Applications of Vision-Language Foundation Models

AI602: Recent Advances in Deep Learning

Lecture 5

Jinwoo Shin

KAIST AI

Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

Prompt-to-Prompt Image Editing with Cross-Attention Control [Hertz et al., 2023]

Motivation: Image editing is challenging in text-driven synthesis diffusion models

- Small modification in text prompt leads to different outcome
- Prior works require a spatial mask for localized image editing

Contribution: Textual editing method via **Prompt-to-Prompt manipulations**

- Text-only editing (w/o spatial mask) based on cross-attention maps

Cross-attention maps: High-dim tensors binding **pixels and tokens** from the prompt

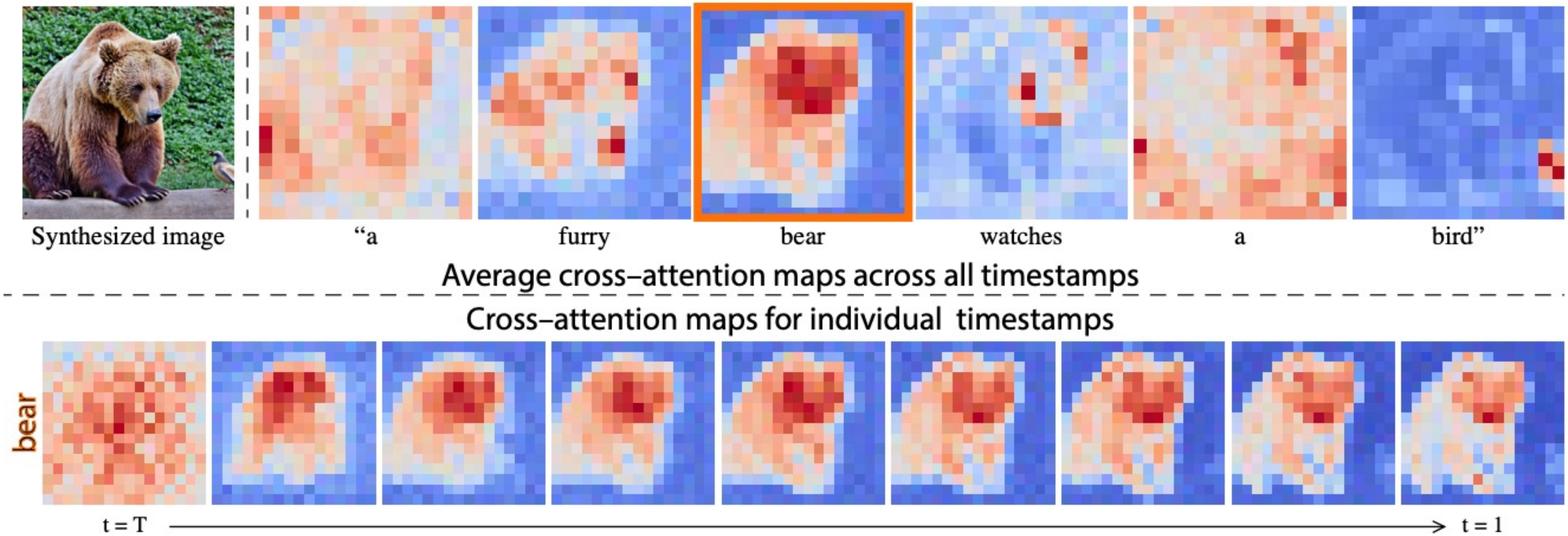
- Contain semantic relations which affects the generated images

Observation: **Spatial layout** and **geometry** depend on the cross-attention maps

- Pixels are more attracted to the words describing them (e.g., bear)

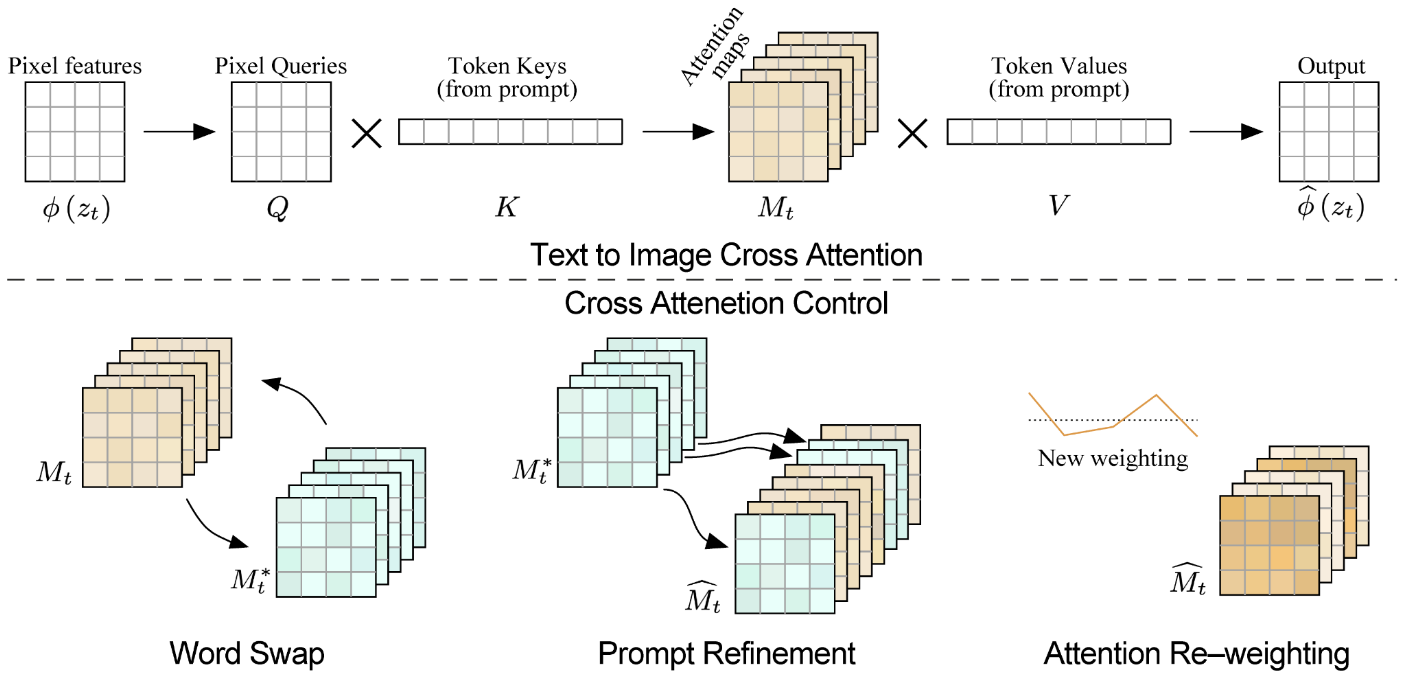
🤔 How to utilize **cross-attention maps** for image editing?

💡 **Inject the attention maps** of original prompt to the modified prompt



Main Idea: Injecting **cross-attention maps** during the diffusion process

- **Word swap**: attention injection of the source image
  - E.g., “a big bicycle” → “a big car”
- **Prompt refinement**: attention injection over the common tokens
  - E.g., “a castle” → “children drawing of a castle”
- **Attention Re-weight**: increase / decrease the attention weights of specified tokens
  - E.g., more or less ”fluffy” on “a fluffy ball”



Prompt-to-Prompt edits high-quality images with only text modification

Word Swap

Prompt Refinement

Attention Re-weighting

InstructPix2Pix: Learning to Follow Image Editing Instructions [Brooks et al., 2023]

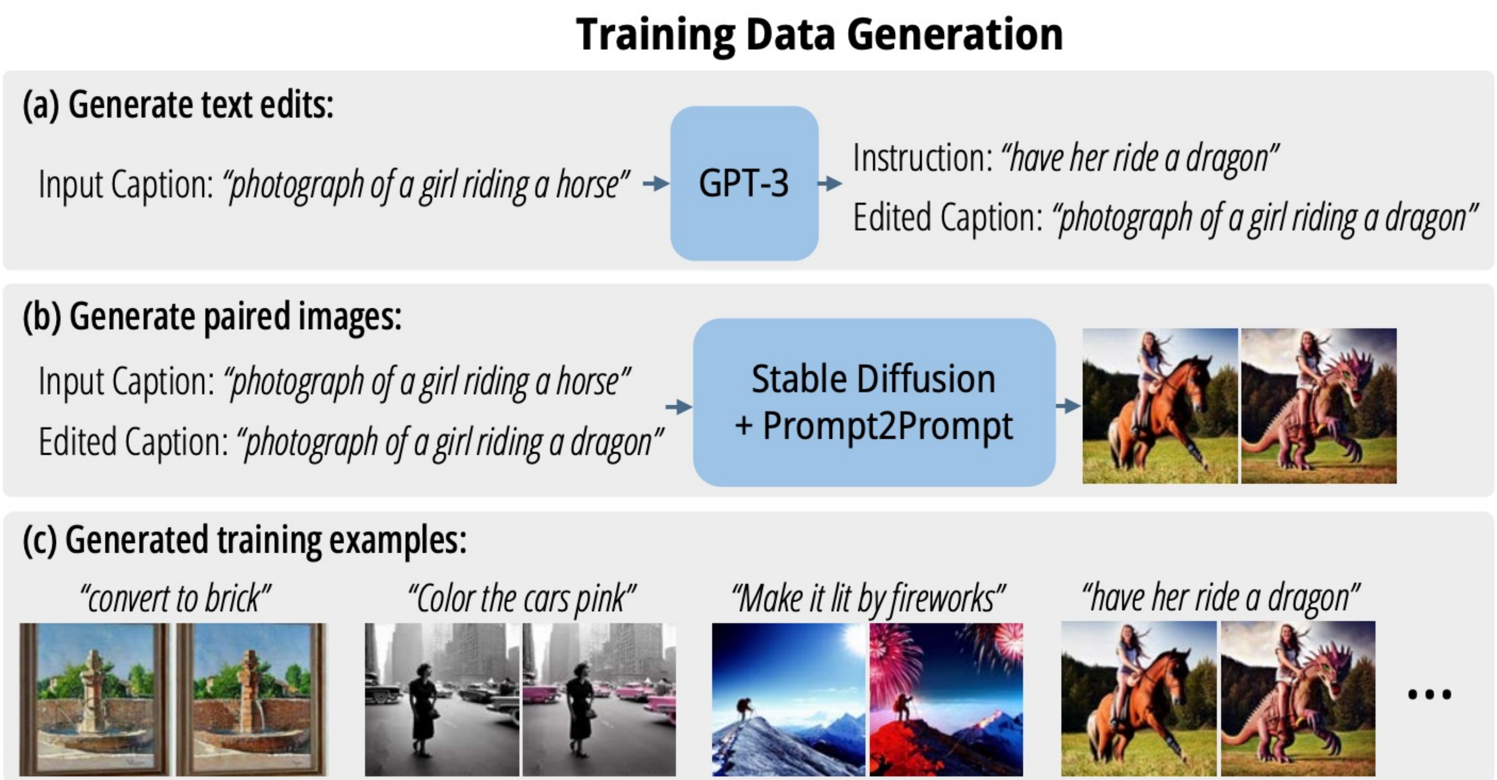
Motivation: Image editing with **detailed prompt** or **extra information** are cumbersome

💡 How about editing images with **human instructions** (e.g., make it big)?

