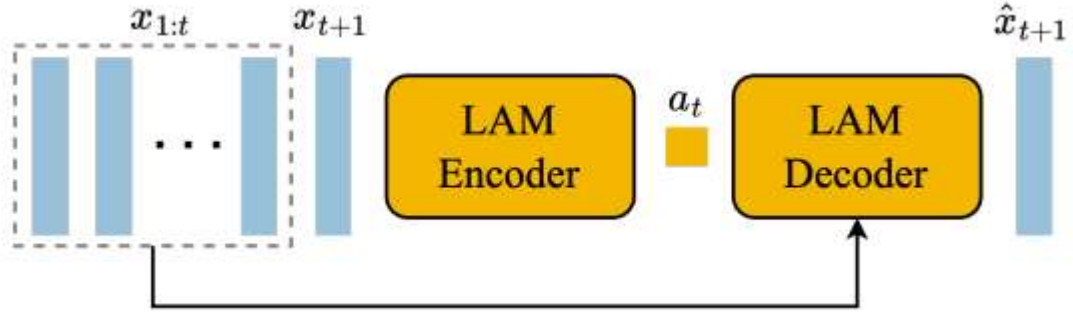
Genie: Generative Interactive Environments



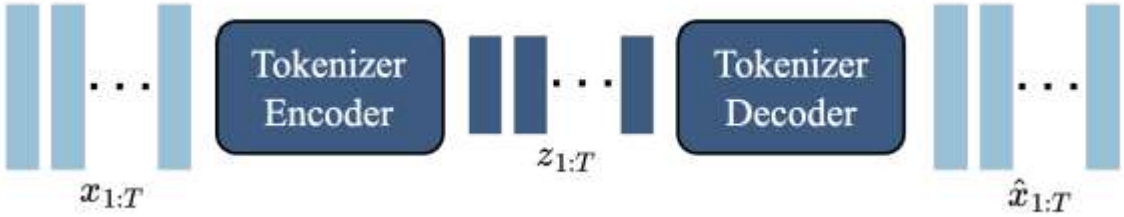
ST-transformer architecture



Dynamics model



Latent action model



Video tokenizer

https://sites.google.com/view/genie-2024/home

Genie: Generative Interactive Environments

“Remarkably, Genie learns not only which parts of an observation are generally controllable, but also infers diverse latent actions that are consistent across the generated environments. Note here how the same latent actions yield similar behaviors across different prompt images. ”

