Prompting in Visual Generation

Ziwei Liu

Nanyang Technological University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Prompting in Generation

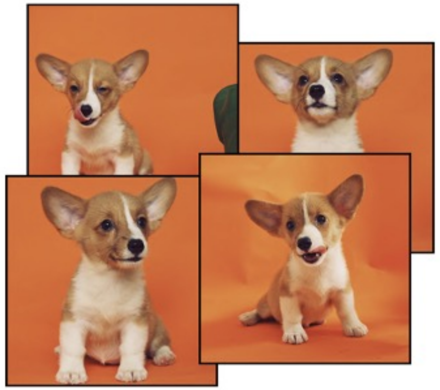


Image Prompt

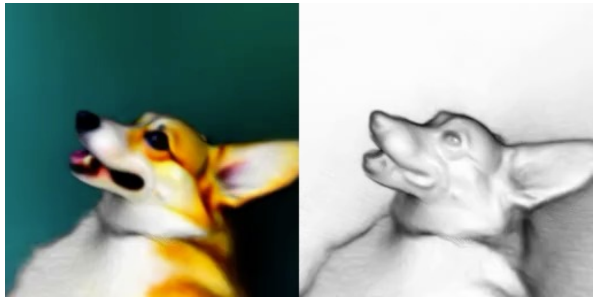
"A Corgi"

Text Prompt

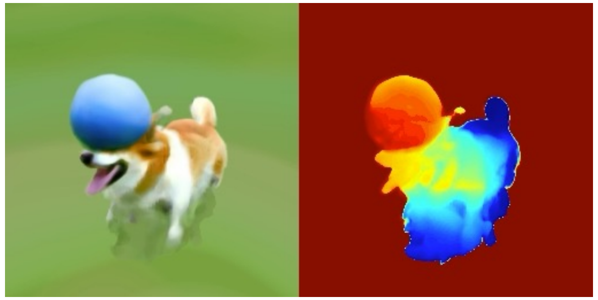
Generative Model

Image

Video



3D



4D Dynamic Scene

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

Text to Image Generation

Text to Image Generation

- Prompt: An astronaut riding a horse in photorealistic style.





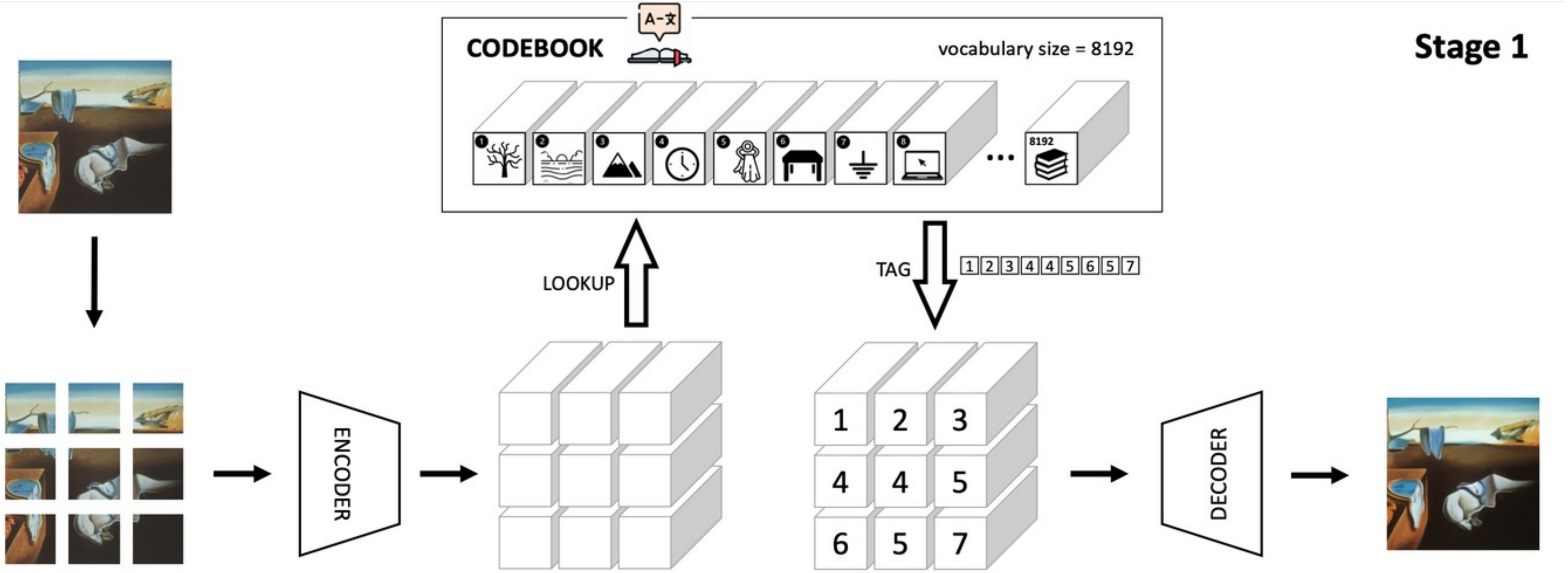
Source: https://openai.com/dall-e-2

Text to Image Generation

- VQGAN-based Methods
  - DALLE
- Diffusion-based Methods
  - GLIDE, DALEE2, Stable Diffusion
- GAN-based Methods
  - GigaGAN
- Generation on Specialized data
  - Text2Human

DALLE

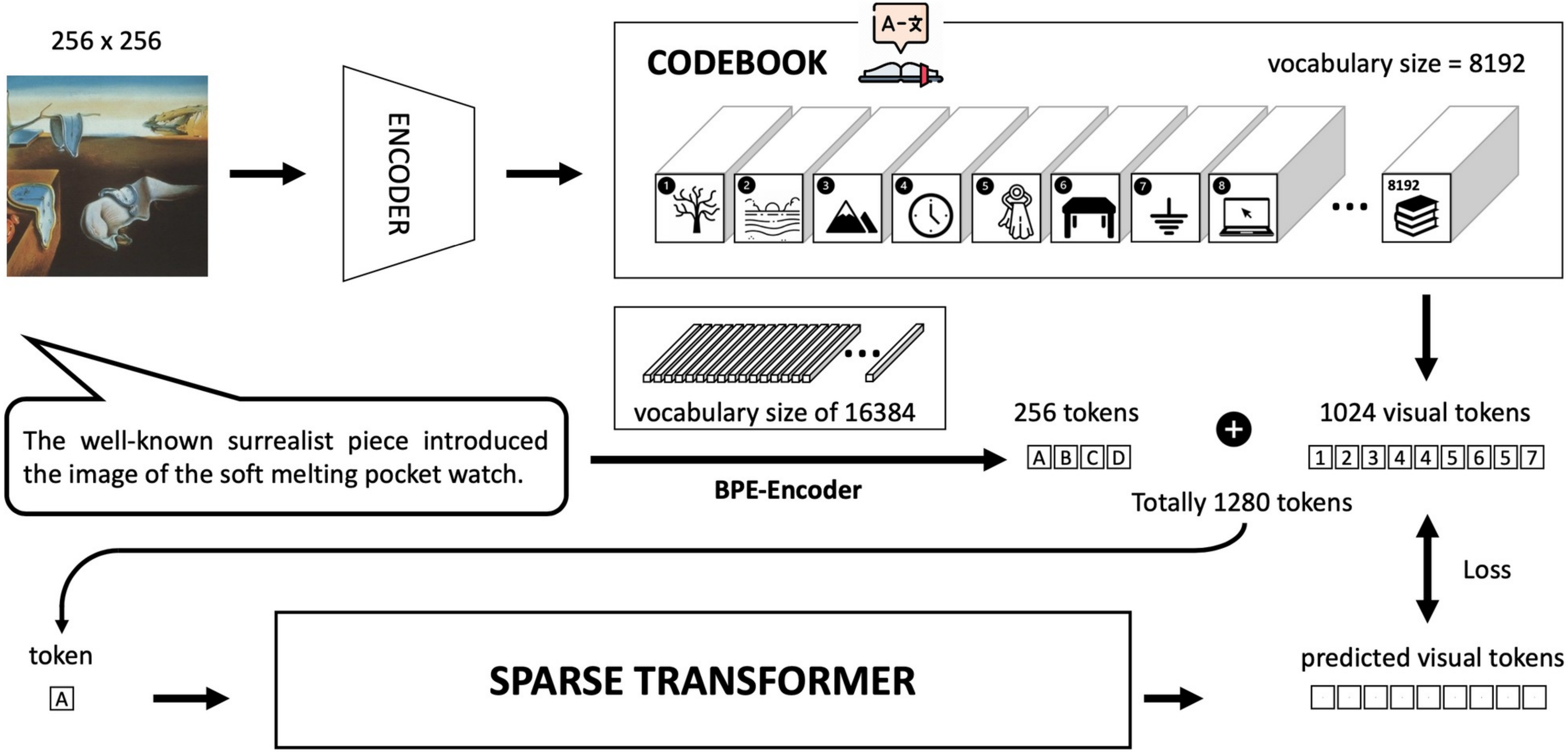
- Stage 1: Learning the Visual Codebook



Ramesh et al., Zero-Shot Text-to-Image Generation, 2021

DALLE

- Stage 2: Learning the Prior



Ramesh et al., Zero-Shot Text-to-Image Generation, 2021

GLIDE

- Diffusion Models
  - Markov chain of latent variables by progressively adding Gaussian noise to samples

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I})$$

  - Learn a model to approximate the true posterior

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$$

  - The model is trained to predict the added noise

$$L_{\text{simple}} := E_{t\sim[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(0,\mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

- Guided Diffusion

$$\hat{\mu}_\theta(x_t|y) = \mu_\theta(x_t|y) + s \cdot \Sigma_\theta(x_t|y)\nabla_{x_t} \log p_\phi(y|x_t)$$

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2022

GLIDE

- Classifier-free guidance

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset))$$

- CLIP Guidance

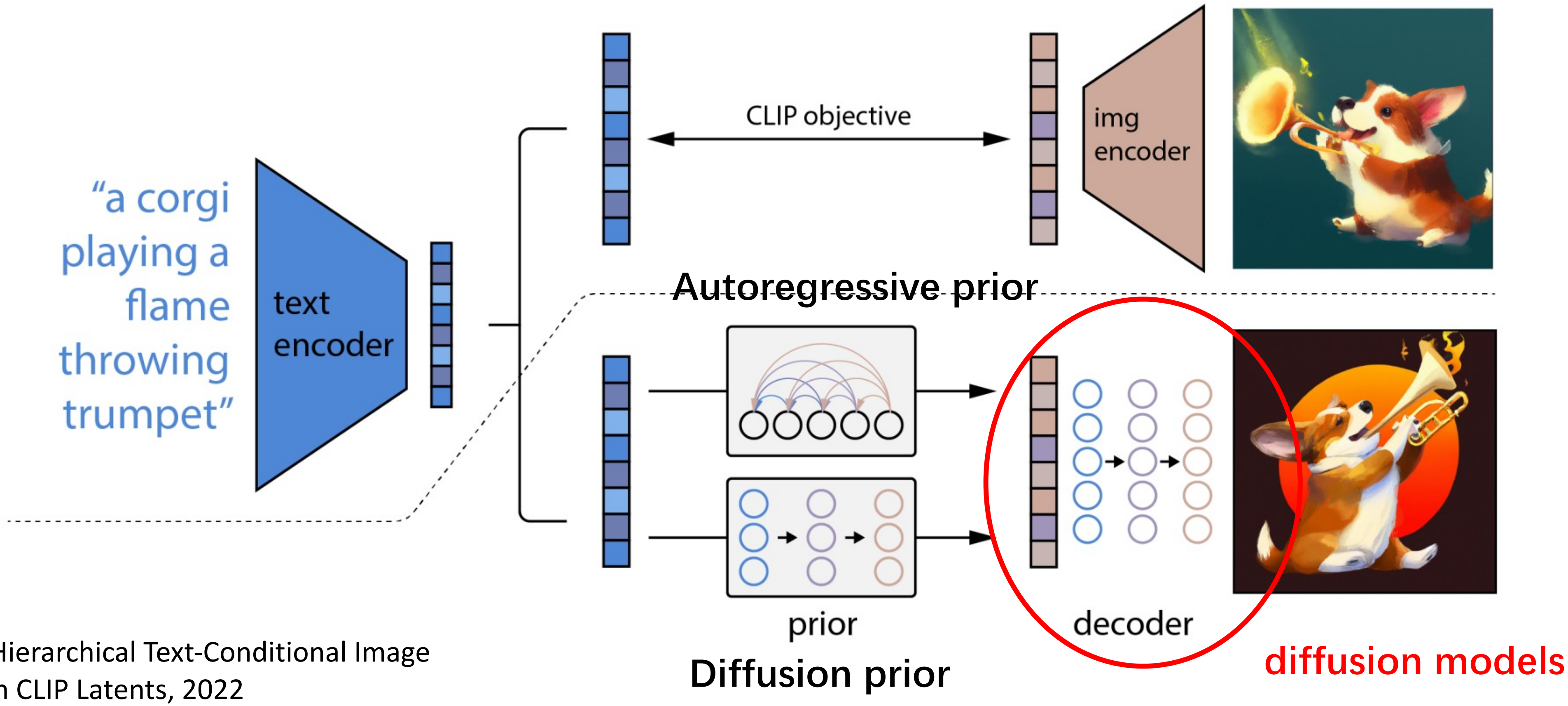
$$\hat{\mu}_\theta(x_t|c) = \mu_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c)\nabla_{x_t}\left(f(x_t) \cdot g(c)\right)$$

- Conclusion: Classifier-free guidance is preferred by human evaluators for both photorealism and caption similarity