Contribution: Fine-tune a generative model to follow **human instructions**

Main Idea: Treat instruction-based image editing as a **supervised problem**

- **Dataset generation**: Text editing instructions and images before/after the edit
  - Two large-scale models on different modalities: GPT-3 and Stable Diffusion
  - **GPT-3**: Fine-tuned to produce the instructions and the edited caption
  - **Stable Diffusion**: Transform a pair of captions into a pair of images (w/ p2p)



Main Idea: Treat instruction-based image editing as a **supervised problem**

- **Dataset generation:** Text editing instructions and images before/after the edit
  - Two large-scale models on different modalities: GPT-3 and Stable Diffusion
  - **GPT-3**: Fine-tuned to produce the instructions and the edited caption
  - **Stable Diffusion**: Transform a pair of captions into a pair of images (with PtP)
- **Training:** Train Stable diffusion on generated paired dataset

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T))\|_2^2\right]$$

: Input image conditioning : Text instruction conditioning

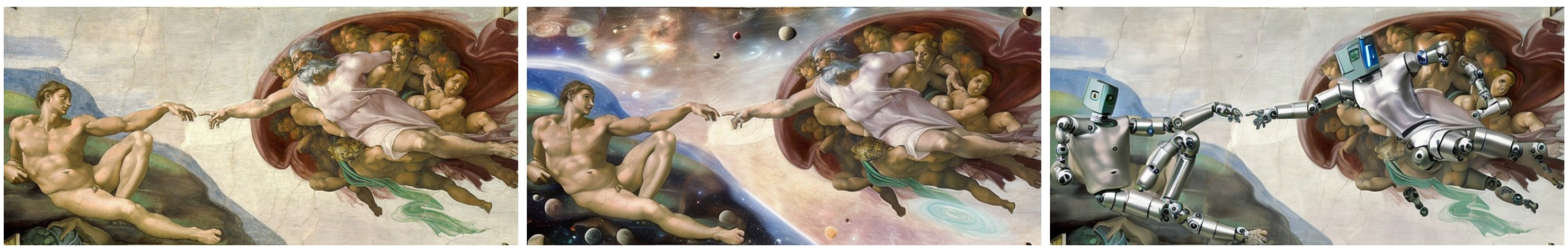
- **Classifier-free guidance for two conditionings**
  - Leverage classifier-free guidance w.r.t. **input image c_I** and **text instruction c_T**

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned}$$

InstructPix2Pix performs many challenging edits

- E.g., replacing object, changing seasons, replacing backgrounds and etc.

Input | *"Make it a Modigliani painting"* | *"Make it a Miro painting"* | *"Make it an Egyptian sculpture"* | *"Make it a marble roman sculpture"*

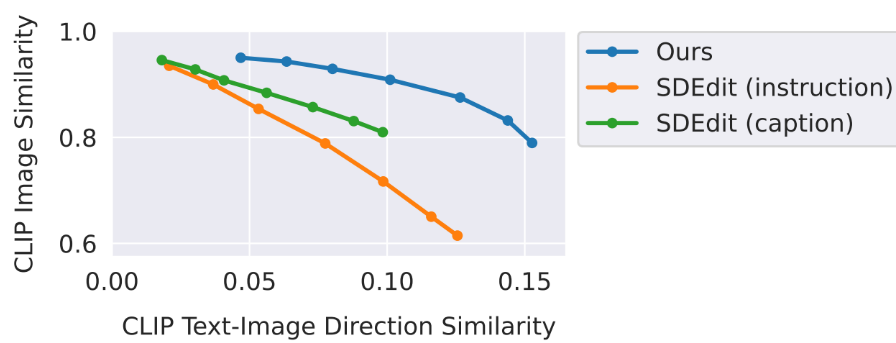


Input | *"Put them in outer space"* | *"Turn the humans into robots"*

Trade-off in consistency

- Consistency with the input images (y-axis)
- Consistency with the edit (x-axis)

→ **Higher image consistency**



Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

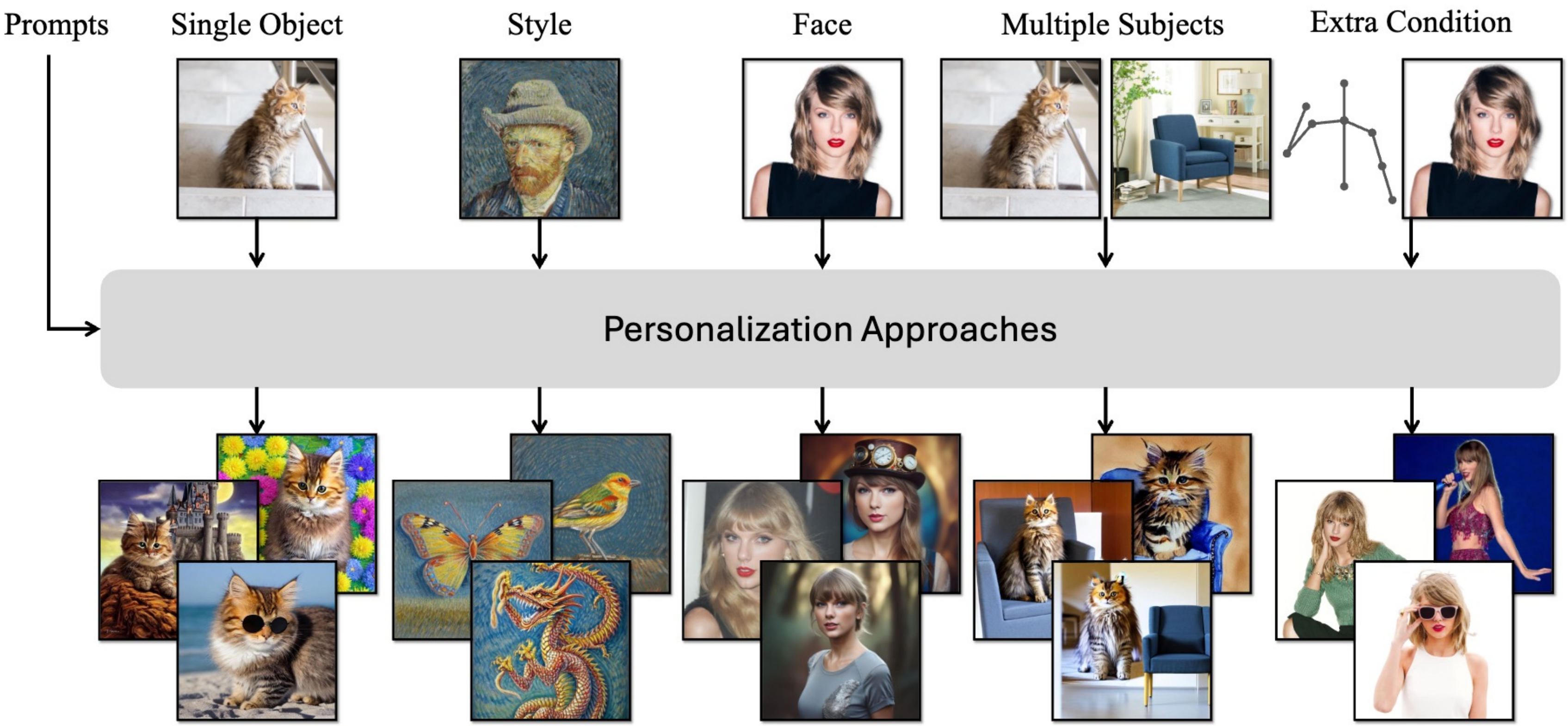
- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

Model Personalization: introduce new concept with a small set of user-provided examples and generate variations of the new concept

- Concept of interest encompasses object, faces, styles and other semantic elements

Main Challenge: Difficulty in introducing **new concept** into large scale models

- Small set of data often results to overfitting or catastrophic forgetting



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion [Gal et al., 2023]

Motivation: Difficulty in introducing **new concepts** into large scale models

- Re-training requires huge amount of cost
- Fine-tuning on few examples leads to catastrophic forgetting

Contribution: Personalized text-to-image generation (given 3-5 images)