https://sites.google.com/view/genie-2024/home

Genie: Generative Interactive Environments

The future of generative virtual worlds

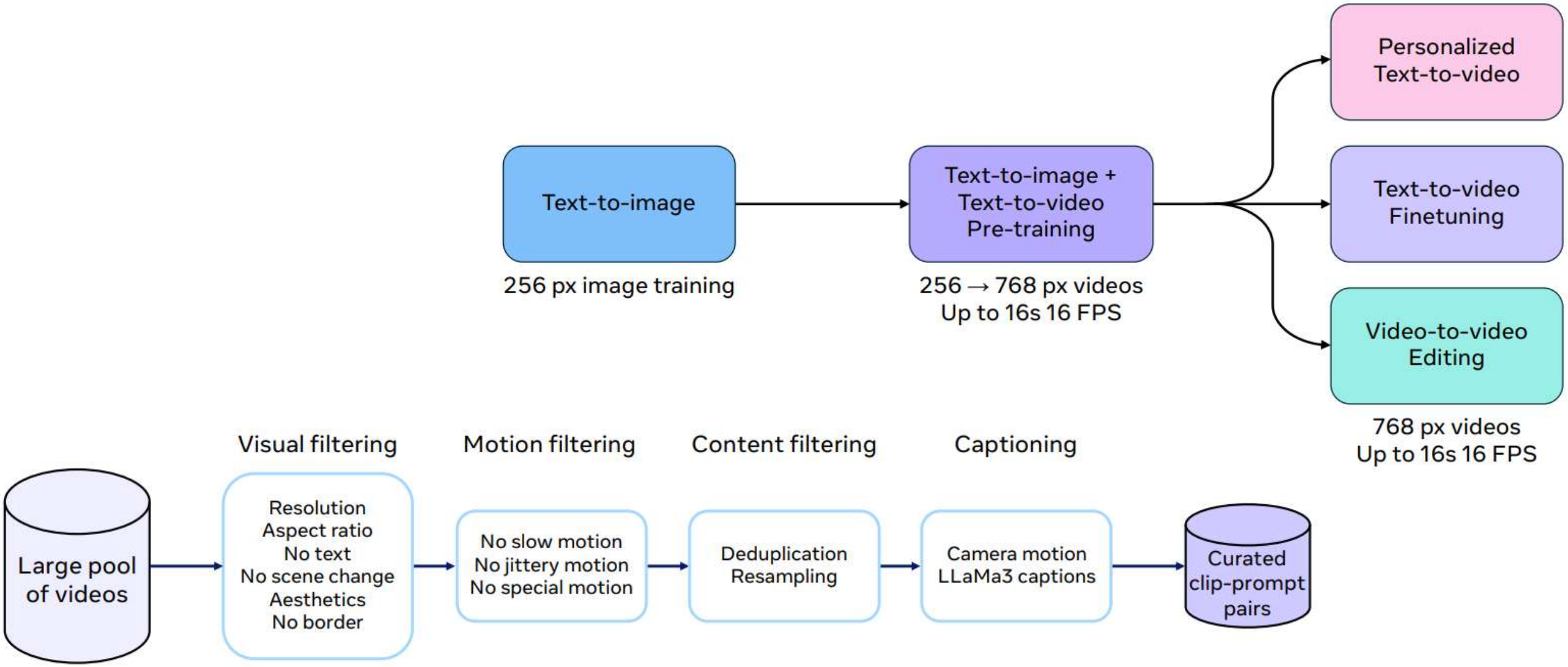
“Trajectories with the same latent action sequence typically display similar behaviors. ”