Nichol et al., GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, 2022

DALLE2

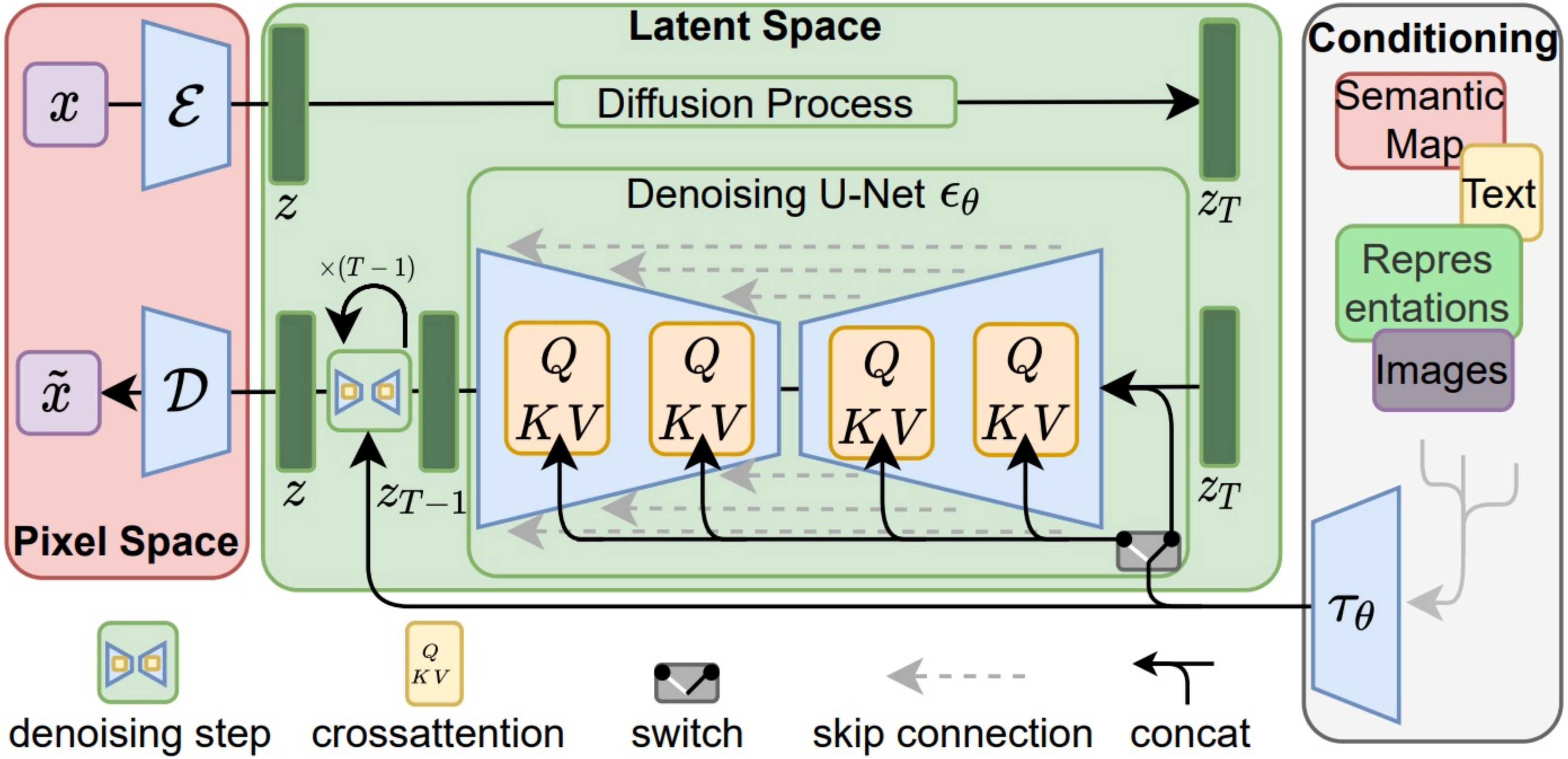
- Two key components:
  - Prior: produces CLIP Image Embeddings conditioned on captions
  - Decoder: produces images conditioned on CLIP Image Embeddings



Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents, 2022

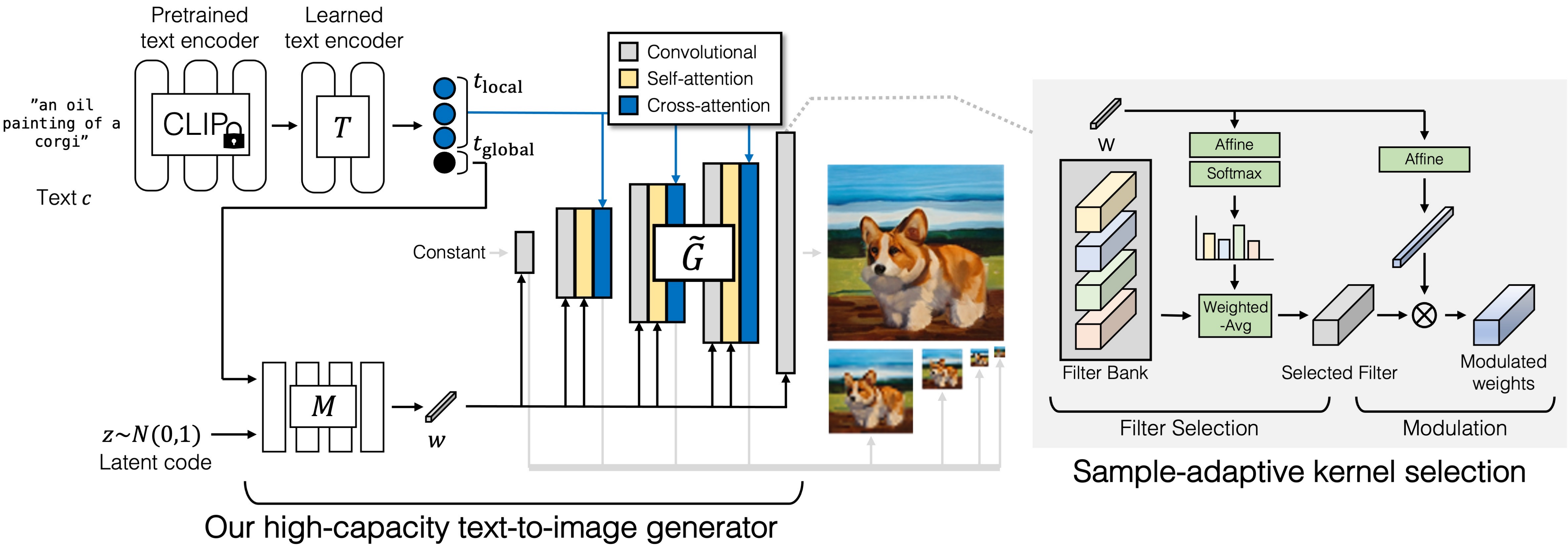
Stable Diffusion

- Encode the images into the latent space



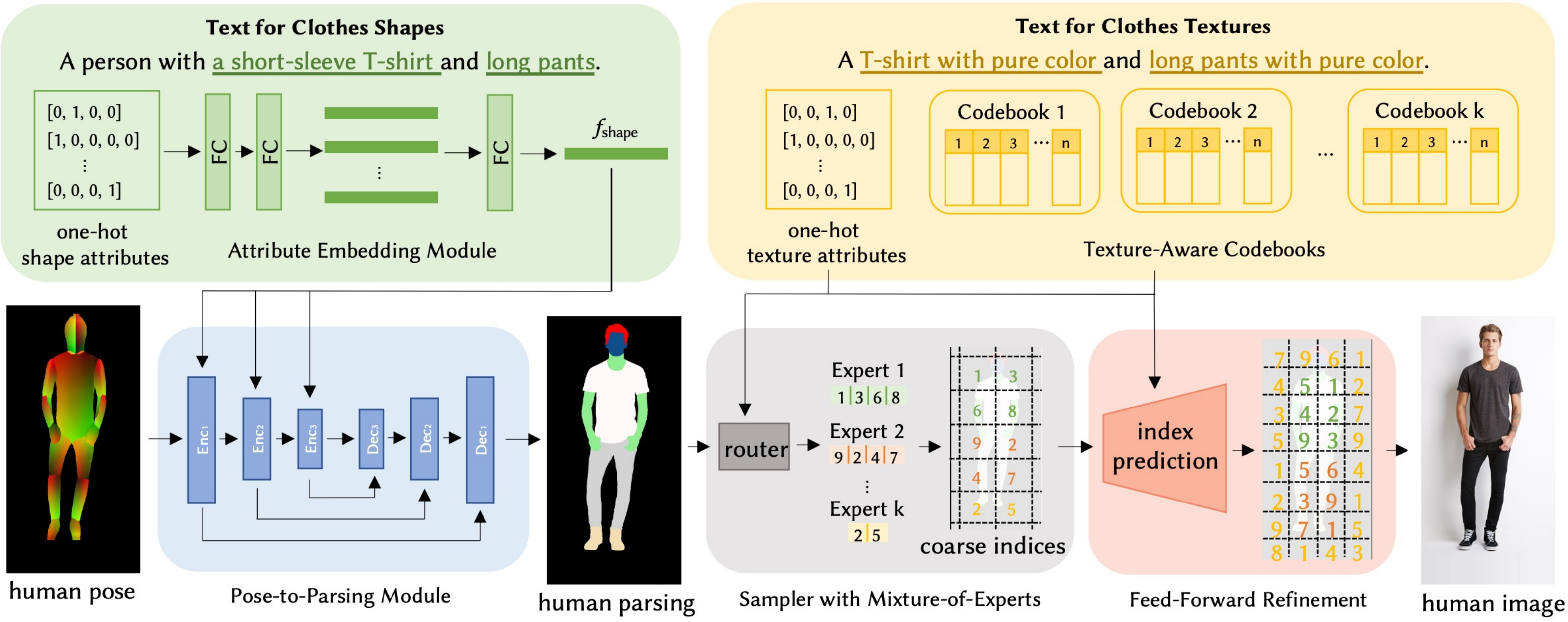
Rombach et al., High-Resolution Image Synthesis with Latent Diffusion Models, 2022

GigaGAN



Kang et al., Scaling up GANs for Text-to-Image Synthesis, 2023

Text2Human



Text for Clothes Shapes
A person with a short-sleeve T-shirt and long pants.
[0, 1, 0, 0]
[1, 0, 0, 0, 0]
[0, 0, 0, 1]
one-hot shape attributes
FC
FC
FC
f_{shape}
Attribute Embedding Module
Text for Clothes Textures
A T-shirt with pure color and long pants with pure color.
[0, 0, 1, 0]
[1, 0, 0, 0, 0]
[0, 0, 0, 1]
one-hot texture attributes
Codebook 1
Codebook 2
Codebook k
1 2 3 ··· n
Texture-Aware Codebooks
Enc1 Enc2 Enc3 Dec3 Dec2 Dec1
human pose
Pose-to-Parsing Module
human parsing
router
Expert 1
1 3 6 8
Expert 2
9 2 4 7
Expert k
2 5
coarse indices
Sampler with Mixture-of-Experts
index prediction
Feed-Forward Refinement
human image

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Image Prompt

- Prompting for Appearance Generation
  - Optimization-Based
    - Textual Inversion
    - DreamBooth
  - Encoder-Based
    - Tuning Encoder
    - ELITE
    - Taming Encoder
- Prompting for Relation Generation
  - ReVersion

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Image Prompt

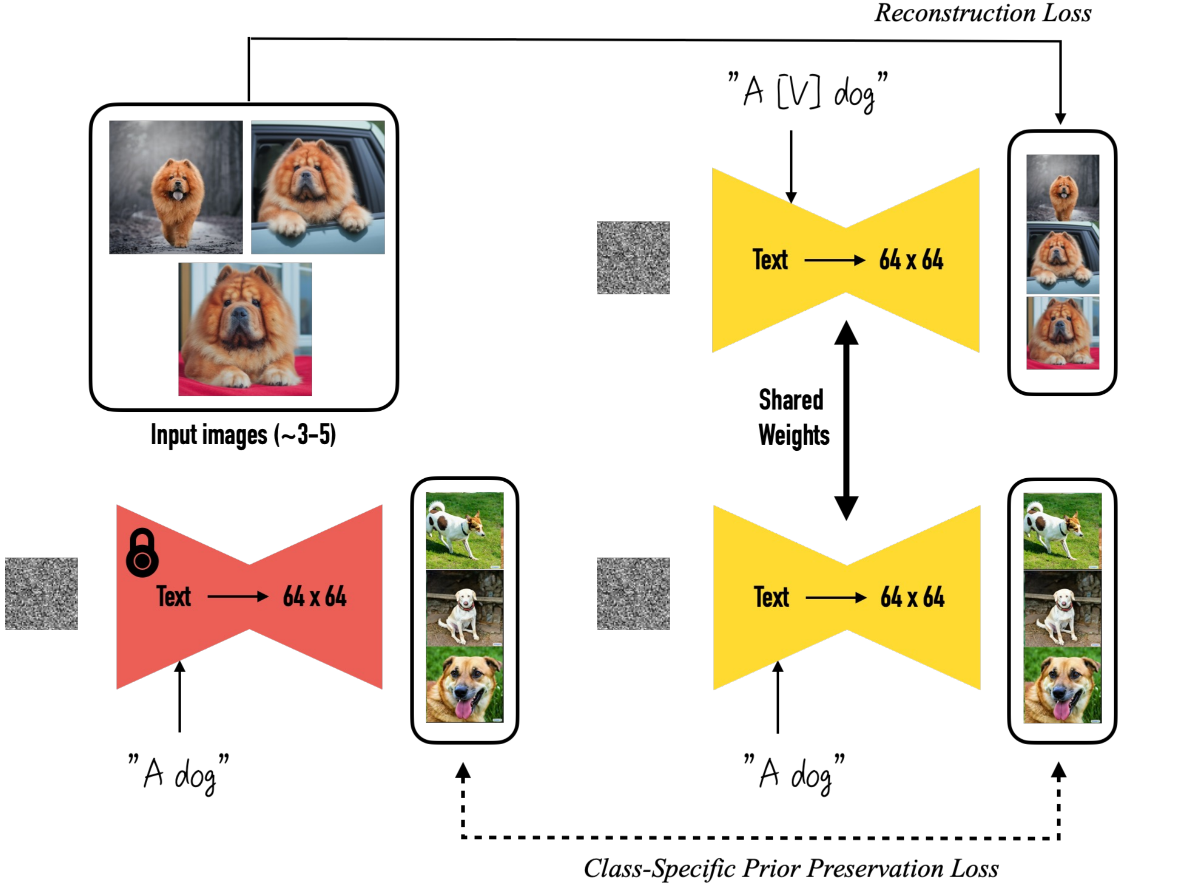
- Prompting for Appearance Generation
  - Optimization-Based
    - Textual Inversion
    - DreamBooth
  - Encoder-Based
    - Tuning Encoder
    - ELITE
    - Taming Encoder
- Prompting for Relation Generation
  - ReVersion

Textual Inversion

- Task: prompting for appearance generation (personalized generation)
- Method: optimize a text token: $v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (ICLR 2023)

DreamBooth



- Task: prompting for appearance generation (personalized generation)
- Method: fine-tune to obtain a personalized text-to-image model

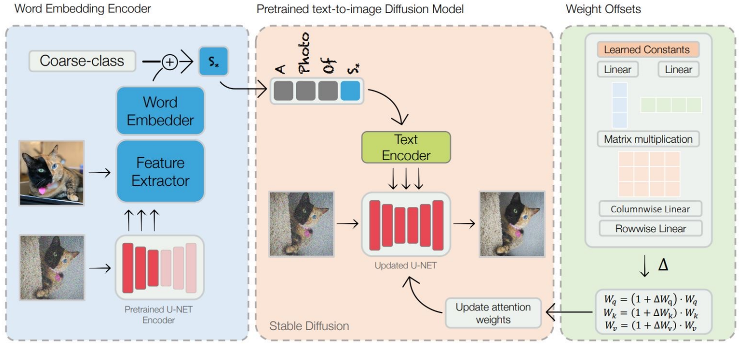
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (CVPR 2023)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