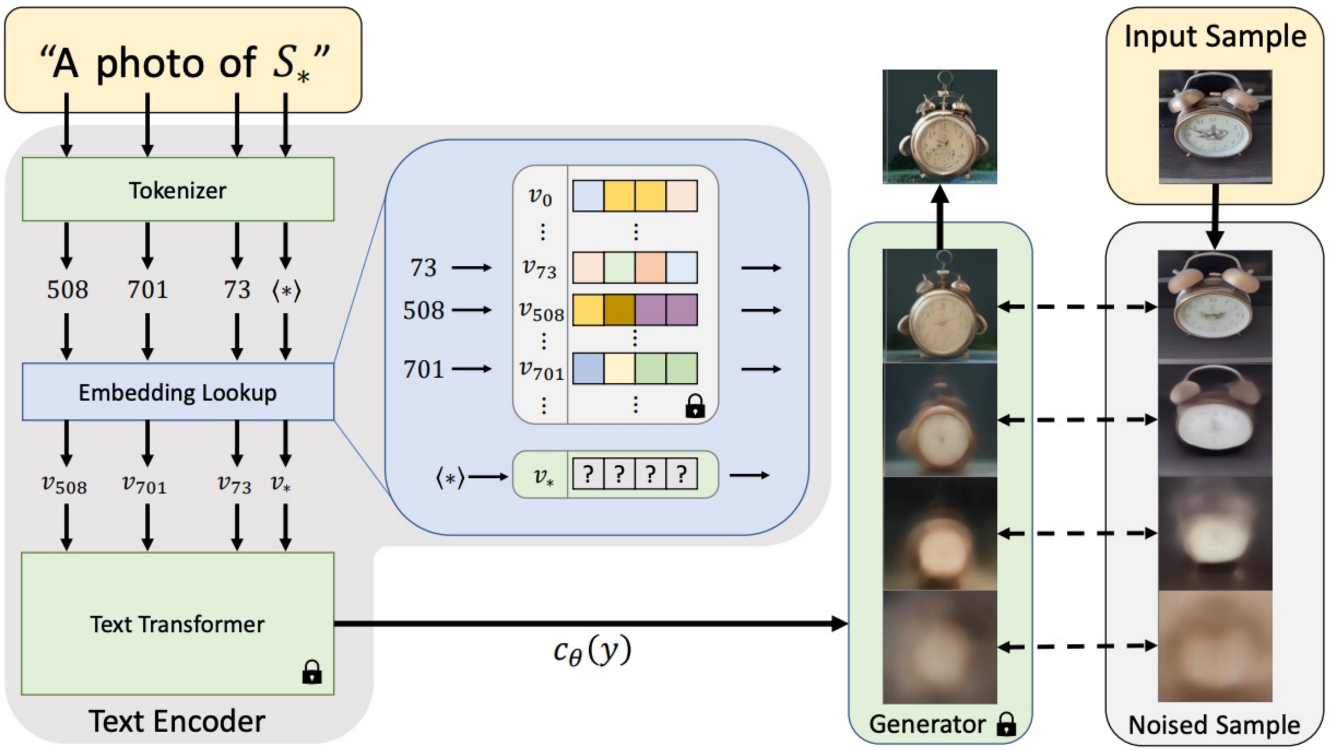
- **Textual inversion**: find new pseudo-words capturing visual semantics and details

Main Idea: Find new **pseudo-word** in text embedding space (in LDMs)

- For pseudo-word $\boldsymbol{S}^*$, directly optimize textual embedding $\boldsymbol{v}^*$ of $\boldsymbol{S}^*$

$$v_* = \arg\min_{v} \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$$

: Learnable new token embedding : Frozen LDM model



Textual Inversion enables **capturing** and **recreating** variations of an object

- Image synthesis guided by a caption lacks fine-grained detail (e.g., color patterns)
- Capture finer details and compose novel scenes w/ only a single token embedding

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation [Ruiz et al., 2023]

Motivation: Lack the ability to synthesize same subjects in different context

- Output domain is limited; detailed textual description yield different appearances

Contribution: Personalization of text-to-image diffusion models (given 3-5 images)

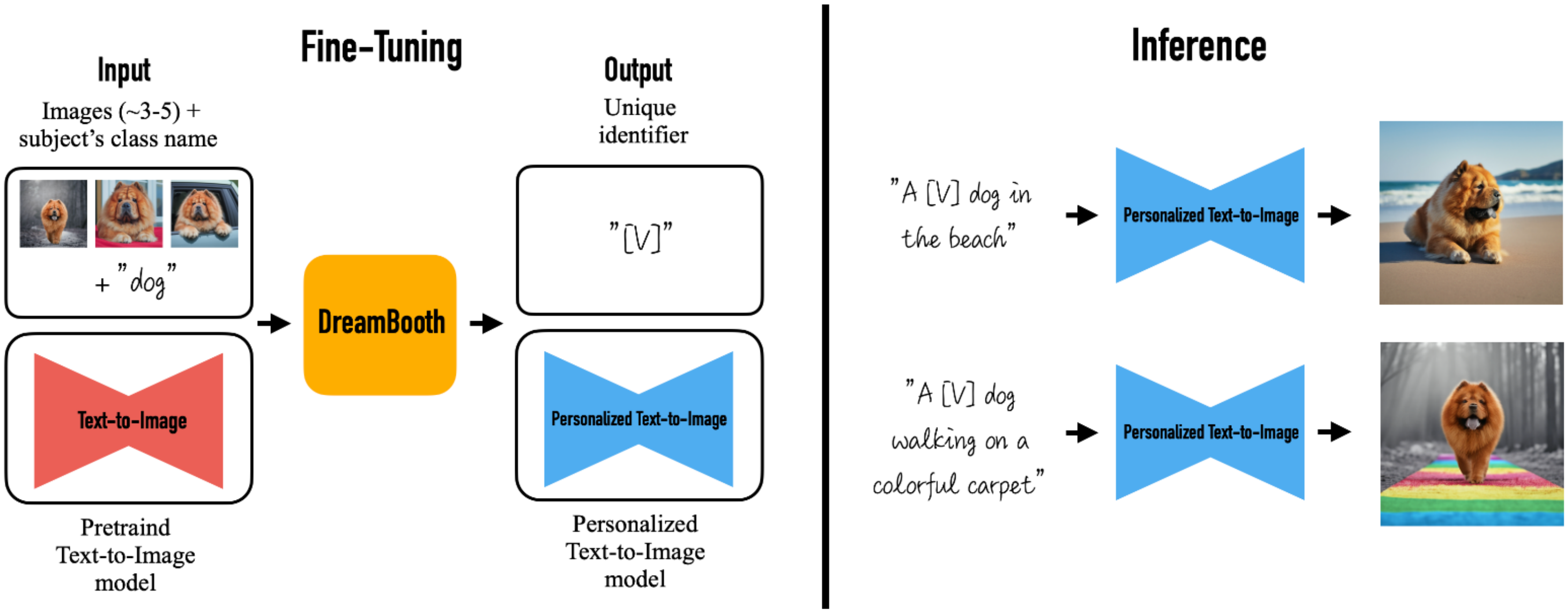
- **Fine-tuning method** to implant the given subject into the model's output domain

Main Idea: Fine-tune text-to-image model w/ **few images** of a subject and **class name**

- Text prompt with unique identifier and the class name (e.g., a $[V]$ dog)
  - **Unique identifier**: class-specific instance
  - **Class name**: prior knowledge on the subject class

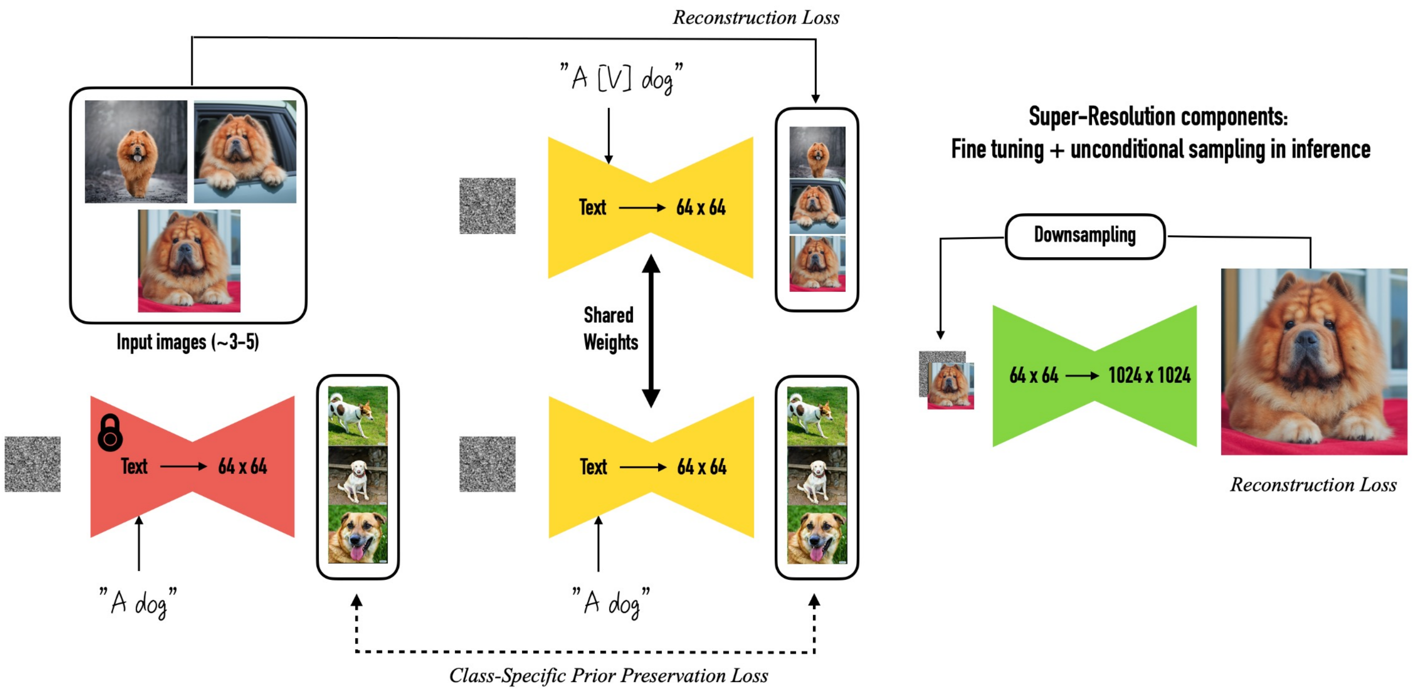
However, fine-tuning text-to-image model with small set may cause:

1. Language drift
2. Reduced output diversity



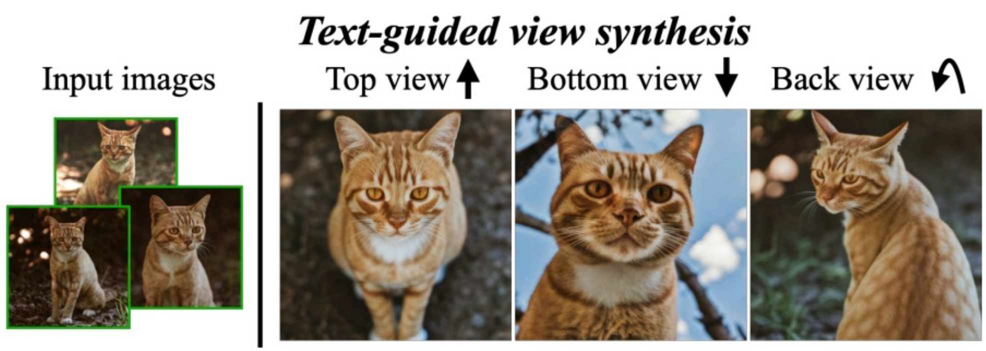
Main Idea: Fine-tune text-to-image model w/ few images of a subject and class name

- Text prompt with unique identifier and the class name (e.g., a $[V]$ dog)
  - **Unique identifier**: class-specific instance
  - **Class name**: prior knowledge on the subject class
- **Class-specific prior preservation loss**
  - Supervise the model w/ own generated samples
  - Leverages the semantic prior that the model has on the class



- Generates image with high preservation of subject details in various context

- Generate novel views with preserving subject identity



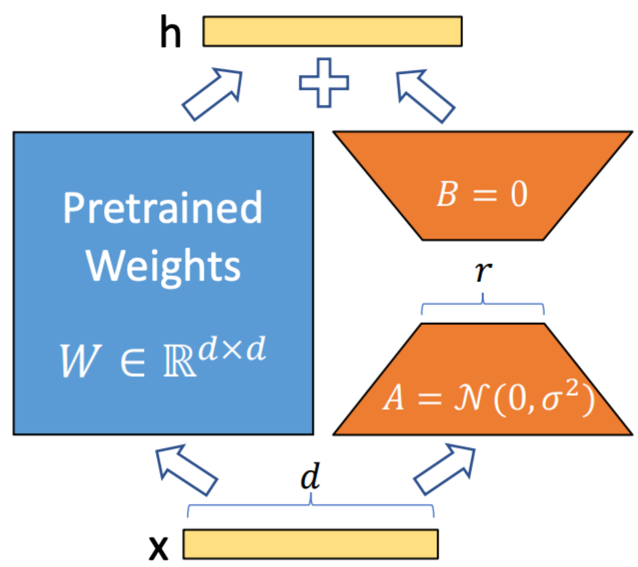
🤔 How to efficiently **fine-tune** large models (e.g., DreamBooth)?

💡 Reduce the **number of trainable parameters**, not fine-tuning all parameters

LoRA: Low-Rank Adaptation of Large Language Models [Hu et al., 2022]

- Freeze the original weights and update only **low-rank decomposed matrices**

$$h = W_0 x + \Delta W x = W_0 x + BAx$$



→ **LoRA enables faster and memory efficient DreamBooth fine-tuning**

Limitations

Major challenges: Tradeoff between textual alignment and concept consistency

- **Textual Inversion:** word embedding is not dense enough to capture visual features
  - Details of subject are often ignored; low concept consistency
- **DreamBooth:** often leads to overfitting and catastrophic forgetting
  - Can't generate diverse images following textual prompts; low textual alignment

Textual Inversion

a photo of V*

V* in Times Square

DreamBooth

a photo of V*

V* on a beach

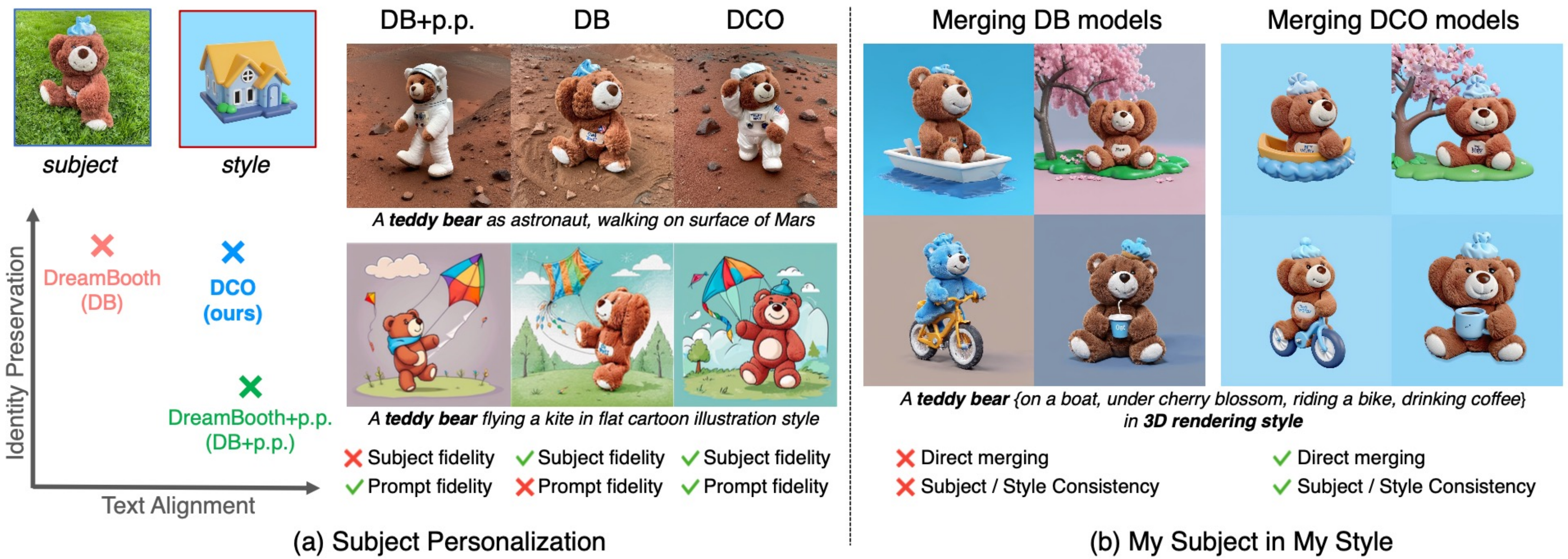
DCO: Direct Consistency Optimization for Robust Customization of Text-to-Image Diffusion Models [Lee et al., 2024]

Motivation: Reduced ability of fine-tuned model compared to pretrained model

- Low textual alignment and compositional generation capability

Contribution: Retaining the pretrained knowledge during low-shot fine-tuning

- **Novel fine-tuning objective** to mitigate the forgetting behavior w/o additional data



Main Idea: Controls the **deviation between fine-tuning and pretrained models**

- Consider the deviation of KL between fine-tuning model and pretrained model

$$\Delta(p_\theta, p_\phi; \mathbf{x}, \mathbf{c}) = \underbrace{D_{\mathrm{KL}}\big(q(\mathbf{z}_{0:1}|\mathbf{x}) \,\|\, p_\phi(\mathbf{z}_{0:1}|\mathbf{c})\big)}_{\text{ELBO of pretrained model}} - \underbrace{D_{\mathrm{KL}}\big(q(\mathbf{z}_{0:1}|\mathbf{x}) \,\|\, p_\theta(\mathbf{z}_{0:1}|\mathbf{c})\big)}_{\text{ELBO of fine-tuning model}}$$

- DCO aims to control the deviation by following log-loss:

$$\mathcal{L}_\Delta(\theta; \mathbf{x}, \mathbf{c}) = -\log \sigma\big(\underbrace{\beta}_{\text{Control hyperparameter}} \Delta(p_\theta, p_\phi; \mathbf{x}, \mathbf{c})\big)$$

- Efficient implementation of DCO loss:

$$\mathcal{L}_{\mathrm{DCO}}(\theta; \mathbf{x}, \mathbf{c}) = \mathbb{E}_{t,\boldsymbol{\epsilon}}\big[-\log \sigma\big(-\beta_t(\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t; \mathbf{c}, t) - \boldsymbol{\epsilon}\|_2^2 - \|\boldsymbol{\epsilon}_\phi(\mathbf{z}_t; \mathbf{c}, t) - \boldsymbol{\epsilon}\|_2^2)\big)\big]$$

	Fine-tuning model	**Pretrained model**
Noise Prediction Model	ϵ_θ	ϵ_ϕ
Model density	p_θ	p_ϕ

Main Idea: Controls the **deviation between fine-tuning and pretrained models**

- DCO directly **regularize KL-divergence** w.r.t. reference images
  - Prior preservation loss which uses auxiliary data causes undesirable model shift

Algorithm 1 Regular fine-tuning

Require: Dataset $\mathcal{D}_{\text{ref}}$, fine-tuning model $\boldsymbol{\epsilon}_\theta$, learning rate $\eta > 0$

1: **while** not converged **do**
2: Sample $(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\text{ref}}$
3: Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: Sample $t \sim \mathcal{U}(0, 1)$
5: $\mathbf{z}_t \leftarrow \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$
6: $\mathcal{L}_{\text{DM}}(\theta) \leftarrow \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t; c, t) - \boldsymbol{\epsilon}\|_2^2$
7: Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{DM}}(\theta)$
8: **end while**

Algorithm 2 Fine-tuning with DCO loss

Require: Dataset $\mathcal{D}_{\text{ref}}$, fine-tuning model $\boldsymbol{\epsilon}_\theta$, pre-trained model $\boldsymbol{\epsilon}_\phi$, temperature $\beta_t > 0$, learning rate $\eta > 0$

1: **while** not converged **do**
2: Sample $(\mathbf{x}, \mathbf{c}) \sim \mathcal{D}_{\text{ref}}$
3: Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
4: Sample $t \sim \mathcal{U}(0, 1)$
5: $\mathbf{z}_t \leftarrow \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$
6: $\ell(\theta) \leftarrow \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t; c, t) - \boldsymbol{\epsilon}\|_2^2$
7: $\ell(\phi) \leftarrow \|\boldsymbol{\epsilon}_\phi(\mathbf{z}_t; c, t) - \boldsymbol{\epsilon}\|_2^2$ (no gradient)
8: $\mathcal{L}_{\text{DCO}}(\theta) \leftarrow -\log \sigma\big(-\beta_t\big(\ell(\theta) - \ell(\phi)\big)\big)$
9: Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{DCO}}(\theta)$
10: **end while**