https://sites.google.com/view/genie-2024/home

Movie Gen: A Cast of Media Foundation Models

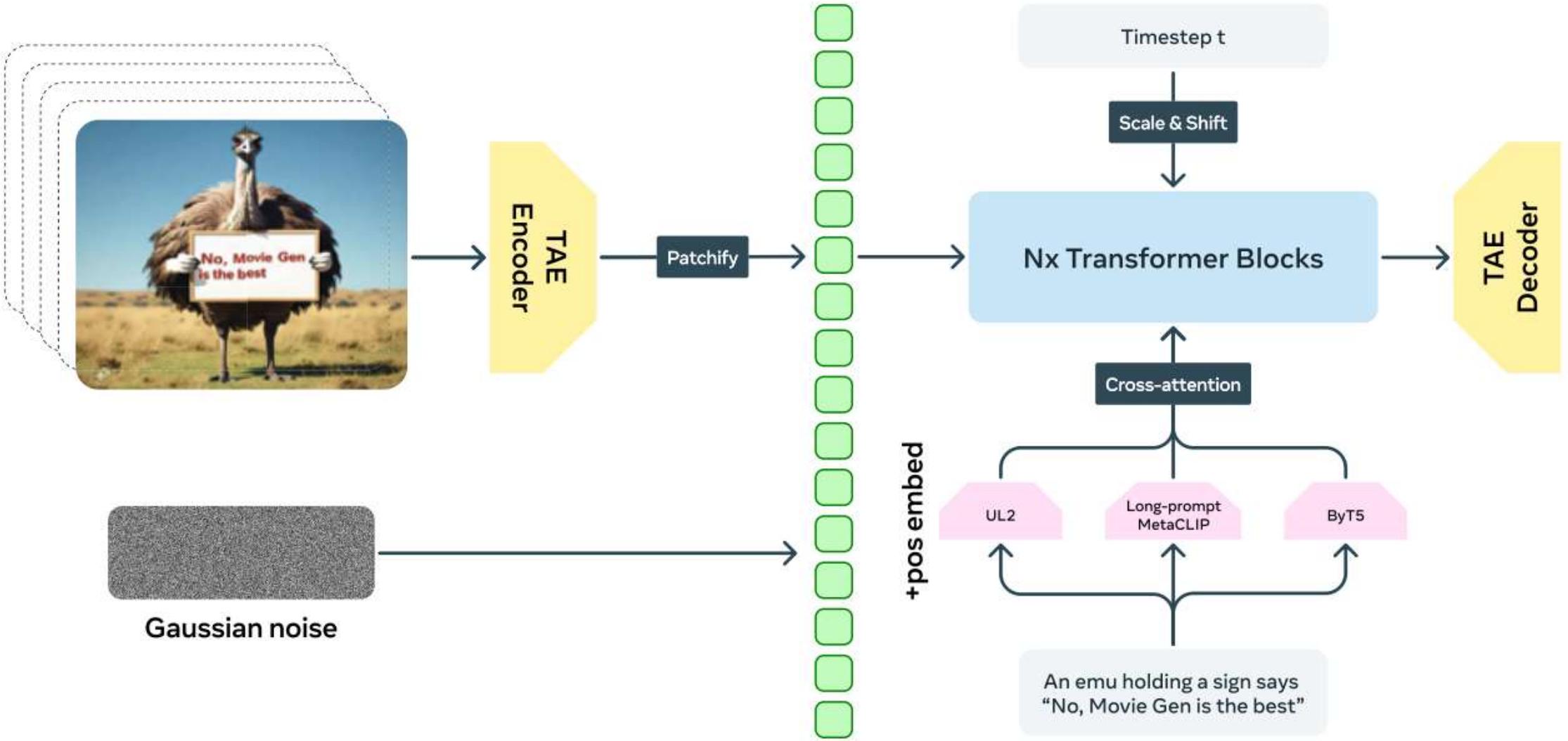
https://ai.meta.com/blog/movie-gen-media-foundation-models-generative-ai-video/