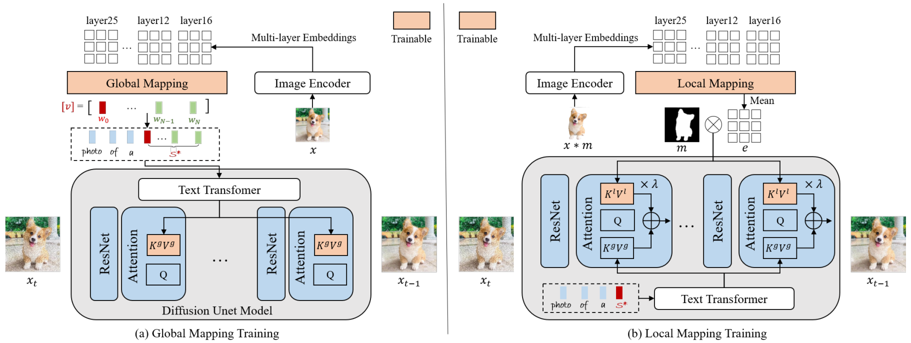
Image Prompt

- Prompting for Appearance Generation
  - Optimization-Based
    - Textual Inversion
    - DreamBooth
  - Encoder-Based
    - Tuning Encoder
    - ELITE
    - Taming Encoder
- Prompting for Relation Generation
  - ReVersion

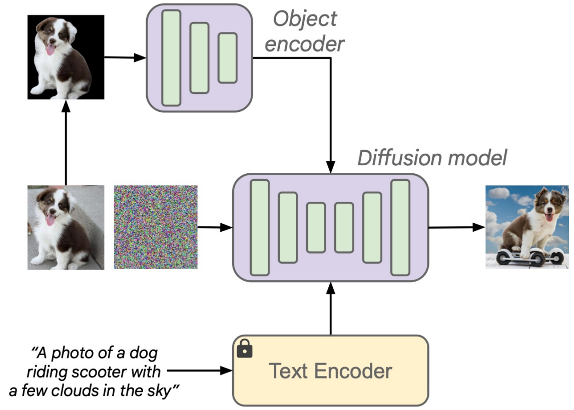
Encoder-Based



Tuning Encoder



ELITE

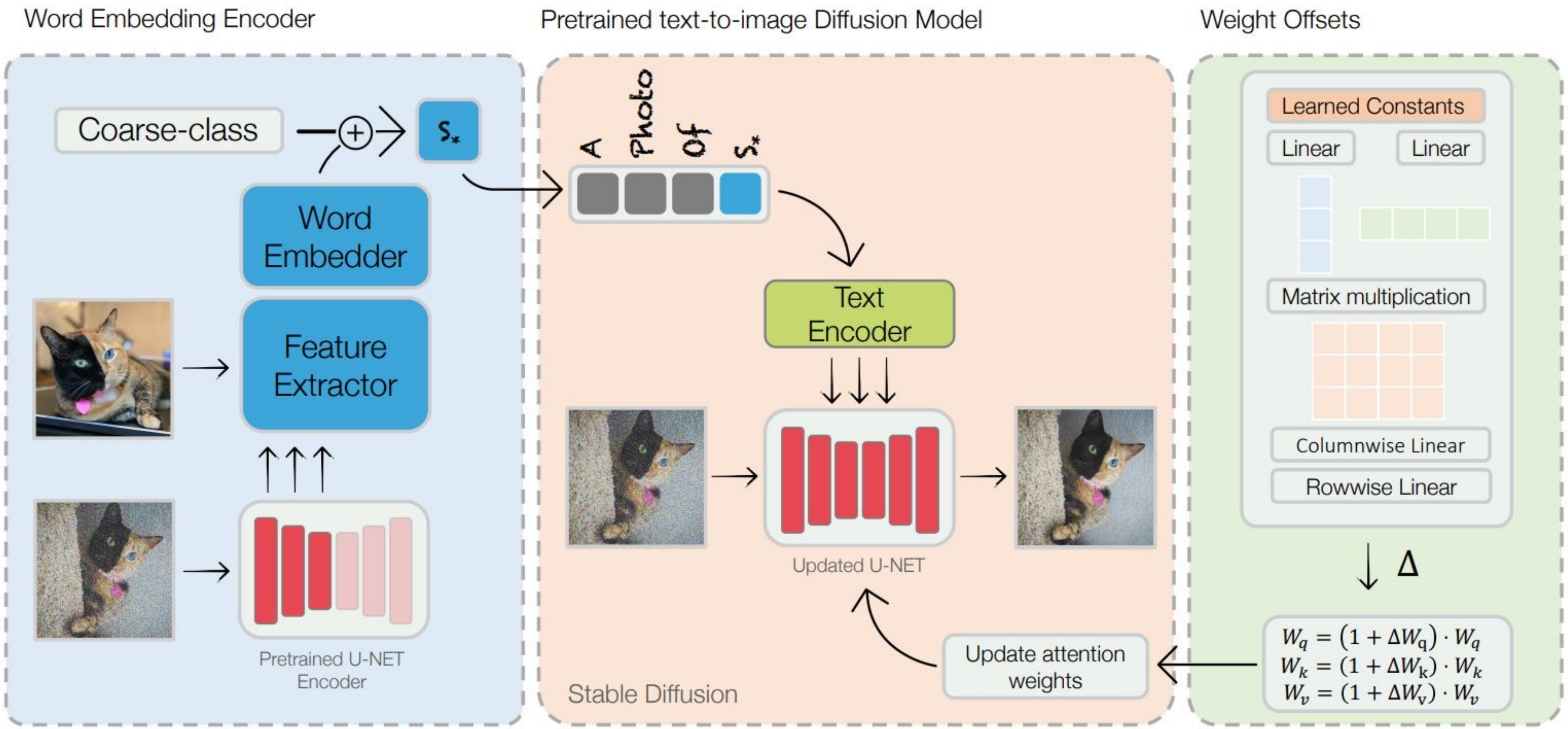


Taming Encoder

- Fast: a few optimization steps
- Memory Efficient
- One-Shot

Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models (2023)
ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation (2023)
Taming encoder for zero fine-tuning image customization with text-to-image diffusion models (2023)

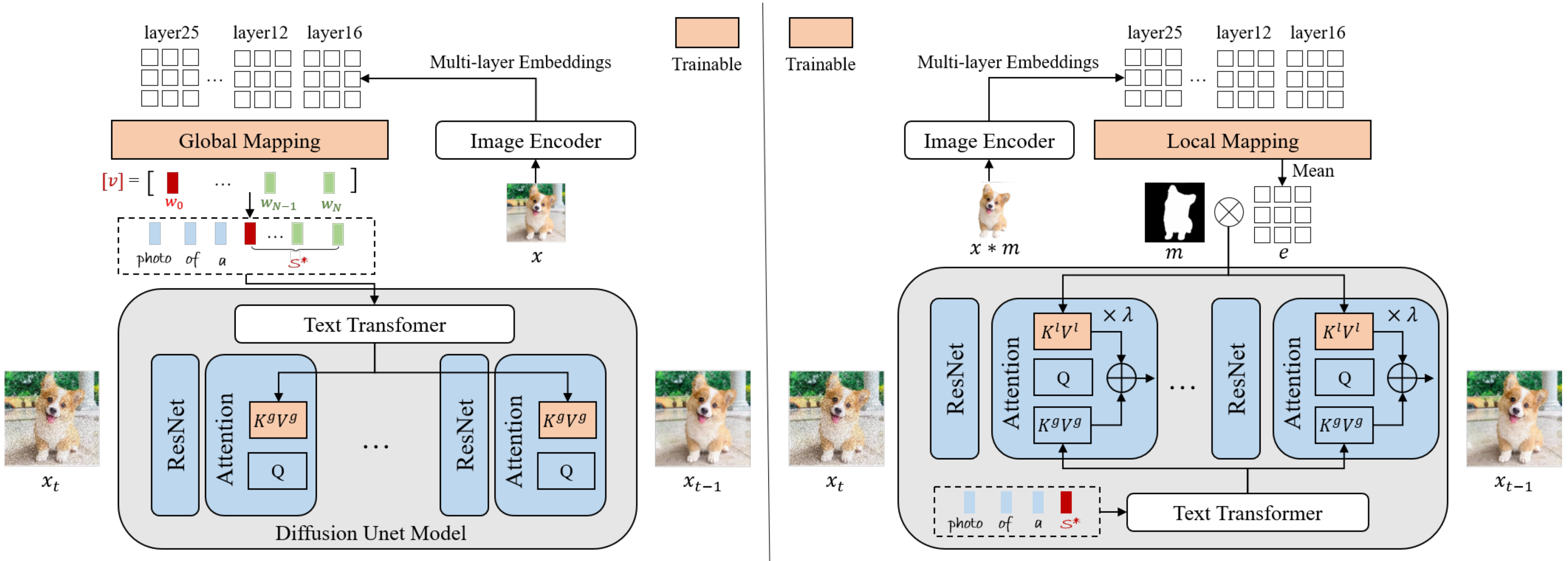
Tuning Encoder



- Domain-Specific Encoder
- Weight Offsets

Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models (2023)

ELITE

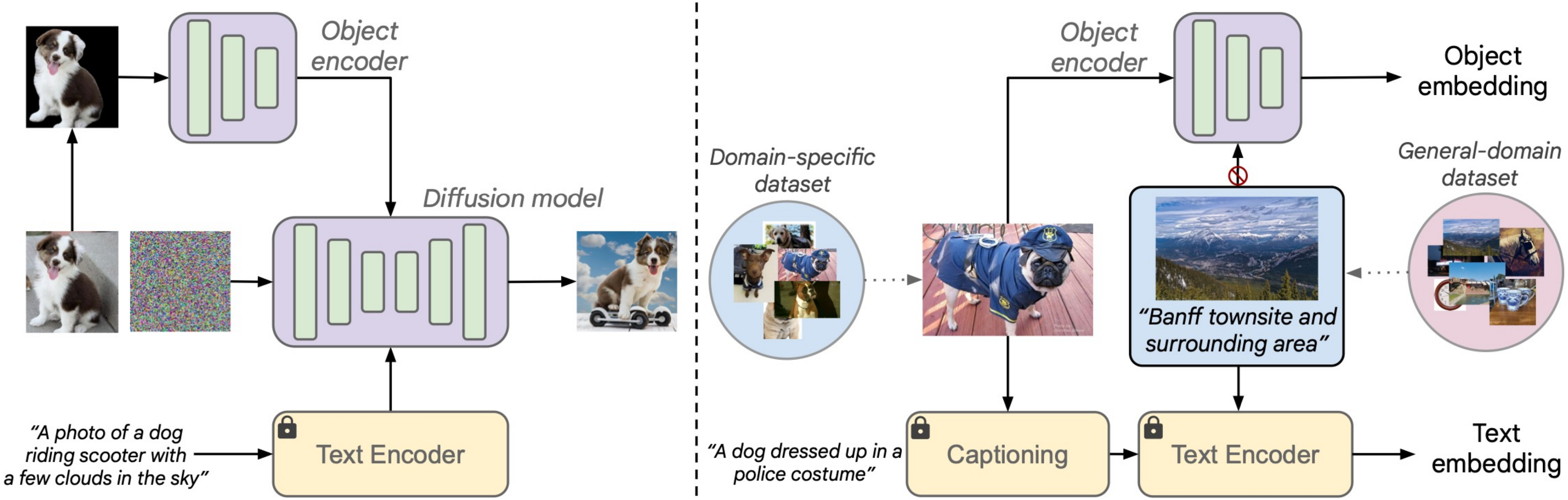


(a) Global Mapping Training
(b) Local Mapping Training

- Global Mapping Network – Text Embeddings
- Local Mapping Network – Details

ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation (2023)

Taming Encoder



- Background Removal + Encoder
- Triplet Preparation Scheme

Taming encoder for zero fine-tuning image customization with text-to-image diffusion models (2023)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Image Prompt

- Prompting for Appearance Generation
  - Optimization-Based
    - Textual Inversion
    - DreamBooth
  - Encoder-Based
    - Tuning Encoder
    - ELITE
    - Taming Encoder
- Prompting for Relation Generation
  - ReVersion

ReVersion

Input

Exemplar Images

*“vegetable **is contained inside** paper bag”*

Output

Relation Prompt

<R>

represent the co-existing relation in exemplar images

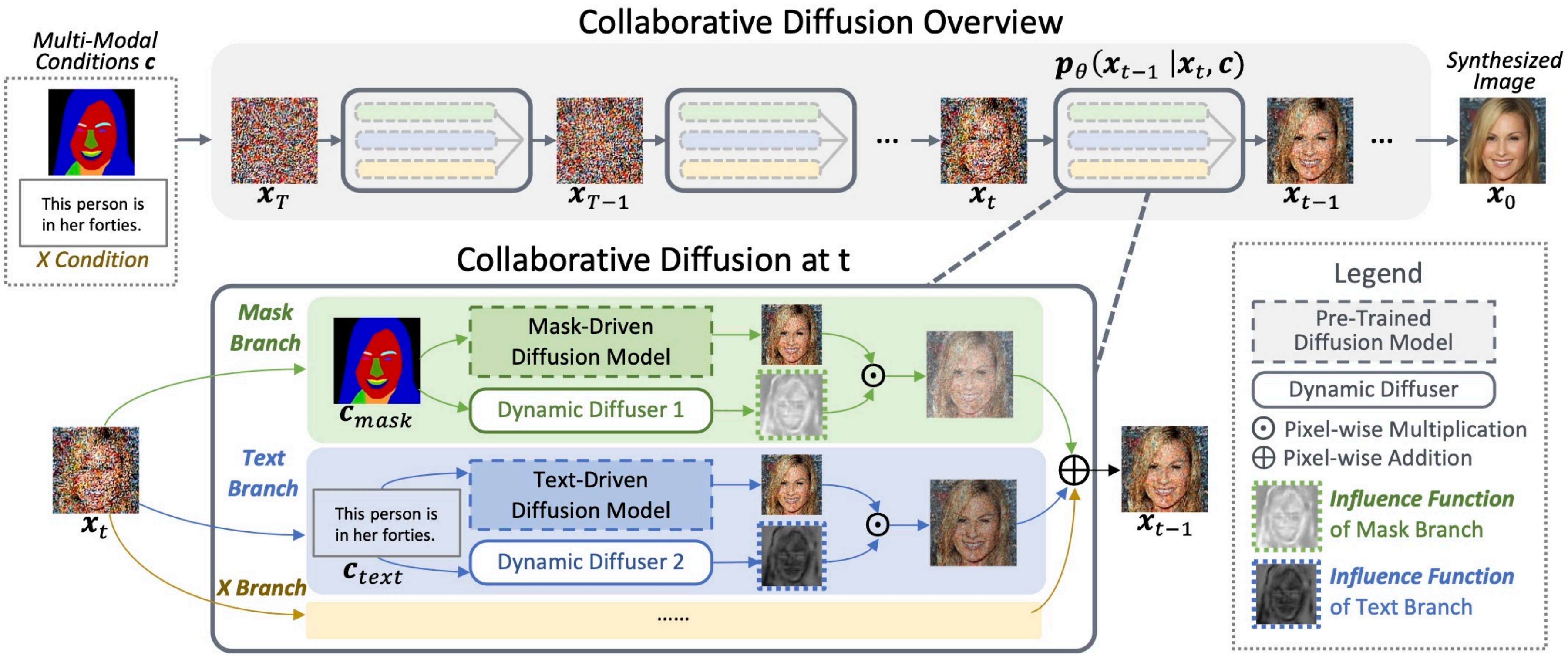
Application

Relation-Specific Text-to-Image Synthesis

*“Spiderman **<R>** paper bag”*

ReVersion: Diffusion-Based Relation Inversion from Images (2023)