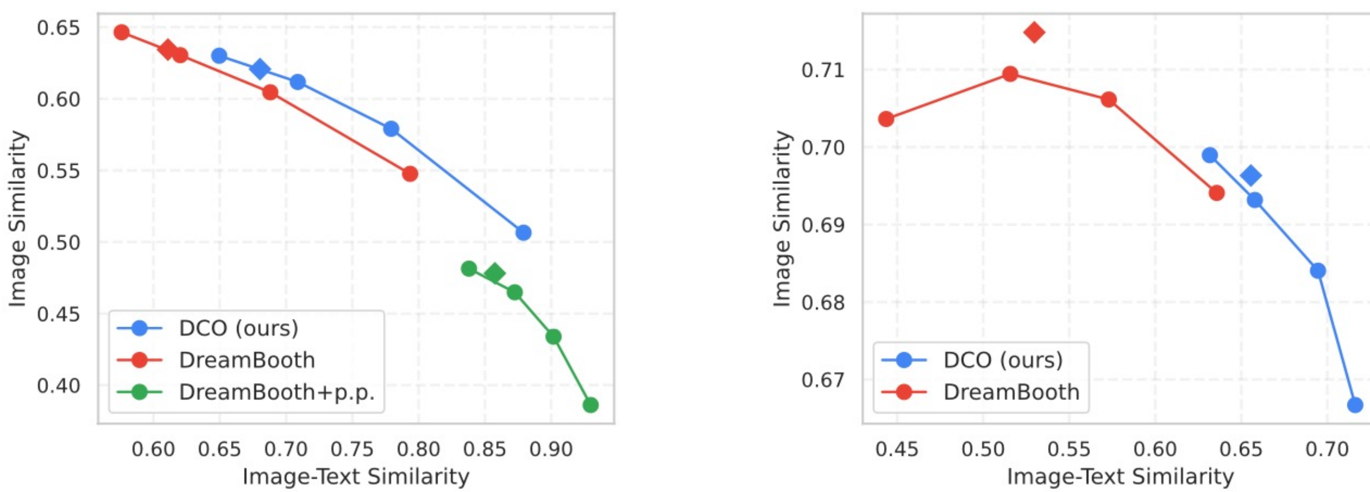
Main Idea: Controls the **deviation between fine-tuning and pretrained models**

- Consistency Guidance Sampling
  - control over the consistency during inference in addition to classifier-free guidance

$$\hat{\epsilon}(\mathbf{z}_t;\mathbf{c},t) = \omega_{\text{con}}\left(\epsilon_\theta(\mathbf{z}_t;\mathbf{c},t) - \epsilon_\phi(\mathbf{z}_t;\mathbf{c},t)\right) + \omega_{\text{text}}\left(\epsilon_\phi(\mathbf{z}_t;\mathbf{c},t) - \epsilon_\phi(\mathbf{z}_t,t)\right) + \epsilon_\phi(\mathbf{z}_t,t)$$

: classifier-free guidance

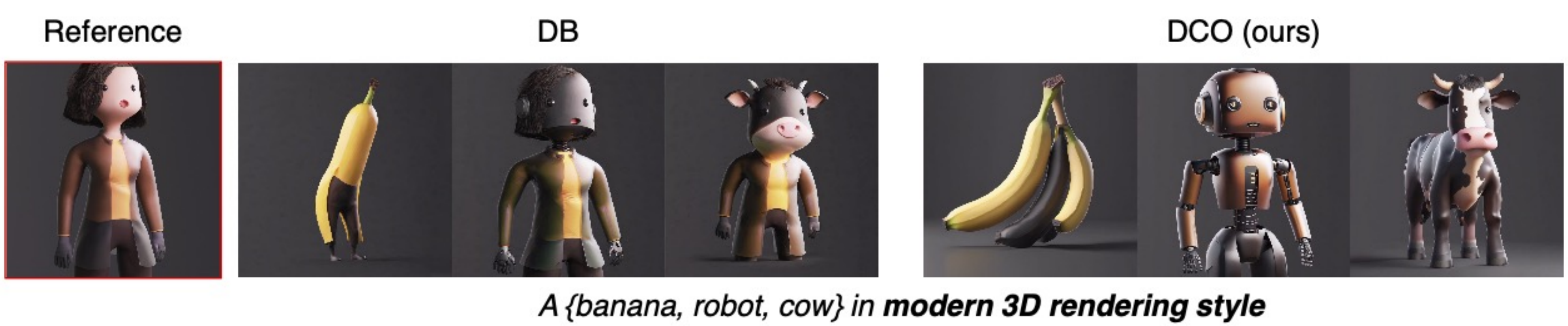
- DCO positions on a superior Pareto frontier between **textual alignment** and **concept consistency**
  - Minimal fine-tuning retain the capability of pretrained model



(a) Subject personalization (b) Style personalization

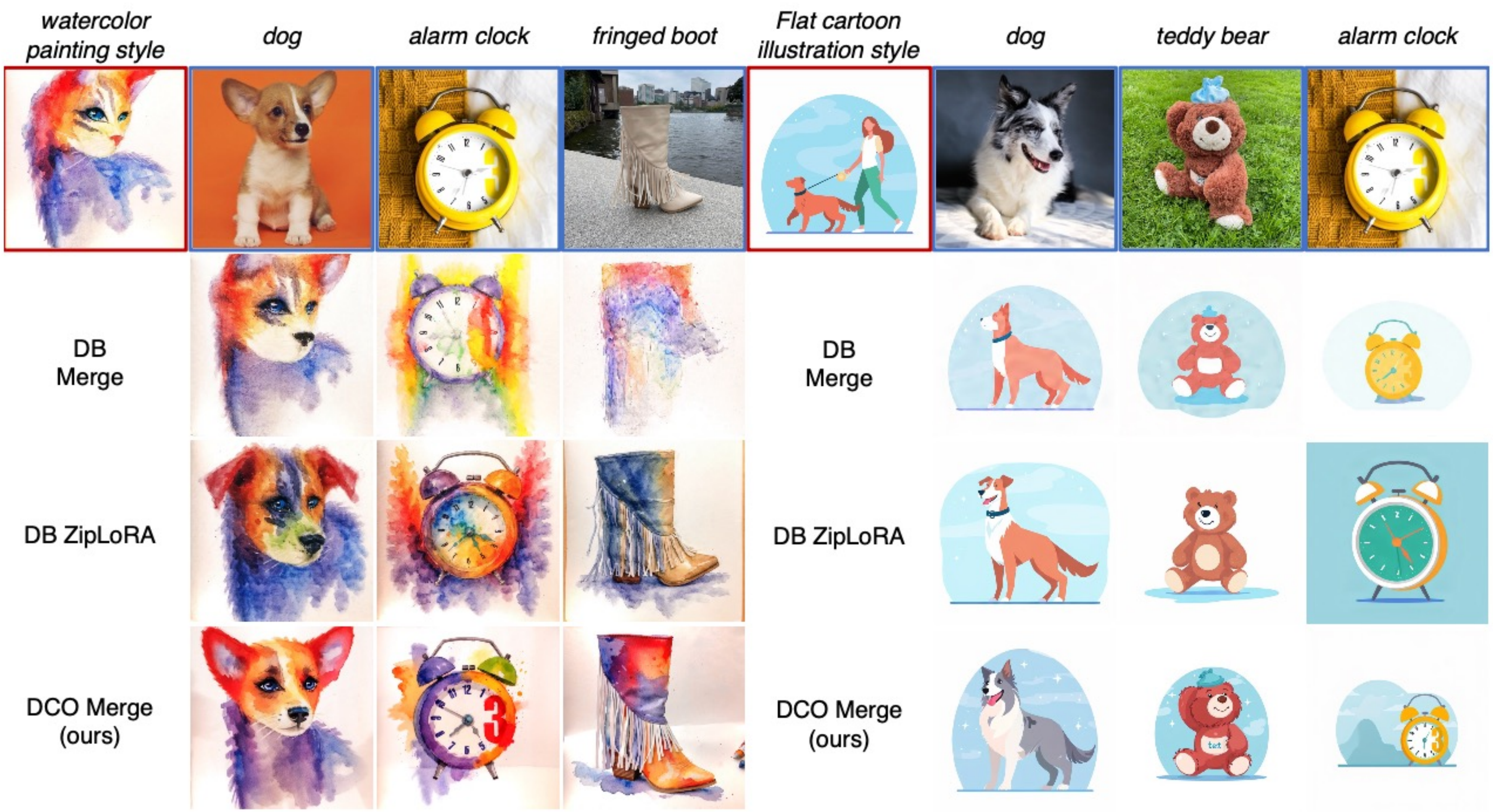
- **Generates various visual attributes as well as into various styles**

- **Generate images with consistent styles w/o entangling content from reference images**