Movie Gen:
A Cast of Media Foundation Models



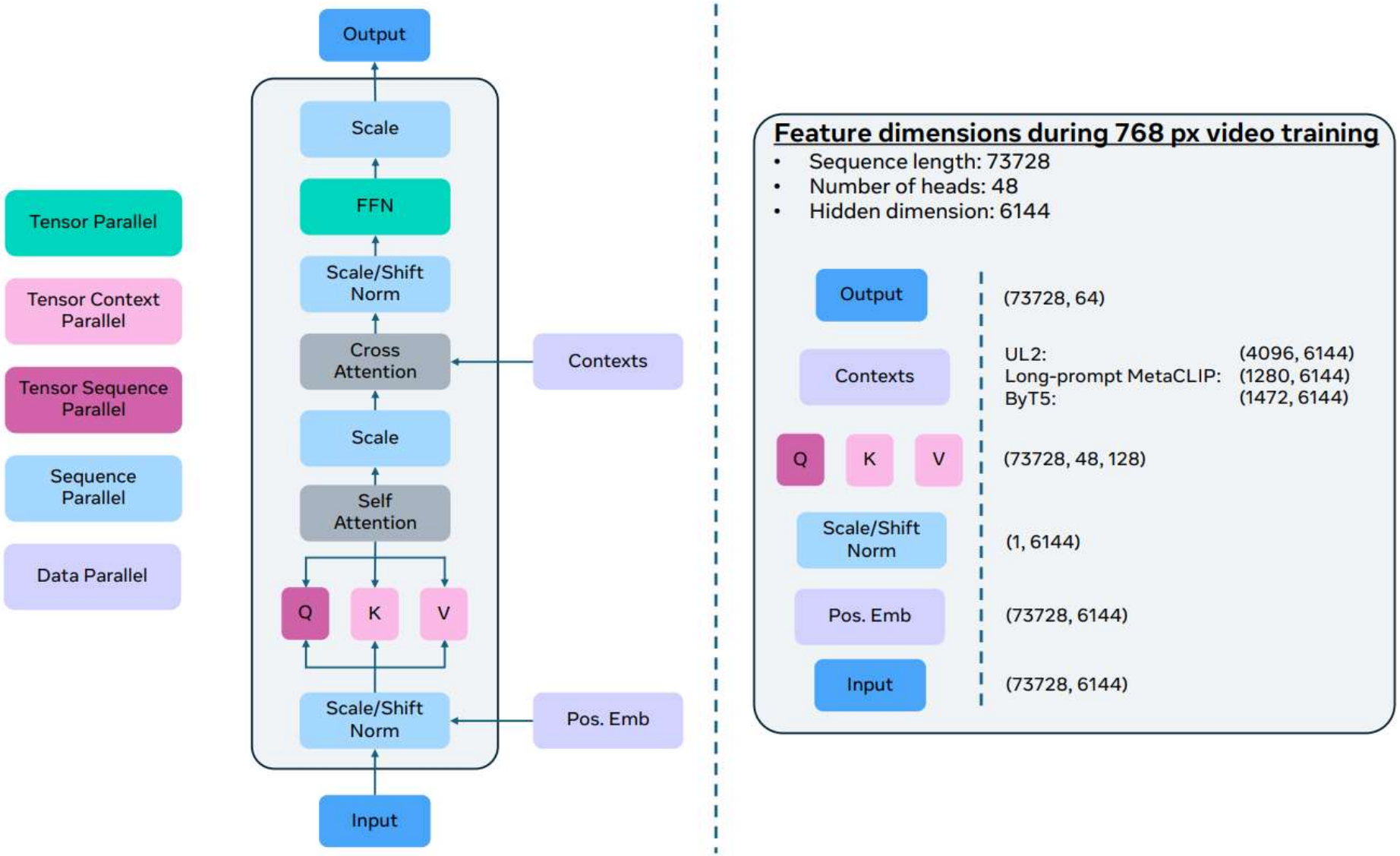
https://ai.meta.com/static-resource/movie-gen-research-paper