Collaborative Diffusion



- Use model collaboration to simultaneously accept different types of prompt: linguistic, visual

Collaborative Diffusion for Multi-Modal Face Generation and Editing (CVPR 2023)

Text to Video Generation

Text to Video Generation

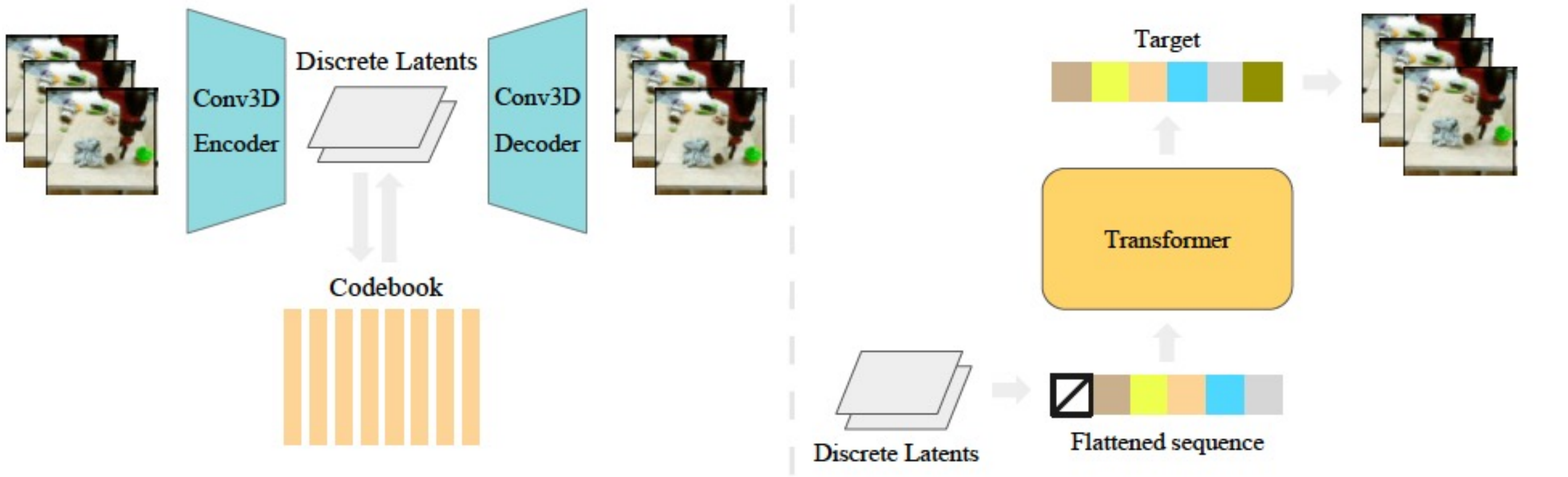
- Auto-regressive methods
  - VideoGPT
  - TATS
  - Phenaki
- Diffusion models
  - Imagen Video
  - Gen1
  - Text2Performer

Text to Video Generation

- Auto-regressive methods
  - VideoGPT
  - TATS
  - Phenaki
- Diffusion models
  - Imagen Video
  - Gen1
  - Text2Performer

T2V: VideoGPT

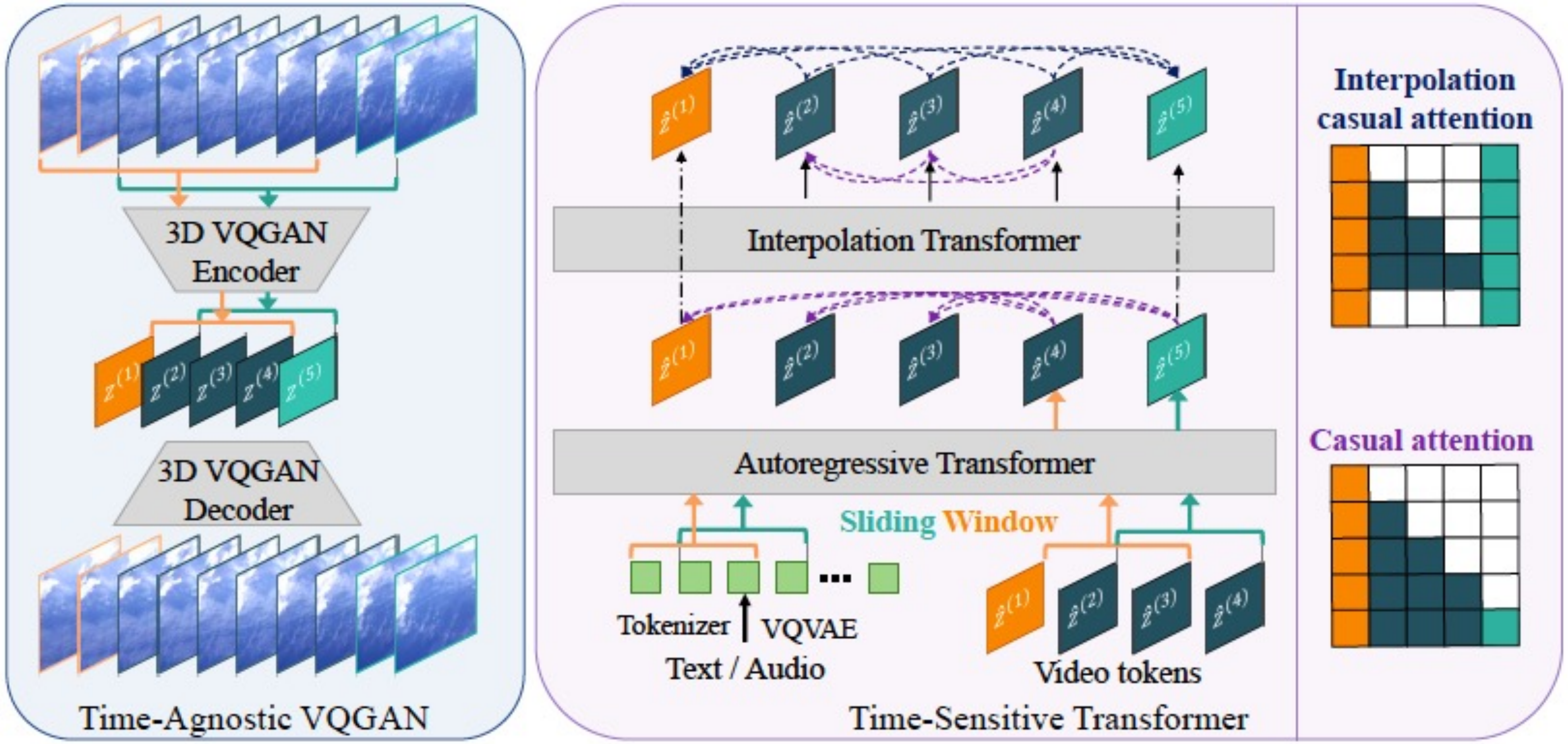
- VQGAN: learn a set of discrete latent codes from raw pixels of the video frames.
- Transformer: learn a prior over the VQ-VAE latent codes.



Yan *et al.* VideoGPT: Video Generation using VQ-VAE and Transformers

T2V: TATS

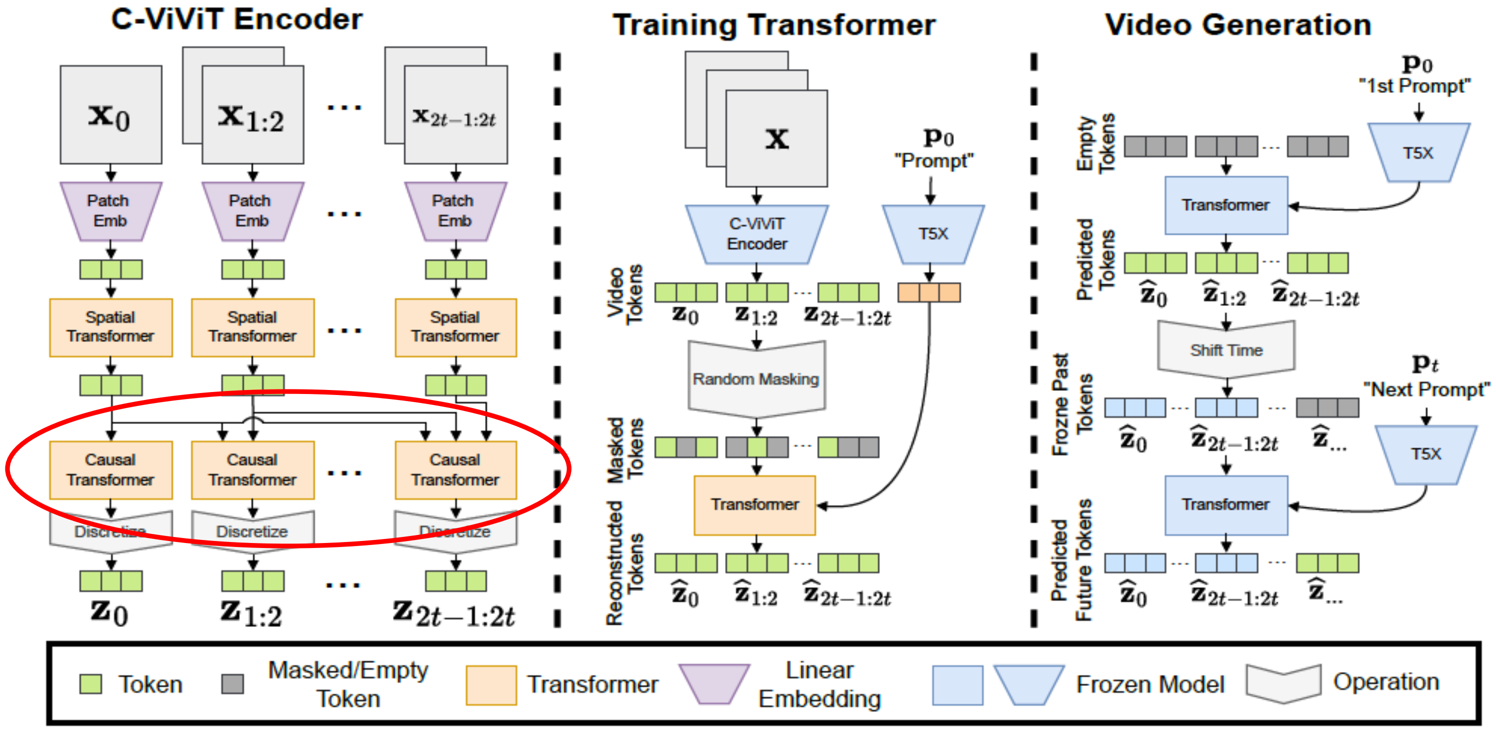
- 3D VQGAN: replacing 2D convolution operations with 3D convolutions for modeling videos.
- Transformer: the hierarchical transformer can model longer time dependence and delay the quality degradation.



Ge *et al.* Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer

T2V: Phenaki

- Encoder-decoder model: compress videos to discrete embeddings.
  - Causal attention makes the C-ViViT encoder autoregressive and enables it to handle a variable number of input frames.
- Transformer model: translate text embeddings to video tokens.



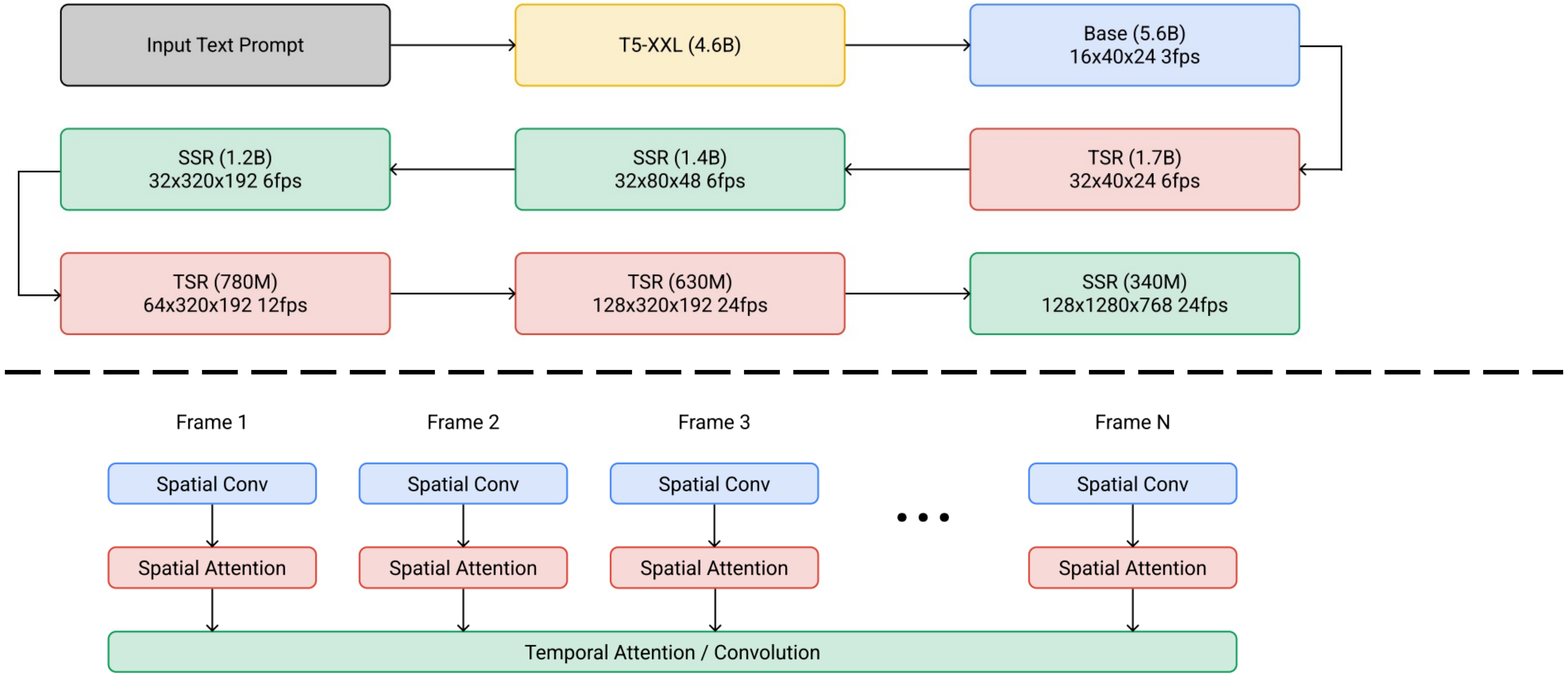
Villegas *et al.* Phenaki: Variable Length Video Generation From Open Domain Textual Description