Models fine-tuned w/ DCO can be merged without interference

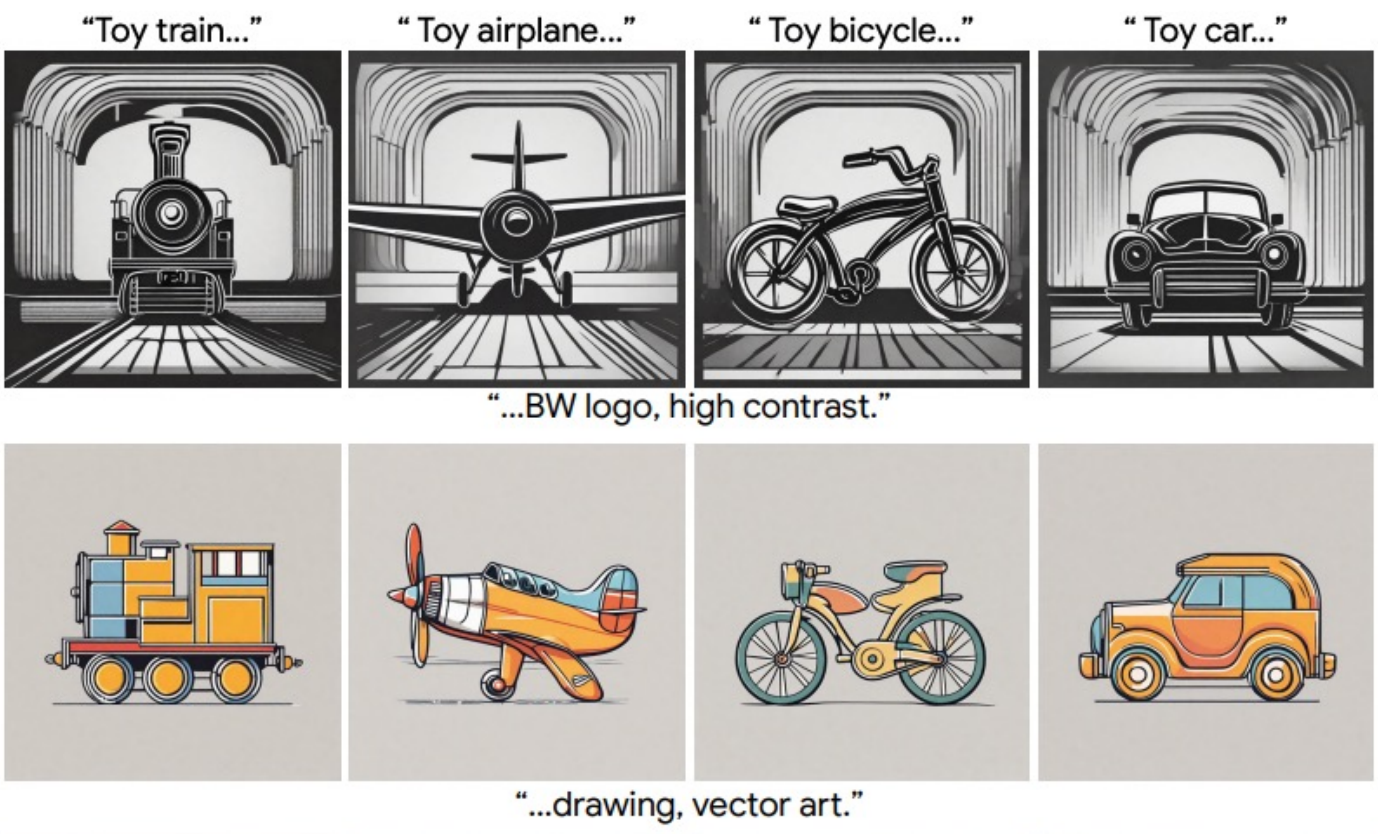
- Enables to generate custom subjects in a custom style w/o post-optimization



Style Aligned Image Generation via Shared Attention [Hertz et al., 2023]

Motivation: Ensuring style consistency requires fine-tuning and manual intervention to dis entangle content and style

Contribution: Training-free style alignment among a series of generated images

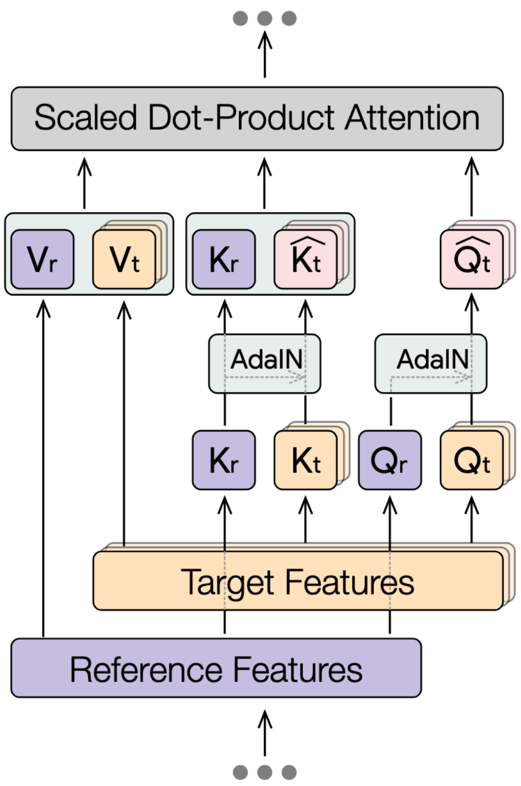


Main Idea: Manipulate self-attention for communication among generated images

- Sharing Keys and values of attention (K_i, V_i) in the batch.
- Normalize Q_t and K_t of the target image using Q_r and K_r of the reference image using AdaIN.

$$\text{Attention}(Q_i, K_{1\ldots n}, V_{1\ldots n})$$

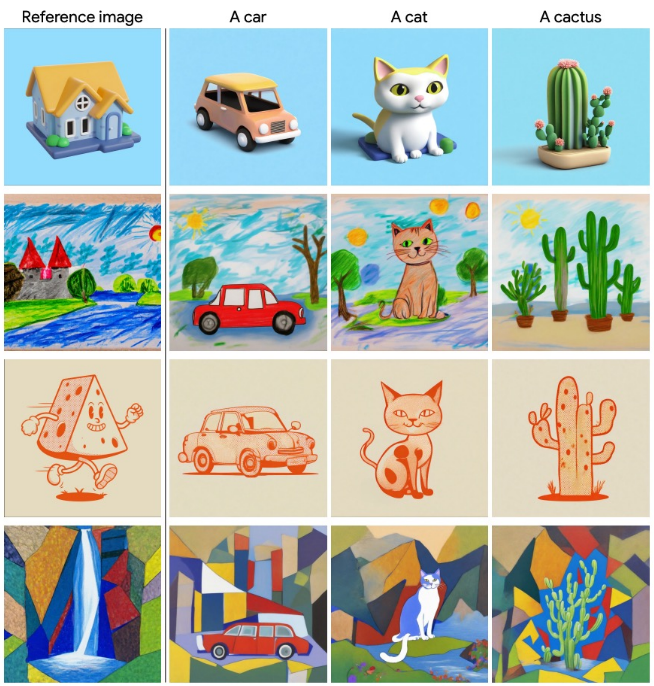
$$\text{where } K_{1\ldots n} = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{bmatrix} \text{ and } V_{1\ldots n} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}$$



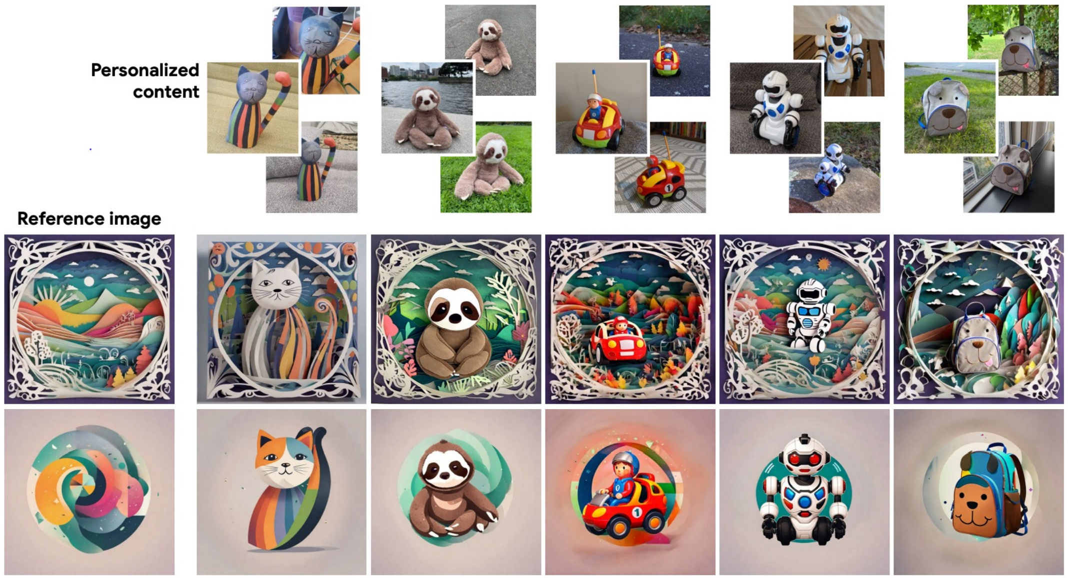
StyleAligned can be integrated into different applications

- Style reference image is given
- Object reference images are given

Style Reference given



Object reference given



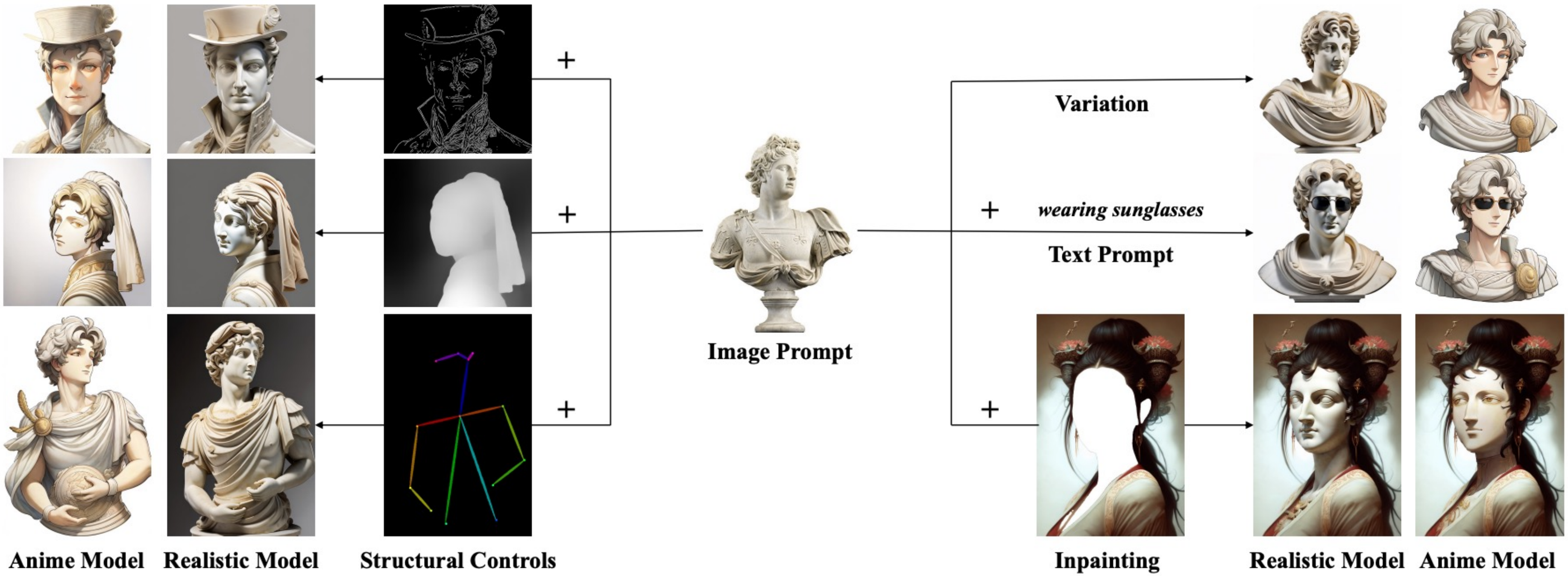
IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models [Ye et al., 2024]