Movie Gen: A Cast of Media Foundation Models



https://ai.meta.com/static-resource/movie-gen-research-paper

Movie Gen: A Cast of Media Foundation Models



https://ai.meta.com/static-resource/movie-gen-research-paper

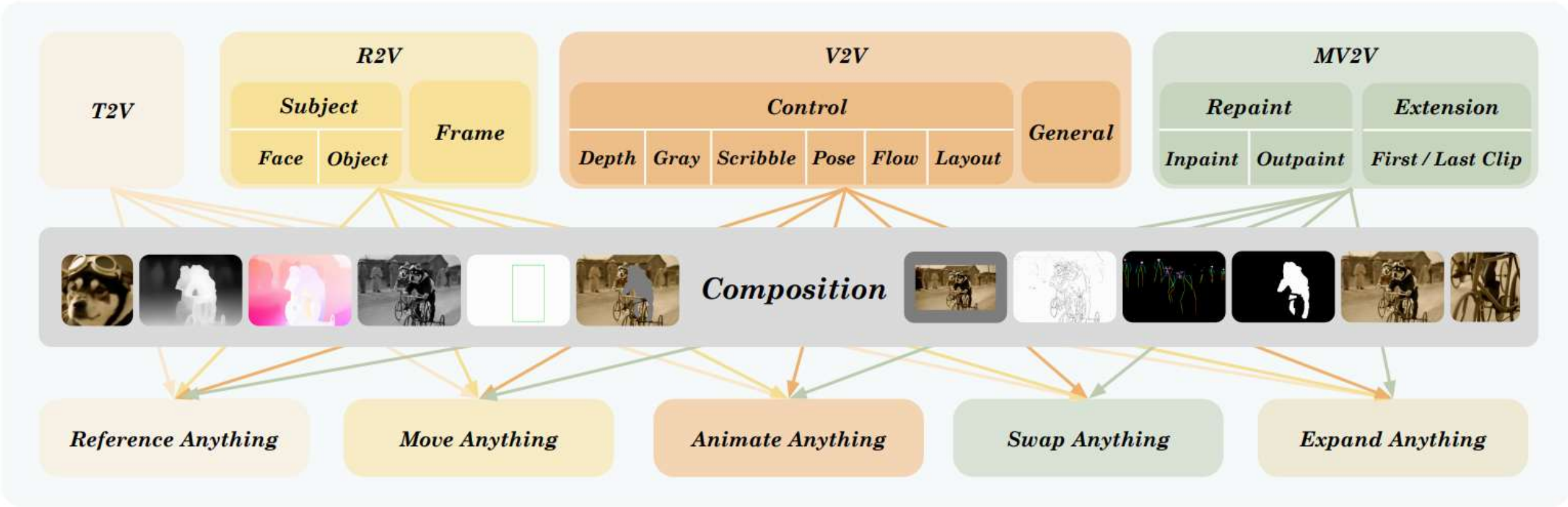
One Model to Workflow

*image taken from ChatGPT 4o

VACE/Wan

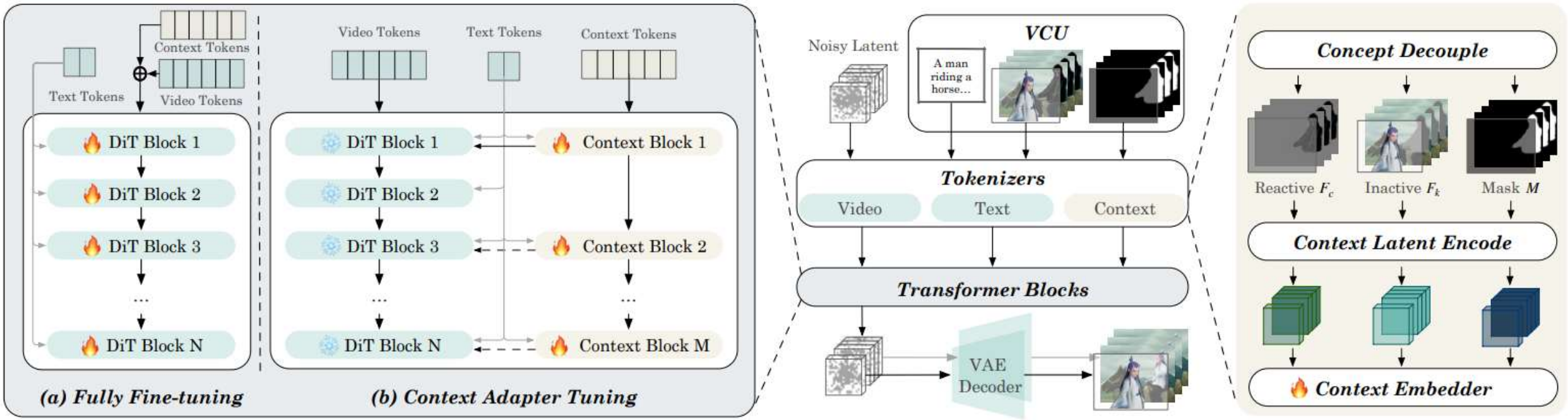
https://www.alibabacloud.com/en/press-room/alibaba-introduces-open-source-model-for-video

VACE



https://arxiv.org/pdf/2503.07598

VACE



https://arxiv.org/pdf/2503.07598

VACE

https://arxiv.org/pdf/2503.07598

Sora2 & Veo3.1: Omni Model / “WORLD MODEL”

From Silent Video to Native Audiovisual Generation

- The target is synchronized audiovisual events.
- Dialogue, ambience, and sound effects must cohere with motion.
- New bottlenecks: lip-sync, event-sound causality, and cross-shot continuity.