Text to Video Generation

- Auto-regressive methods
  - VideoGPT
  - TATS
  - Phenaki
- Diffusion models
  - Imagen Video
  - Gen1
  - Text2Performer

T2V: Imagen video

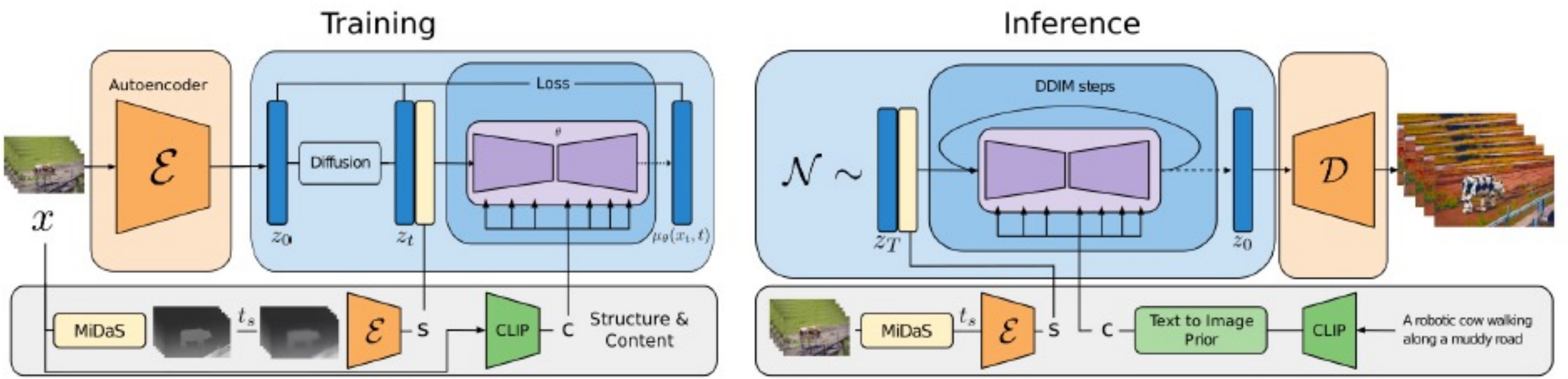
- Cascaded Diffusion Models.
  - 1 frozen text encoder, 1 base video diffusion model, 3 SSR (spatial super-resolution), and 3 TSR (temporal superresolution) models – for a total of 7 video diffusion models



Ho *et al.* IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS

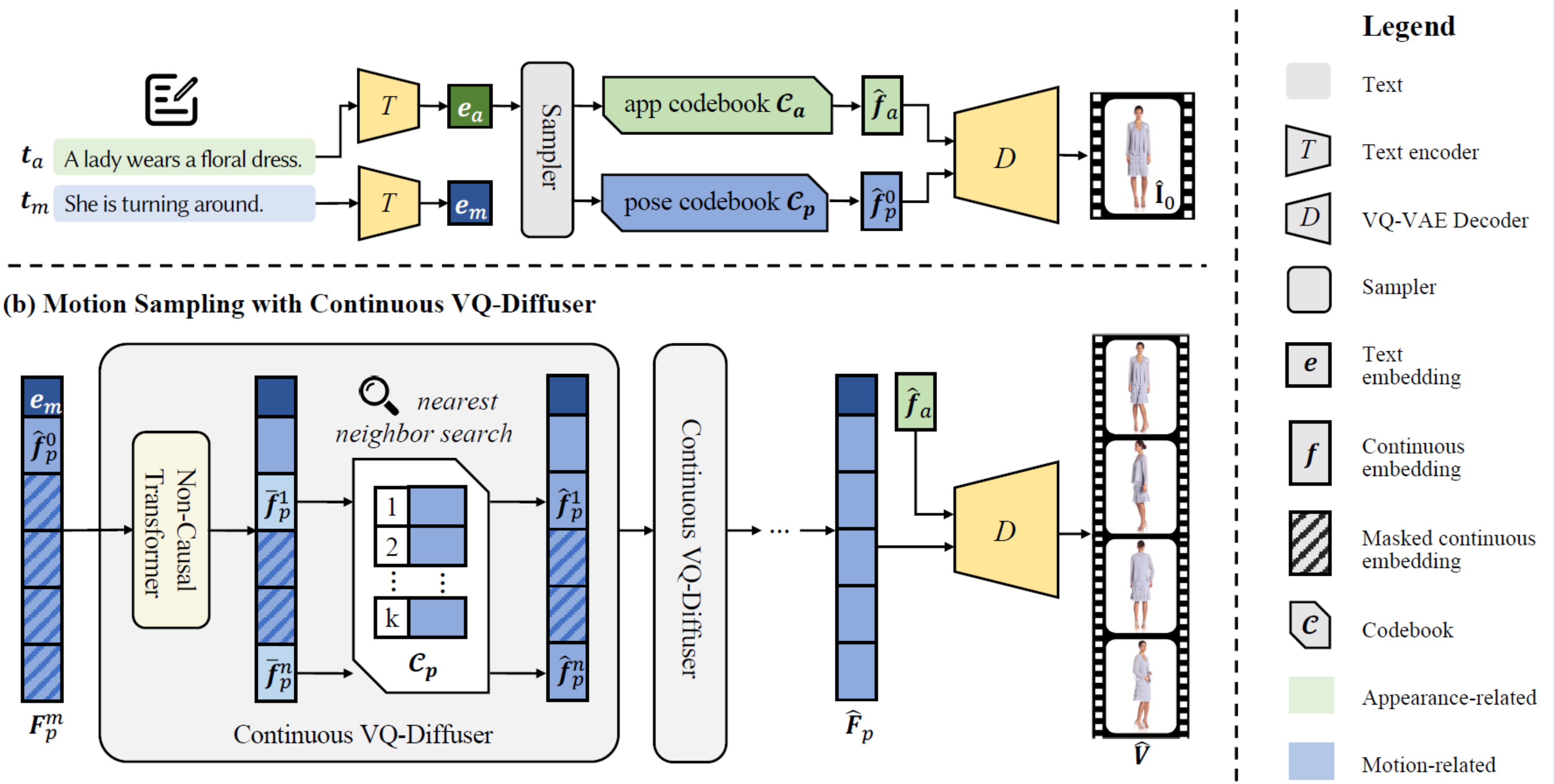
T2V: Gen1

- Diffusion model: introduce temporal layers into a pre-trained image latent diffusion model
- Structure representation: utilize depth maps to provide control over structure and content fidelity.
- Content Representation: utilize CLIP to produce image (training) or text (inference) embeddings.



Esser *et al.* Structure and Content-Guided Video Synthesis with Diffusion Models

T2V: Text2Performer

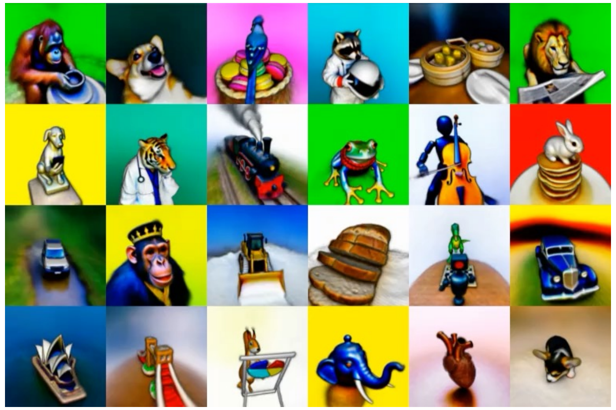
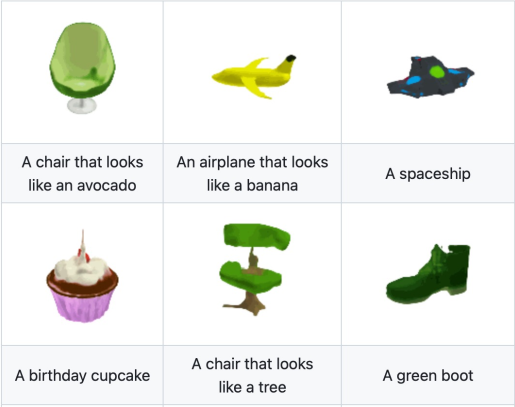
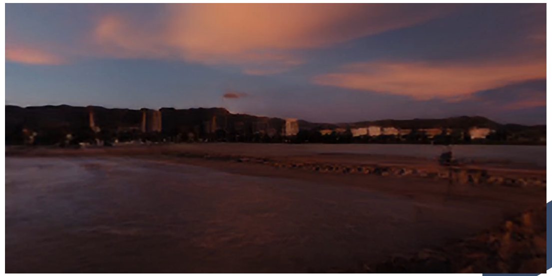


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

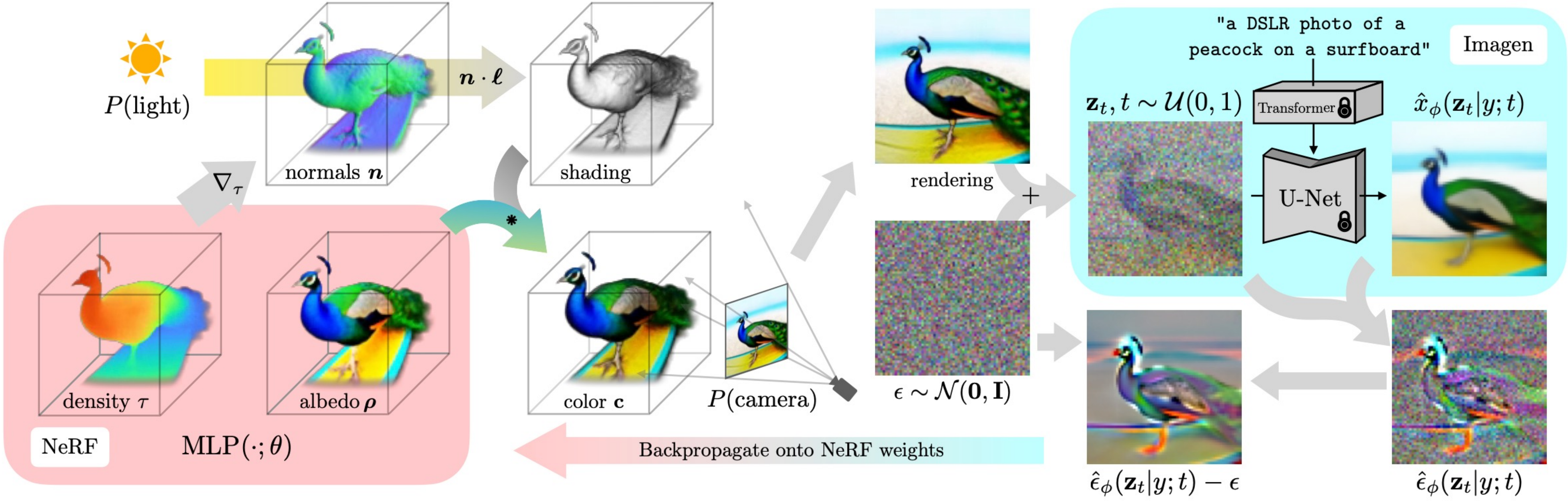
S-LAB FOR ADVANCED INTELLIGENCE

Text to 3D Generation

Overview

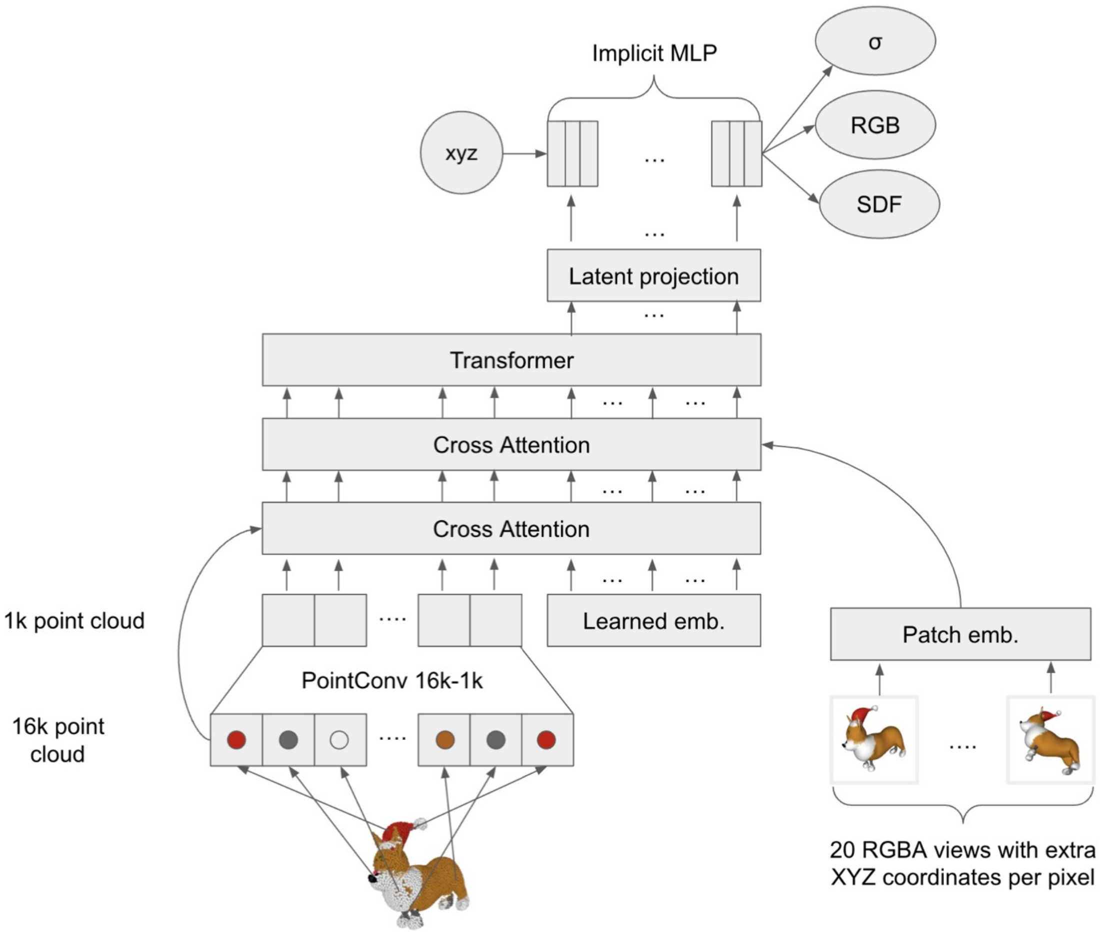
	Object	**Human**	**Scene**
Leveraging 2D Prior from pretrained text-2D models	DreamFusion 	AvatarCLIP 	Text2Room 
Supervised Training from text-3D paired data	Shap-E 	Rodin 	Text2Light 