Motivation: Control w/ text prompt is limited as it involves complex engineering

- Prior works (e.g., direct fine-tuning) requires large computing resources

Contribution: Extended capability of image prompting w/ lightweight adapter

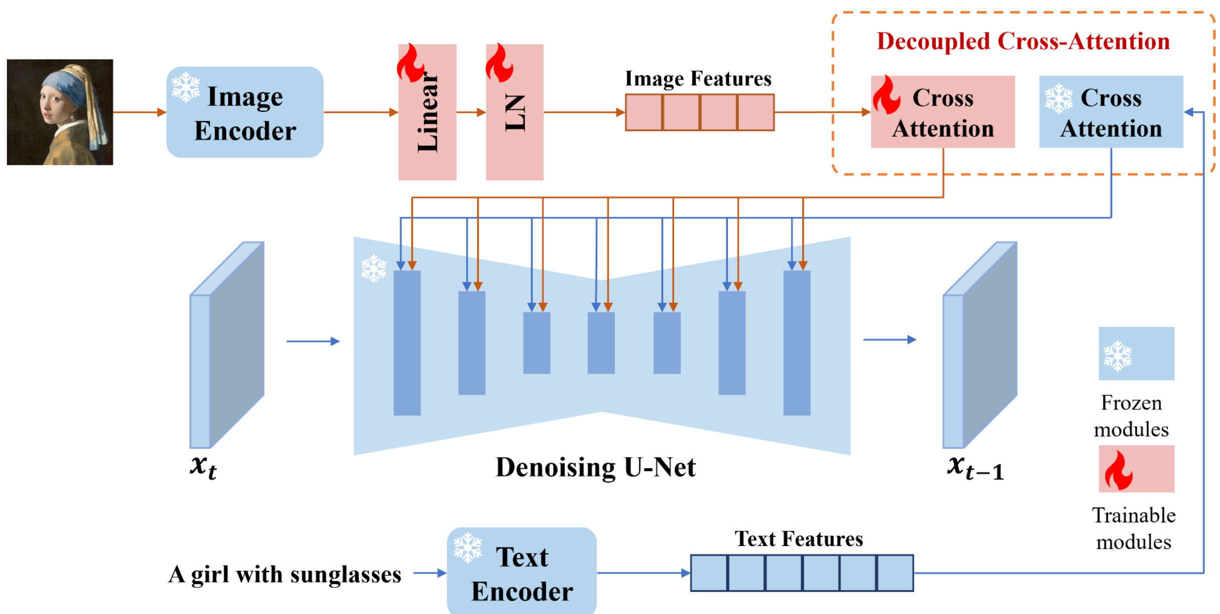
- **Effective adapter design** to incorporate both text and image prompts



Main Idea: Lightweight adapter via **decoupled cross-attention mechanism**

- Frozen image encoder (e.g., CLIP) to extract image features from image prompt
  - Small trainable projection network to project into a sequence of features
- Adapter module with decoupled cross-attention to embed image features

$$\mathbf{Z}^{new} = \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}}_{\text{Text cross-attention}} + \underbrace{\text{Softmax}\left(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}}\right)\mathbf{V}'}_{\text{Image cross-attention}}$$



Main Idea: Lightweight adapter via **decoupled cross-attention mechanism**

- Frozen image encoder (e.g., CLIP) to extract image features from image prompt
  - Small trainable projection network to project into a sequence of features
- Adapter module with decoupled cross-attention to embed image features

$$\mathbf{Z}^{new} = \underbrace{\text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}}_{\text{Text cross-attention}} + \underbrace{\text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}'}_{\text{Image cross-attention}}$$

- **Training:** Same training objective as original T2I models w/ image-text pairs
  - 10 M text-image pairs from LAION-2B and COYO-700M

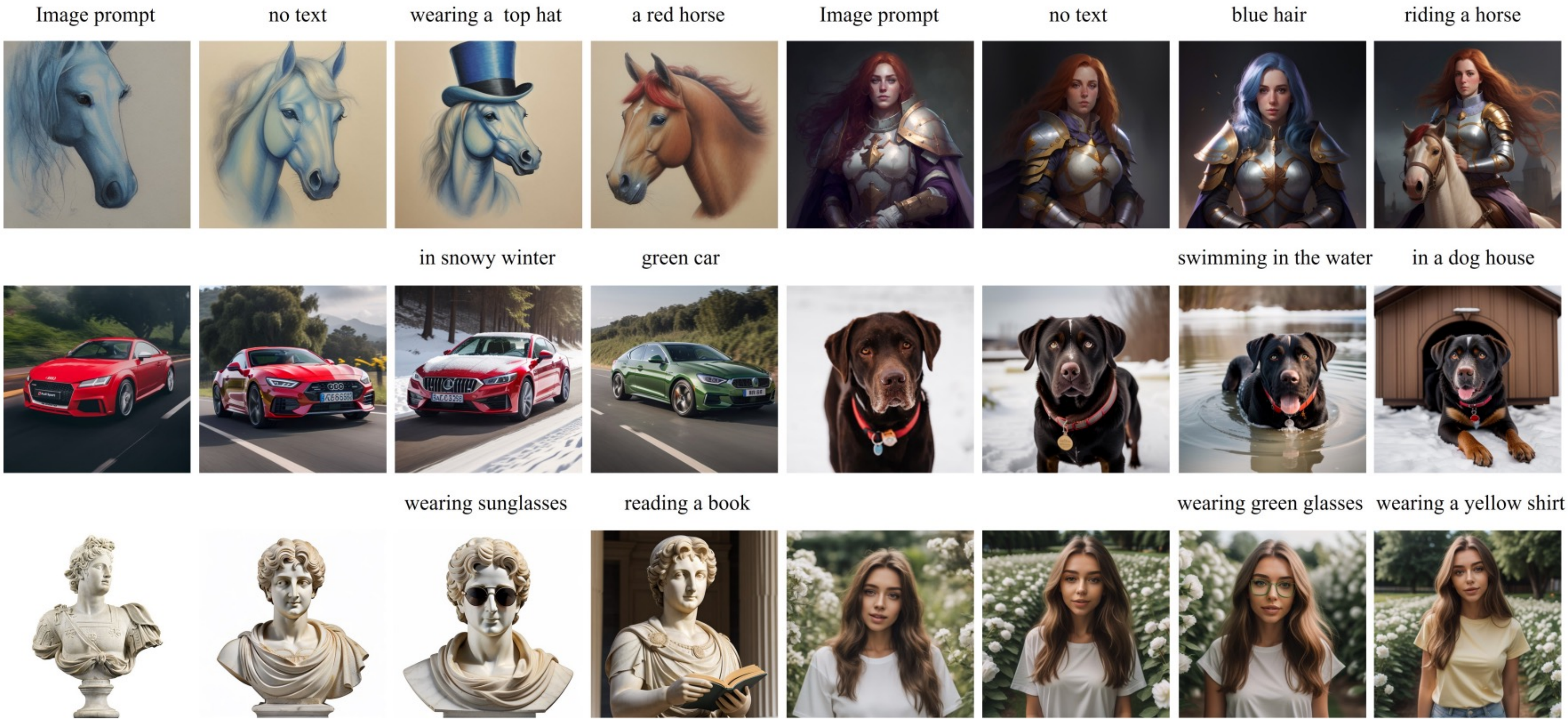
$$L_{\text{simple}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}, \boldsymbol{c}_t, \boldsymbol{c}_i, t} \| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \big(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t \big) \|^2$$