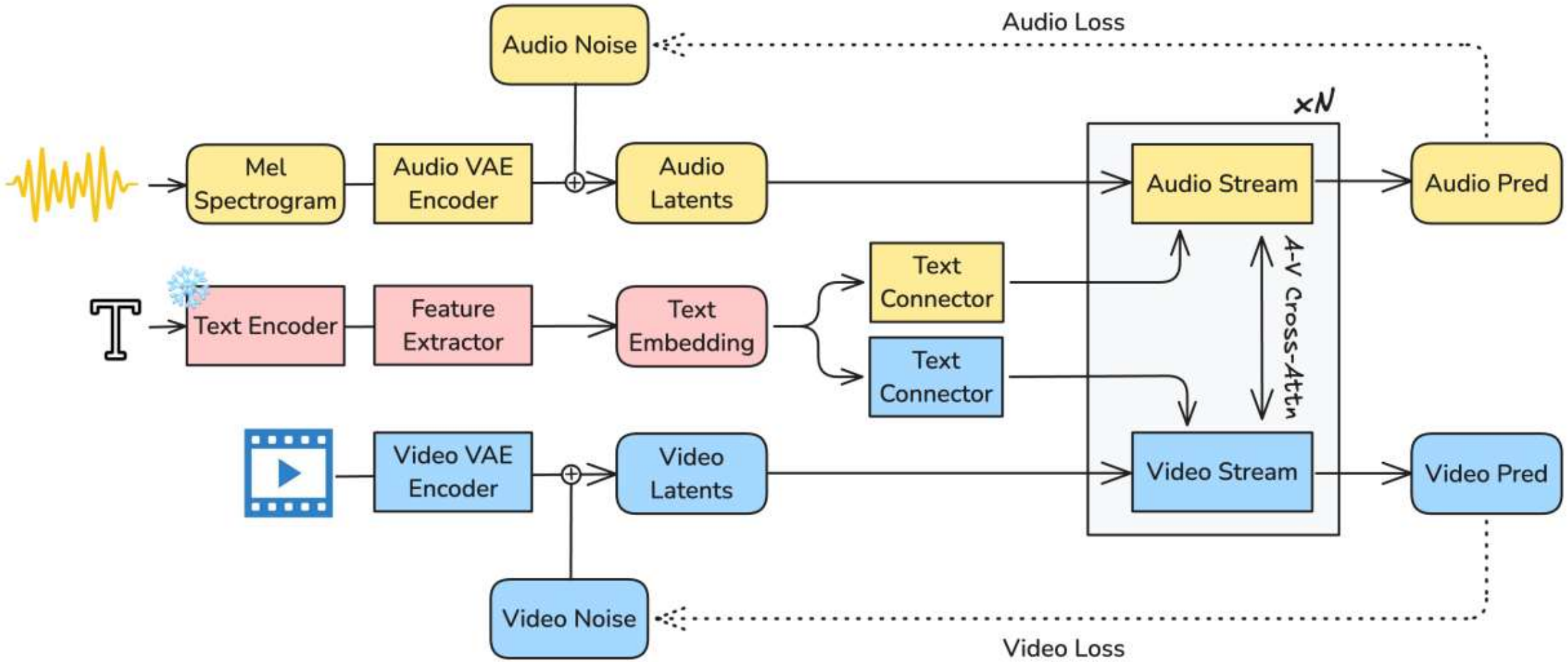
- Sora 2 improves physics, realism, synchronized audio, and steerability.
- Prompts now combine subject, camera, pacing, and audio intent.
- Storyboard, remix, stitch, extend, and characters define the workflow.

https://sora.chatgpt.com/explore

- Veo 3.1 adds richer native audio and stronger prompt adherence.
- Reference images, first/last frames, and extension expose creative control.
- A strong Veo prompt reads like a director's brief.