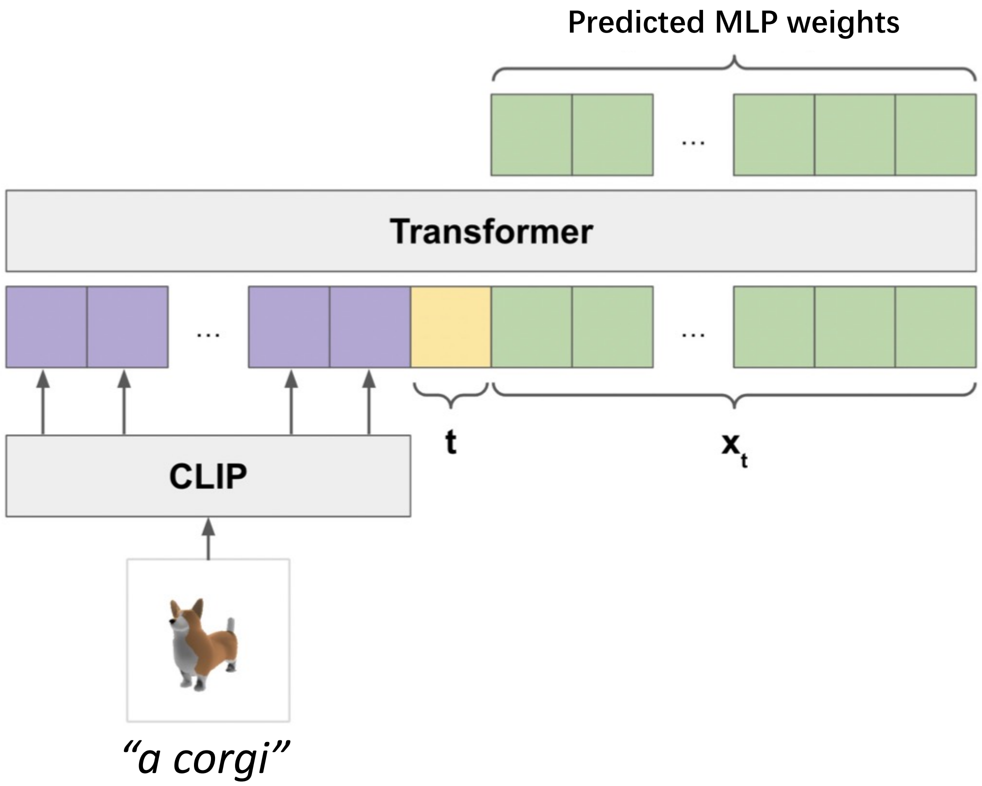
DreamFusion



DreamFusion: Text-to-3D Using 2D Diffusion

Shap-E





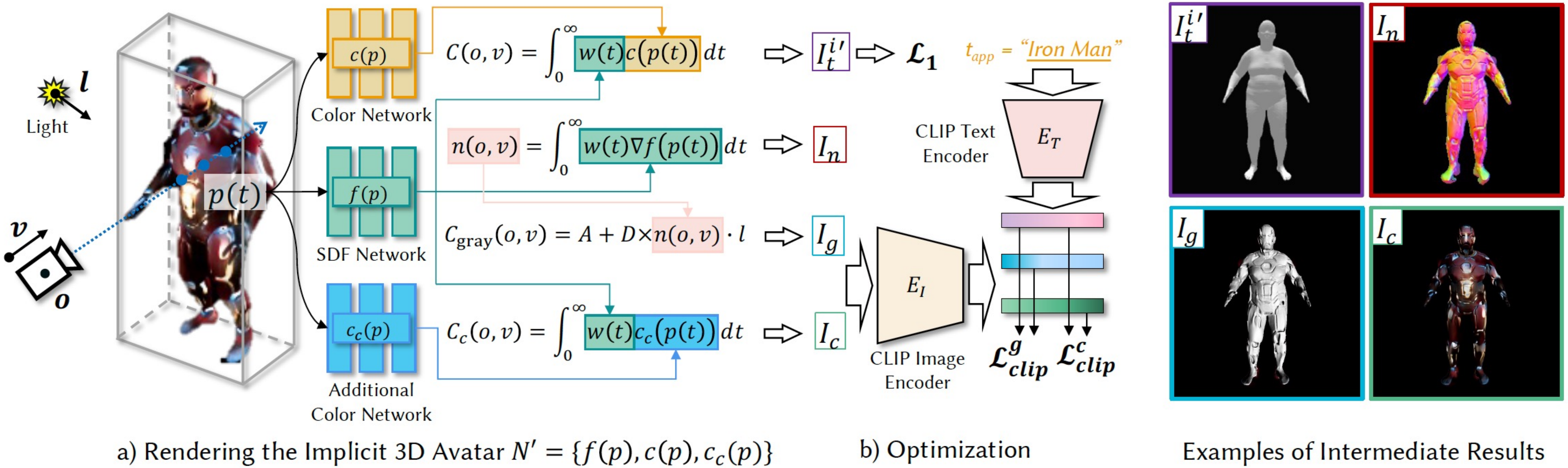
Step 1: Encode 3D Objects into Latent Space

Step 2: Latent Diffusion

Shap·E: Generating Conditional 3D Implicit Functions

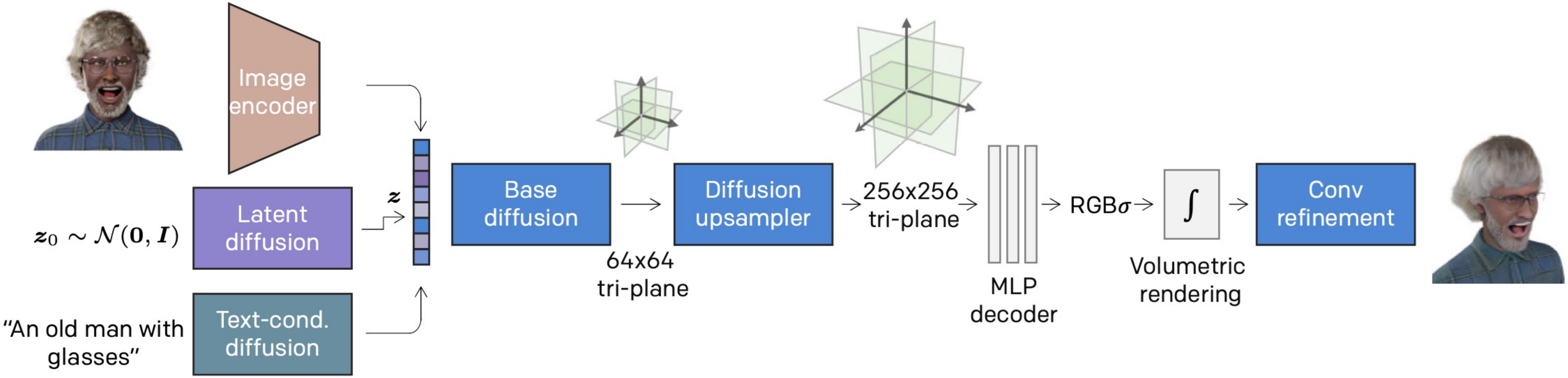
AvatarCLIP

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
S-LAB FOR ADVANCED INTELLIGENCE



AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars

Rodin

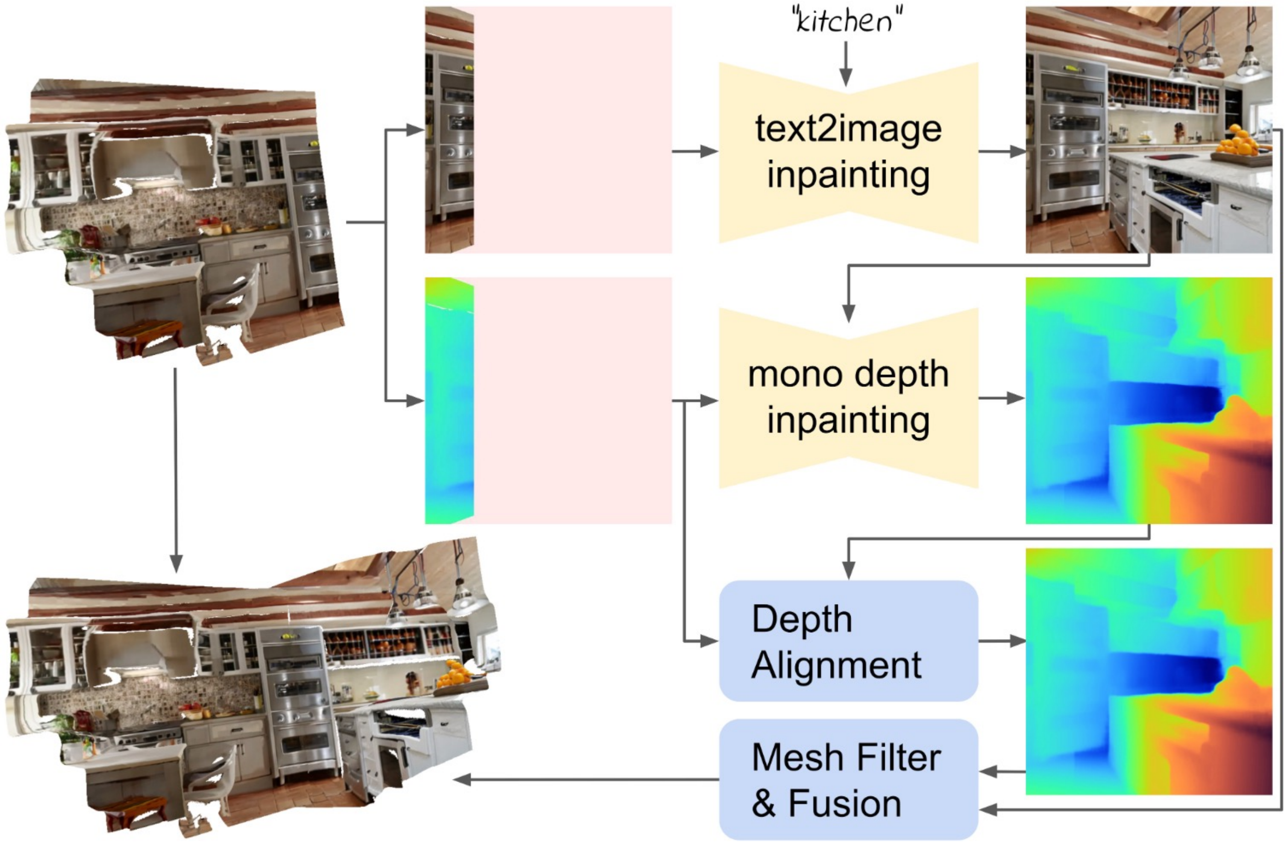


Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion

Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models

“Editorial Style Photo, Rustic Farmhouse, Living Room, Stone Fireplace, Wood, Leather, Wool”

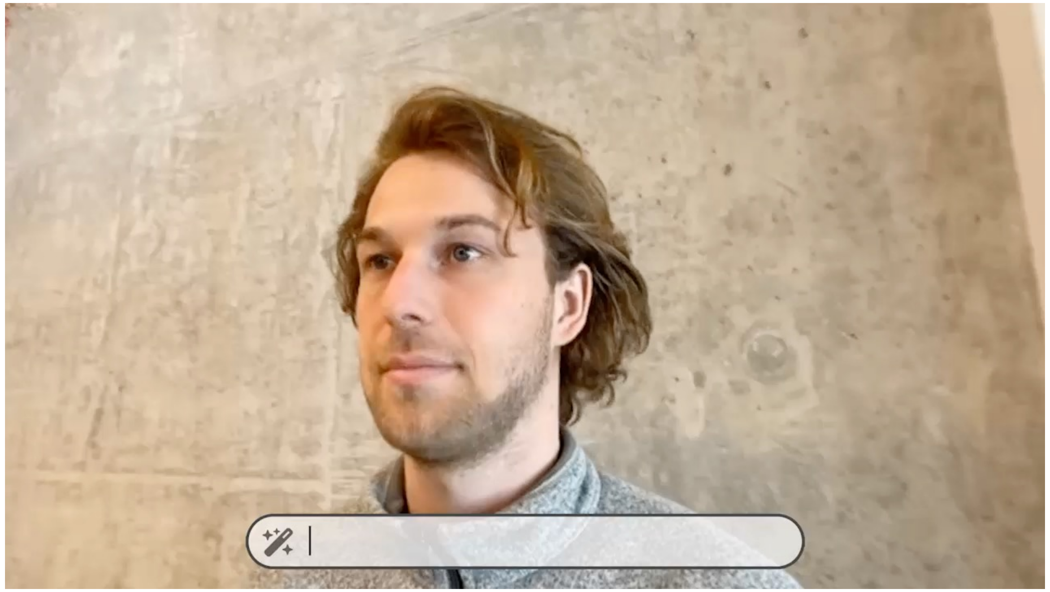
“A living room with a lit furnace, couch, and cozy curtains, bright lamps that make the room look well-lit.”