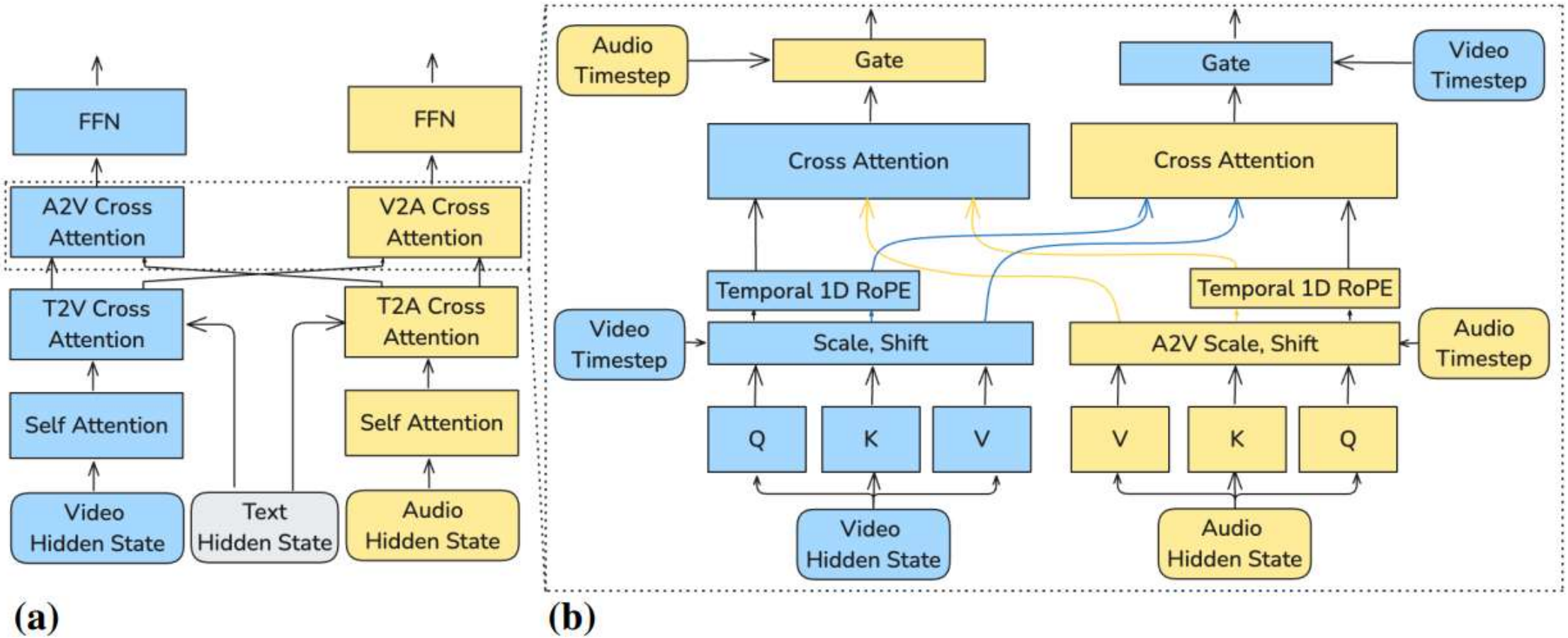
https://blog.google/innovation-and-ai/products/veo-updates-flow/?utm_source=chatgpt.com

LTX



https://arxiv.org/pdf/2601.03233

LTX



https://arxiv.org/pdf/2601.03233

LTX

Figure 3: **Visualization of AV cross-attention maps.** The maps are averaged across attention heads and the model layers; V2A and A2V maps correspond to the first and last 1/3 of inference steps, respectively. Red vertical lines on the audio waveform mark the timestamps of the displayed frames. The visualization demonstrates the model's ability to spatially track a moving vehicle, dynamically shift attention from one speaker to another and then to both simultaneously, and focus on the lip region during close-up speech.

https://arxiv.org/pdf/2601.03233

Seedance

https://seed.bytedance.com/zh/blog/official-launch-of-seedance-2-0