Text Prompts -> 3D Scenes

Optimization based

Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions



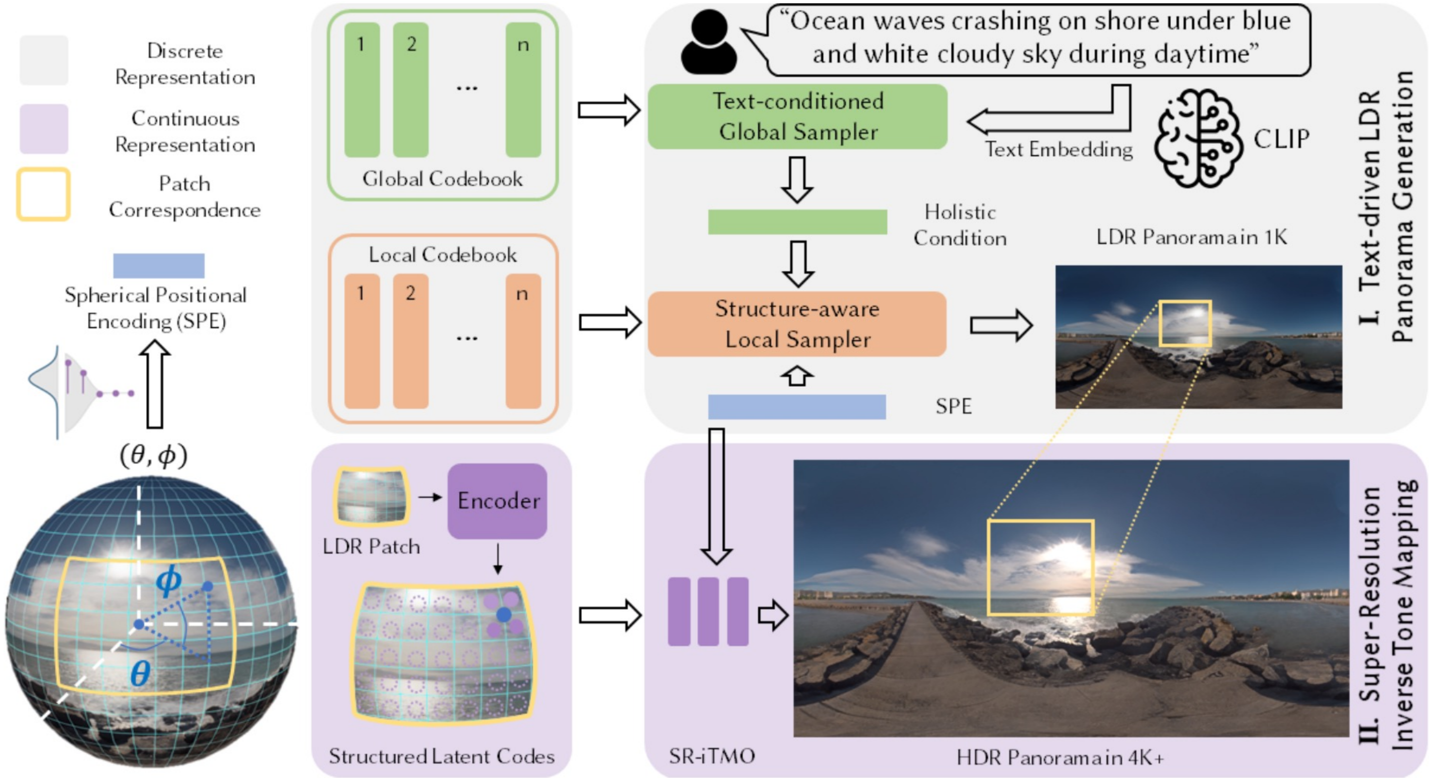
Edit 3D Scenes via Instructions

Text Prompts + Instruction Tuning -> 3D Scenes

Optimization based

Text2Light: Zero-shot Text-driven HDR Panorama Generation

“Sunset by the Ocean”



Text Prompts -> Panoramic 3D Scenes
Feed Forward Generation

Future work

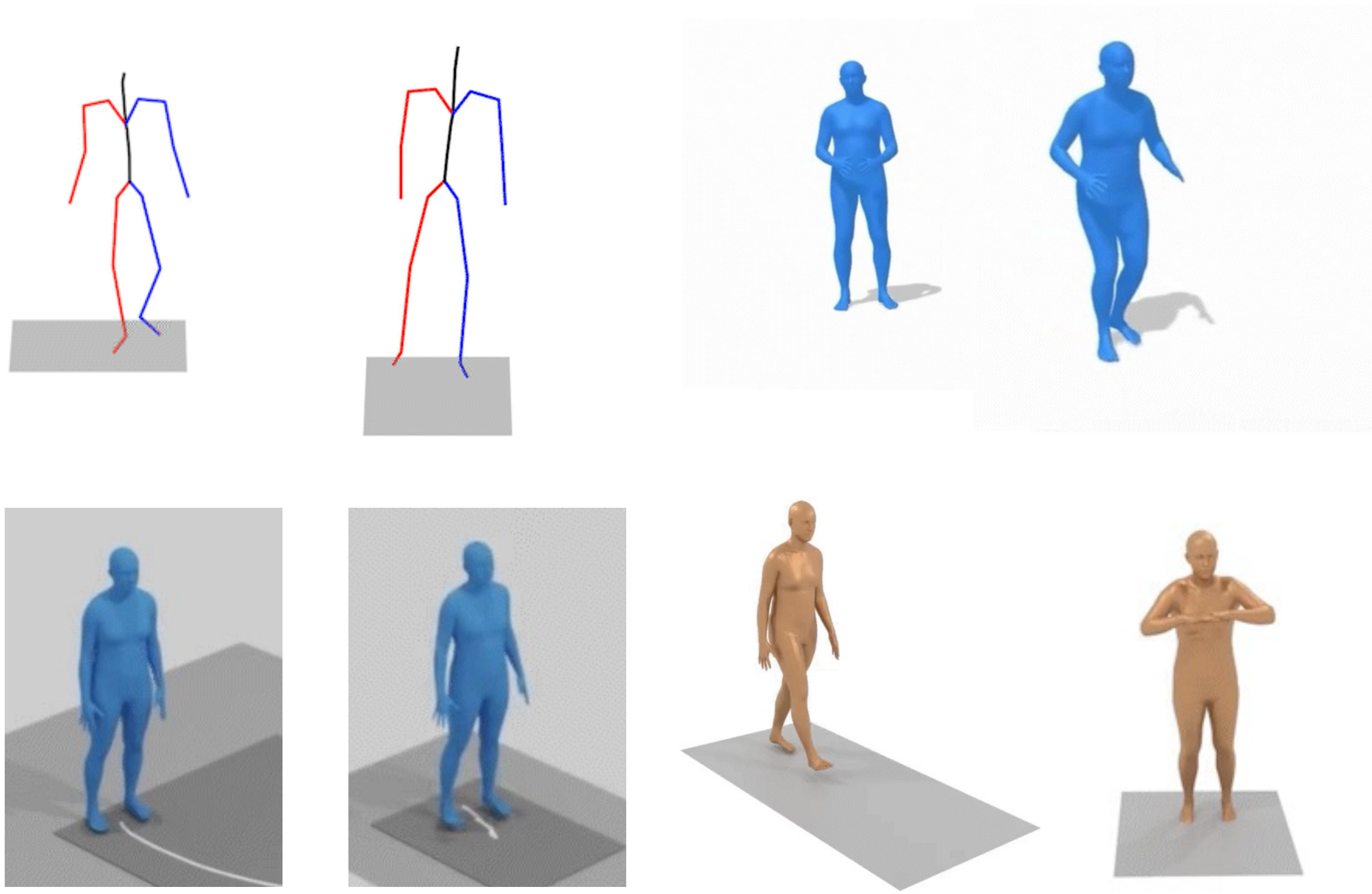
- Faster Generation:
  - Per-scene-optimization is time consuming.
- Higher Quality:
  - The resolution is limited by the resolution of 2D model.
  - Super high guidance weight leads to over-saturation, over-smoothing results.
- More Efficient 3D Representation
  - Directly learning from 3D data is expensive.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

S-LAB FOR ADVANCED INTELLIGENCE

Text to 4D Generation

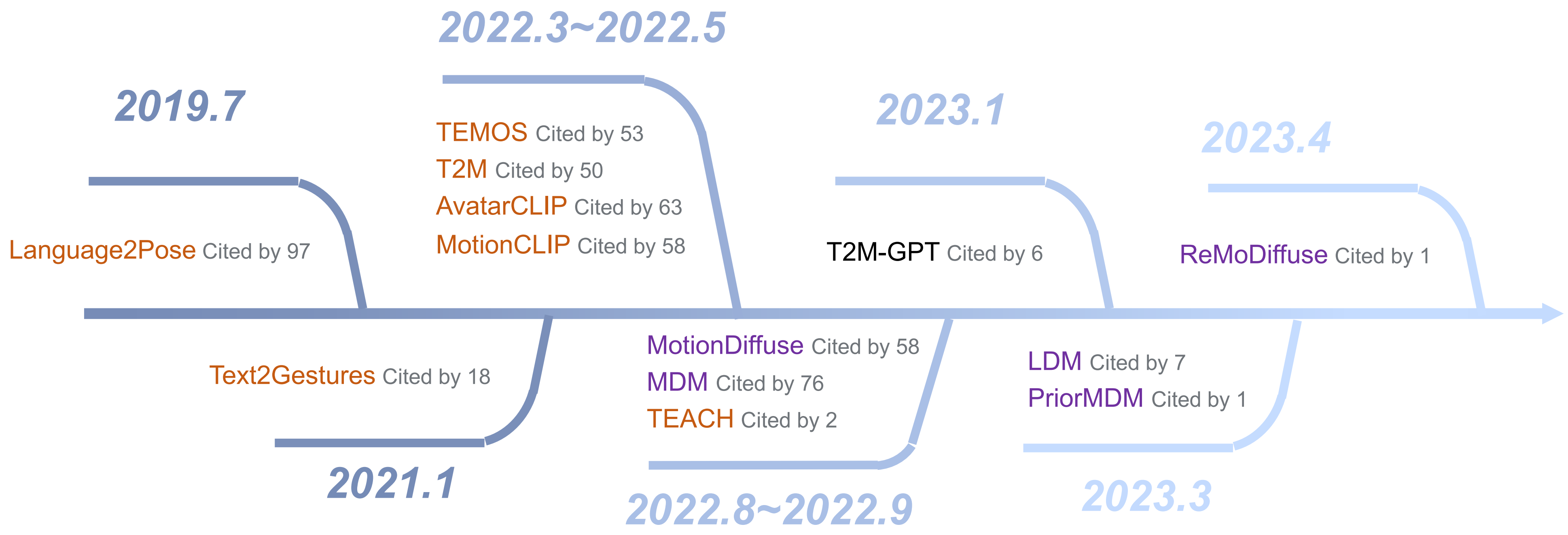
Text-to-4D Generation

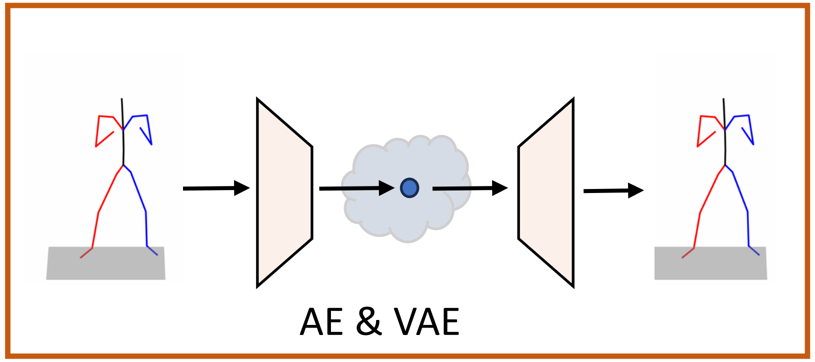


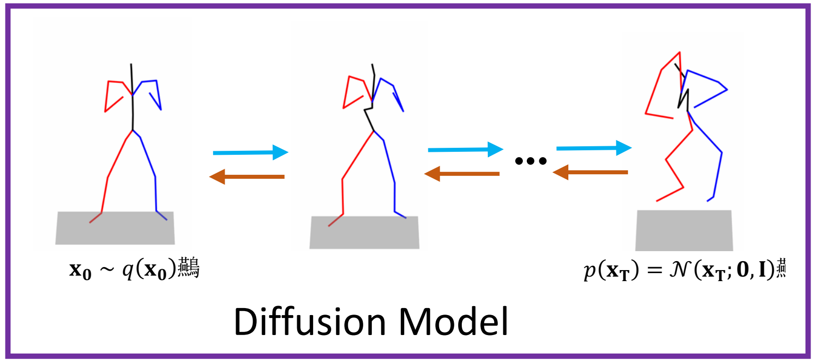
Motion generation

4D scene generation

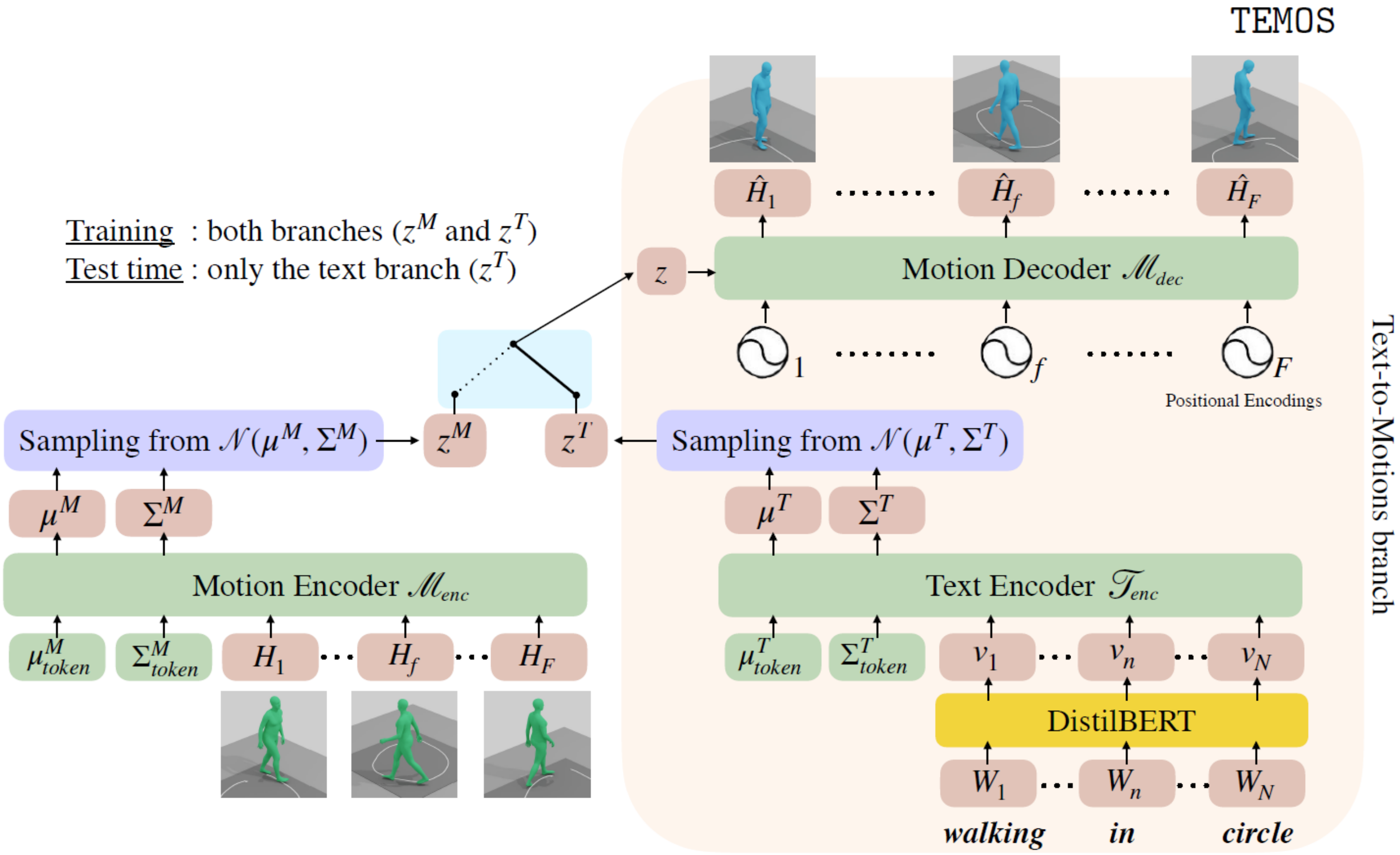
Human Motion Generation







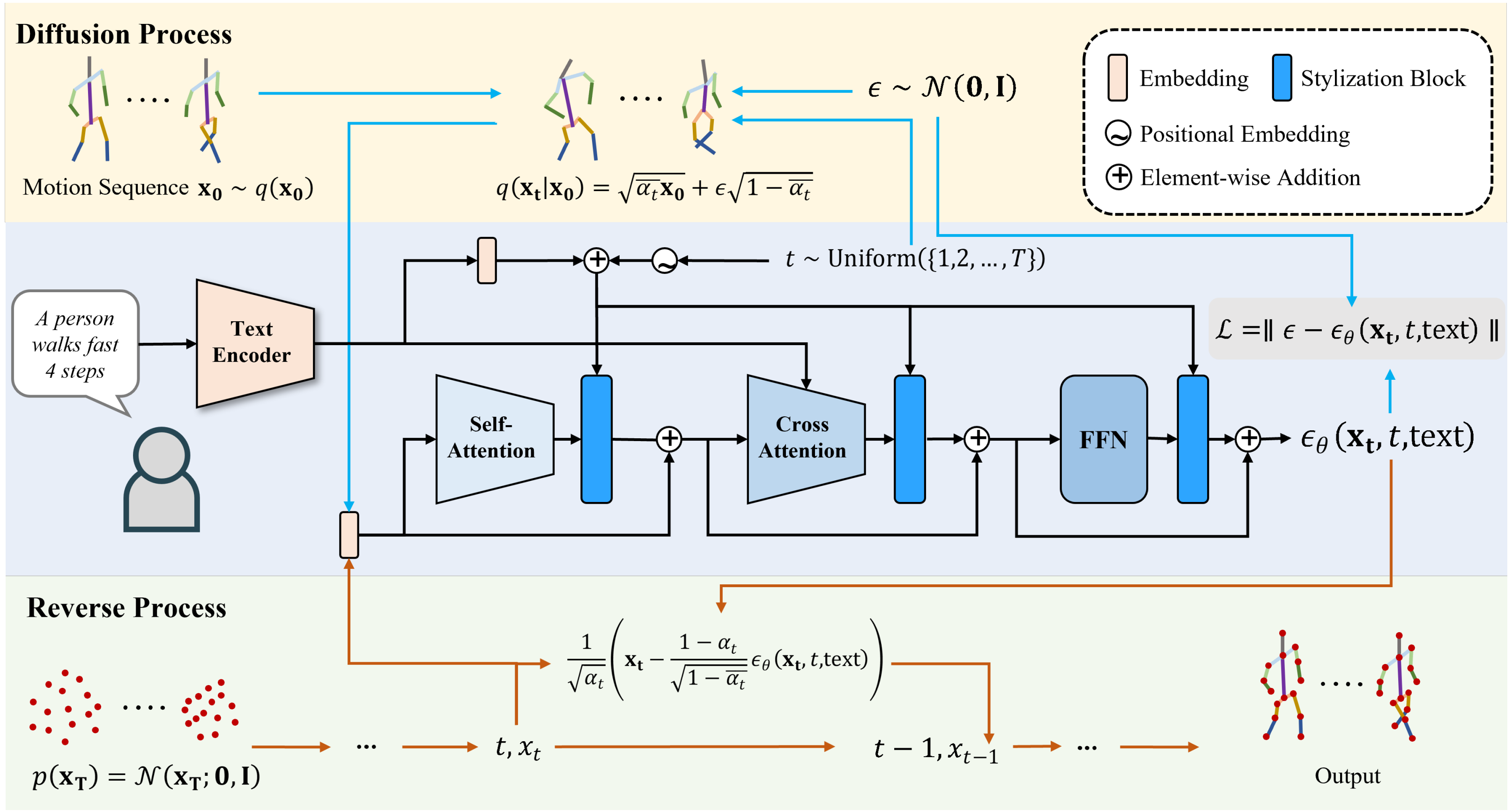
TEMOS

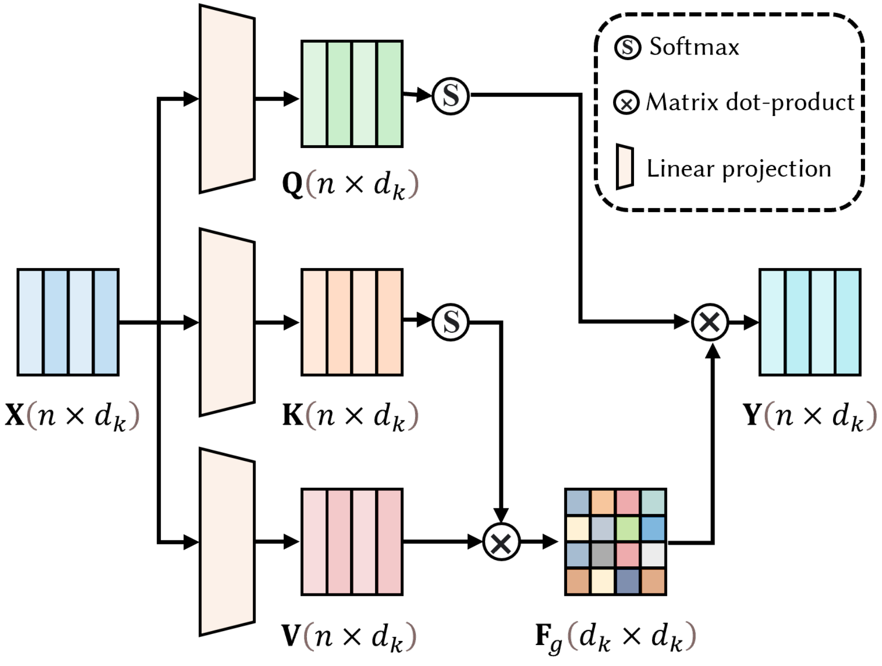


$$\mathcal{L} = L_1(H, \hat{H}^M) + L_1(H, \hat{H}^T) + KL(\phi^T, \phi^M) + KL(\phi^M, \phi^T) + KL(\phi^T, \psi) + KL(\phi^M, \psi)$$

[1] Petrovich et al. Temos: Generating diverse human motions from textual descriptions.

MotionDiffuse

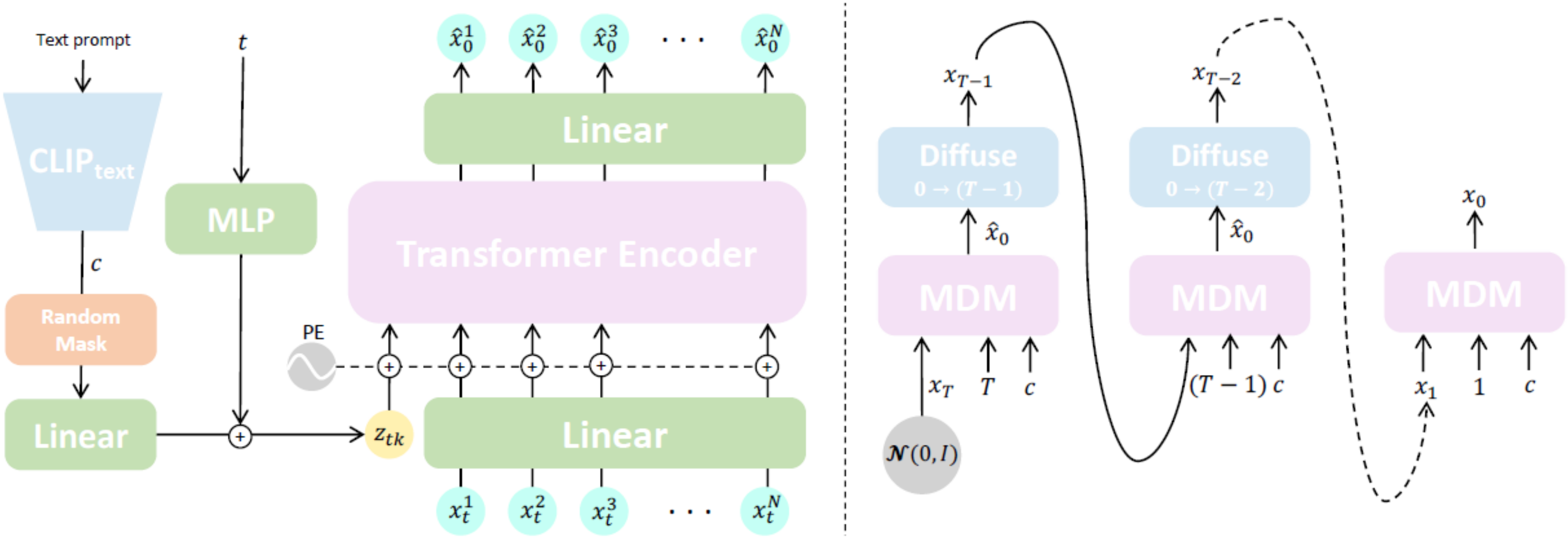




Linear Self-Attention

[2] Zhang et al. Motiondiffuse: Text-driven human motion generation with diffusion model.

MDM



Geometric Loss

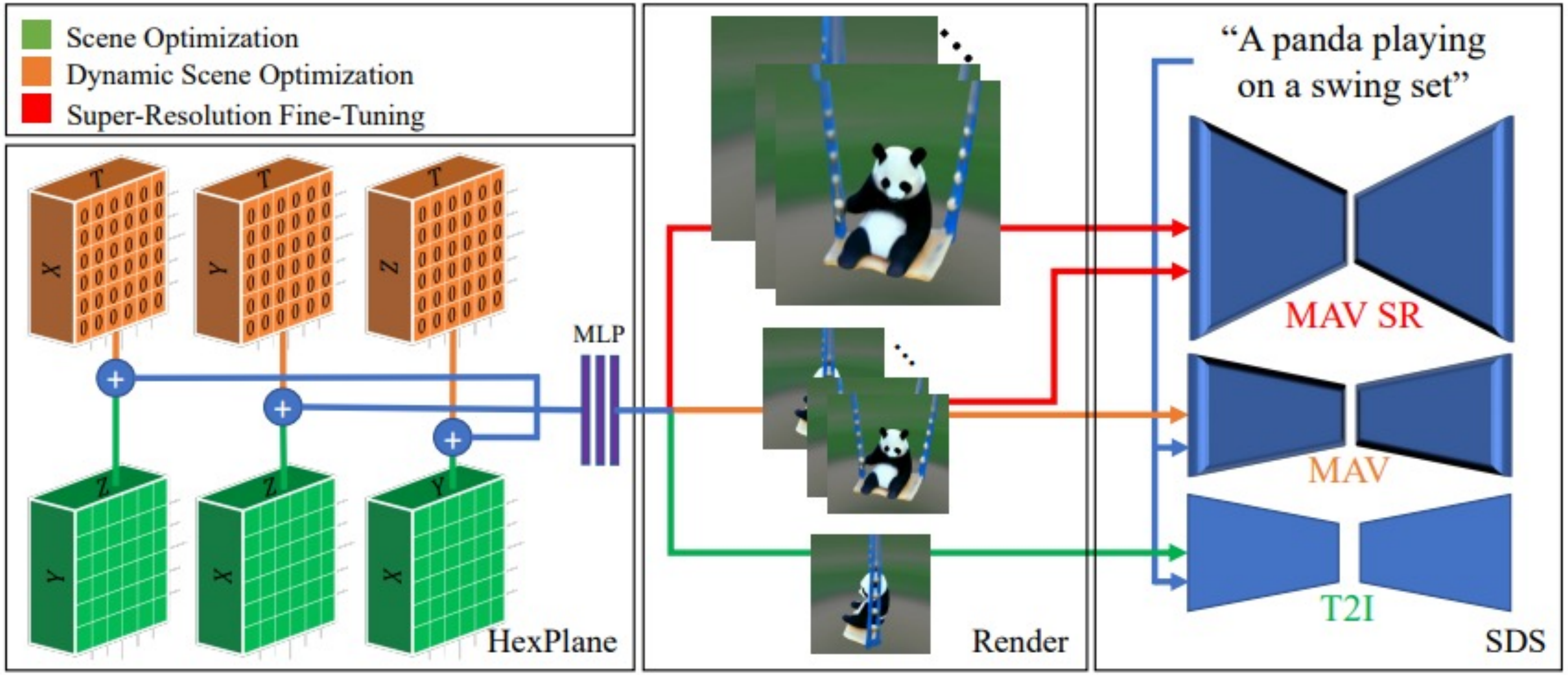
$$\mathcal{L}_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N} \|FK(x_0^i) - FK(\hat{x}_0^i)\|_2^2$$

$$\mathcal{L}_{\text{foot}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(FK(\hat{x}_0^{i+1}) - FK(\hat{x}_0^i)) \cdot f_i\|_2^2$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\|_2^2$$

[3] Tevet et al. Human motion diffusion model.

4D Scene Generation – MAV3D



4D Scene Representation $[P_{xy}^{XYR_1} + P_{zt}^{ZTR_1}; P_{xz}^{XZR_2} + P_{yt}^{YTR_2}; P_{yz}^{YZR_3} + P_{yz}^{XTR_3}]$

Dynamic Scene Optimization $\nabla_\theta \mathcal{L}_{SDS-T} = E_{\sigma,\epsilon}\left[w(\sigma)(\hat{\epsilon}(V_{(\bar{\theta},\sigma,\epsilon)}|y,\sigma)-\epsilon)\frac{\partial V_\theta}{\partial \theta}\right]$

[4] Singer et al. Text-To-4D Dynamic Scene Generation.

Future Direction

1. **More Customized Generation**
2. **More Dynamic Modeling**
3. **More Fine-Grained Alignment**

Acknowledgement

Yuming Jiang

Ziqi Huang

Chenyang Si

Fangzhou Hong

Zhaoxi Chen

Mingyuan Zhang

Ziwei Liu