: text features

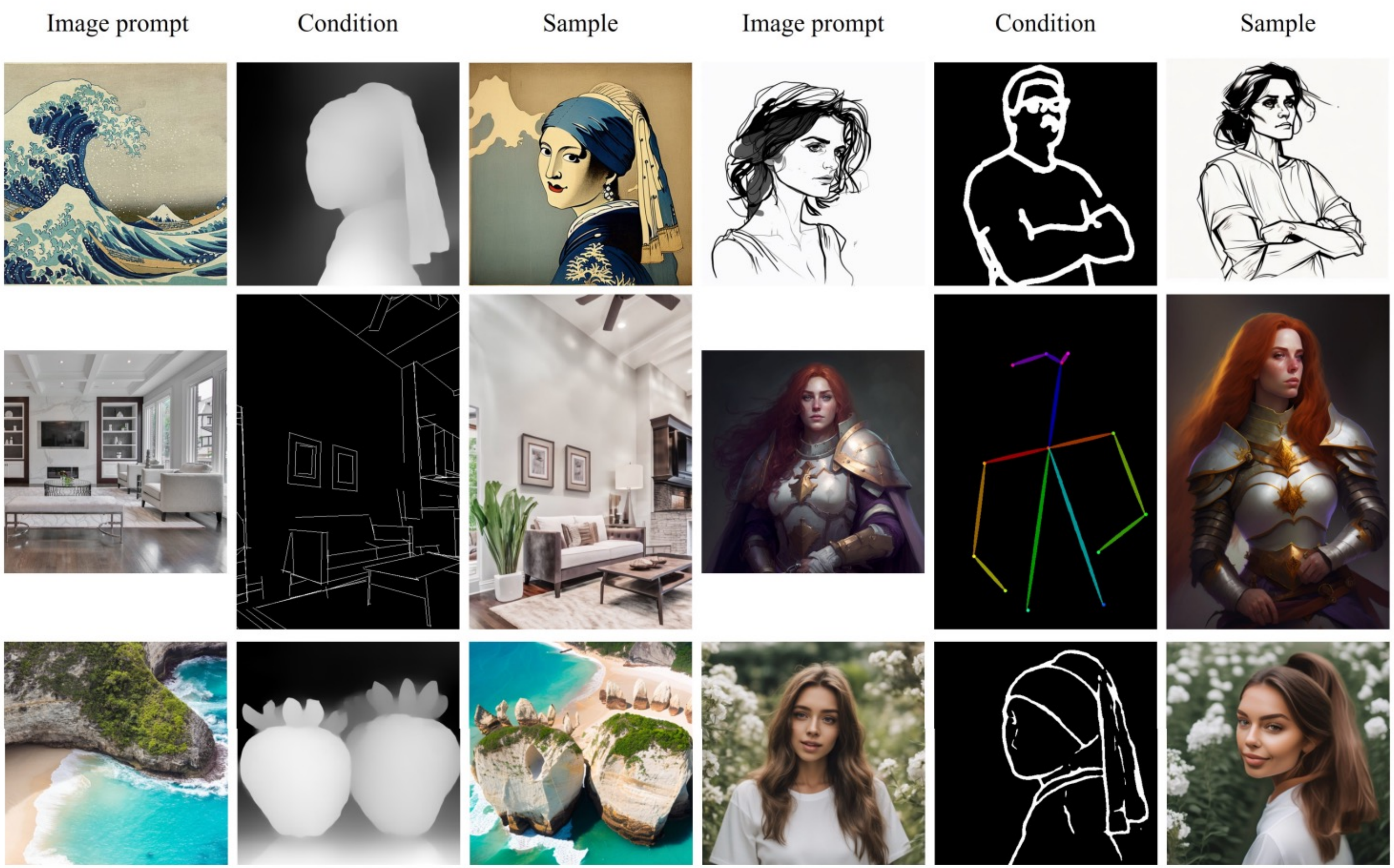
: image features

IP-Adapter [Ye et al., 2024]

- **Generates images with high identity preservation w/ diverse prompts**



- **Enables incorporating additional structural conditions w/o fine-tuning**



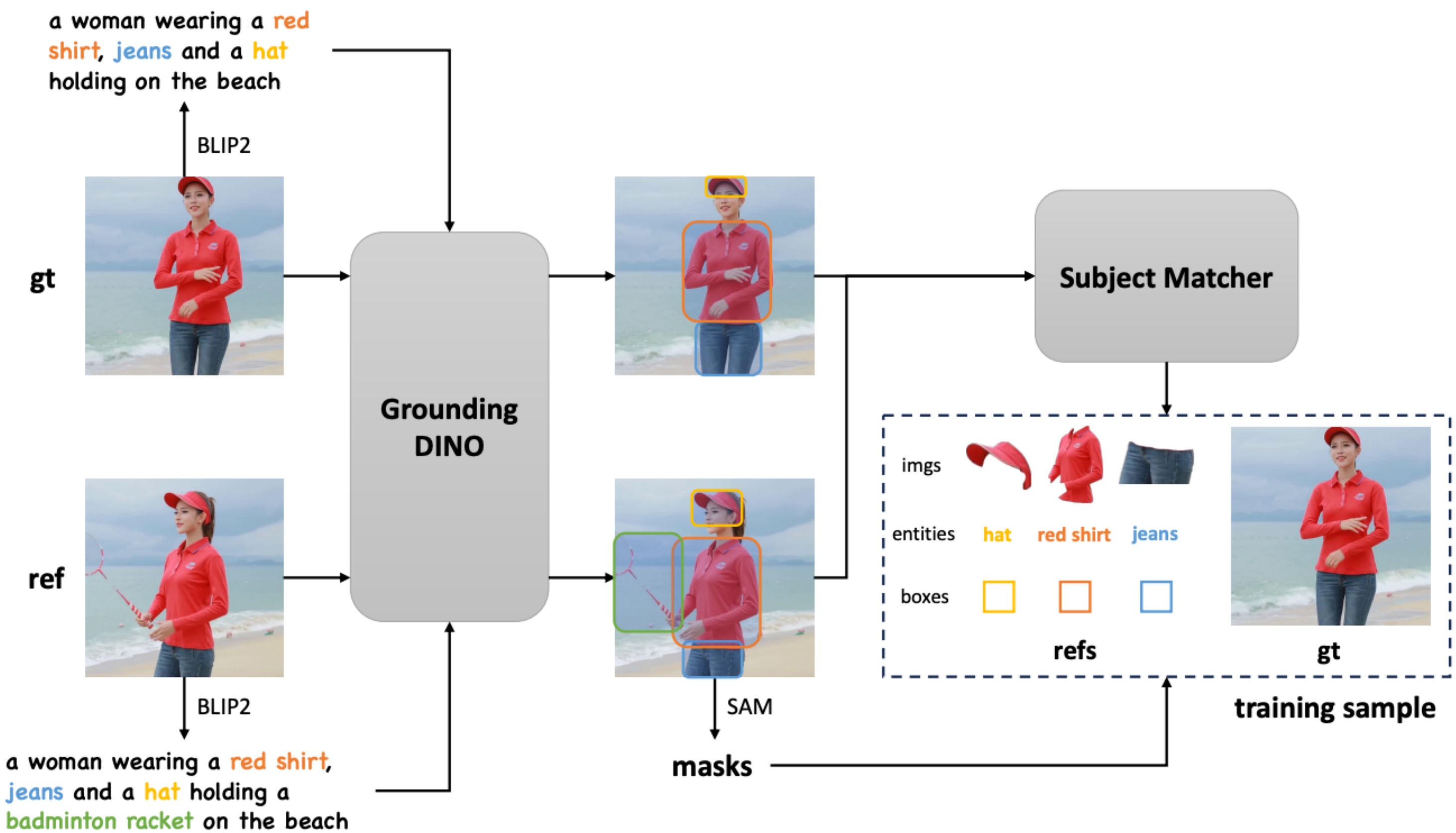
MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance [Wang et al., 2024]

Motivation: Multi-subject personalization still incur notable detail inaccuracies
- e.g., subject blending, subject-subject

Contribution: Layout-guided zero-shot image personalization w/ multiple subjects

Main Idea: Separately extract image features of each subject with paired data

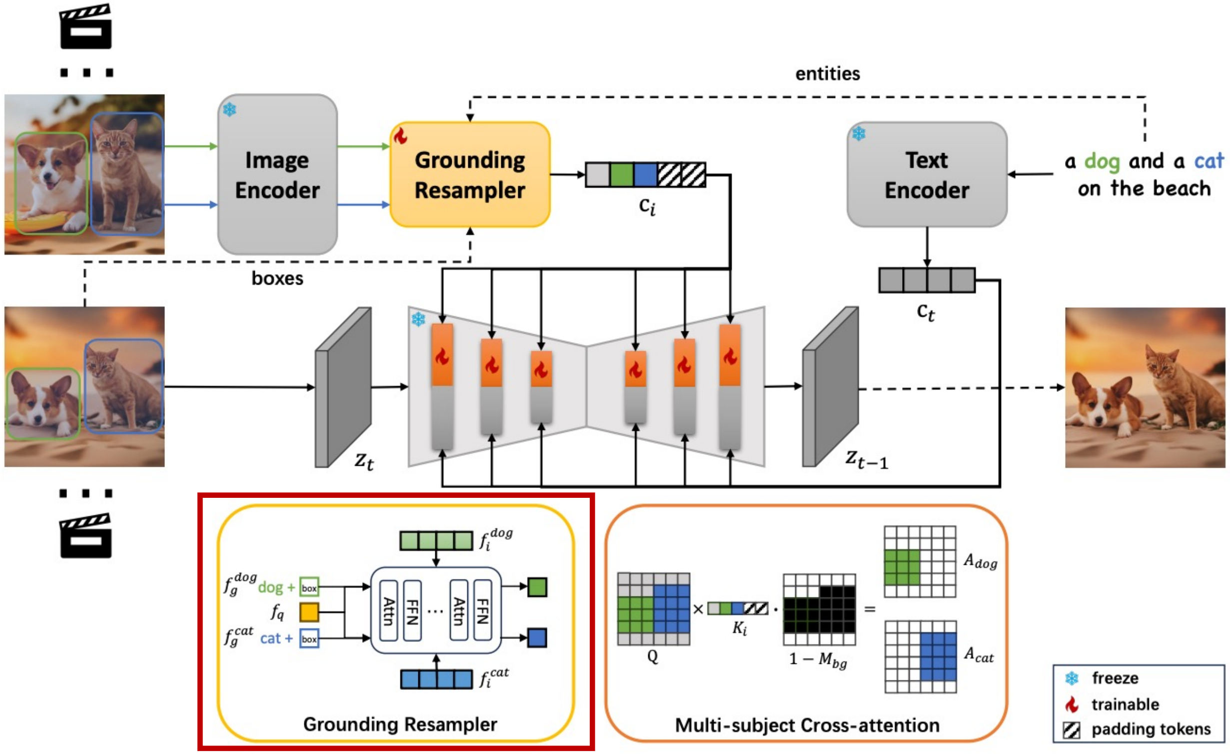
- **Dataset construction for paired data**
  - Stand-alone images often results 'copy-and-paste' artifacts
  - Extract multiple frames in a video for ground truth and reference images



Main Idea: Separately extract image features of each subject with paired data

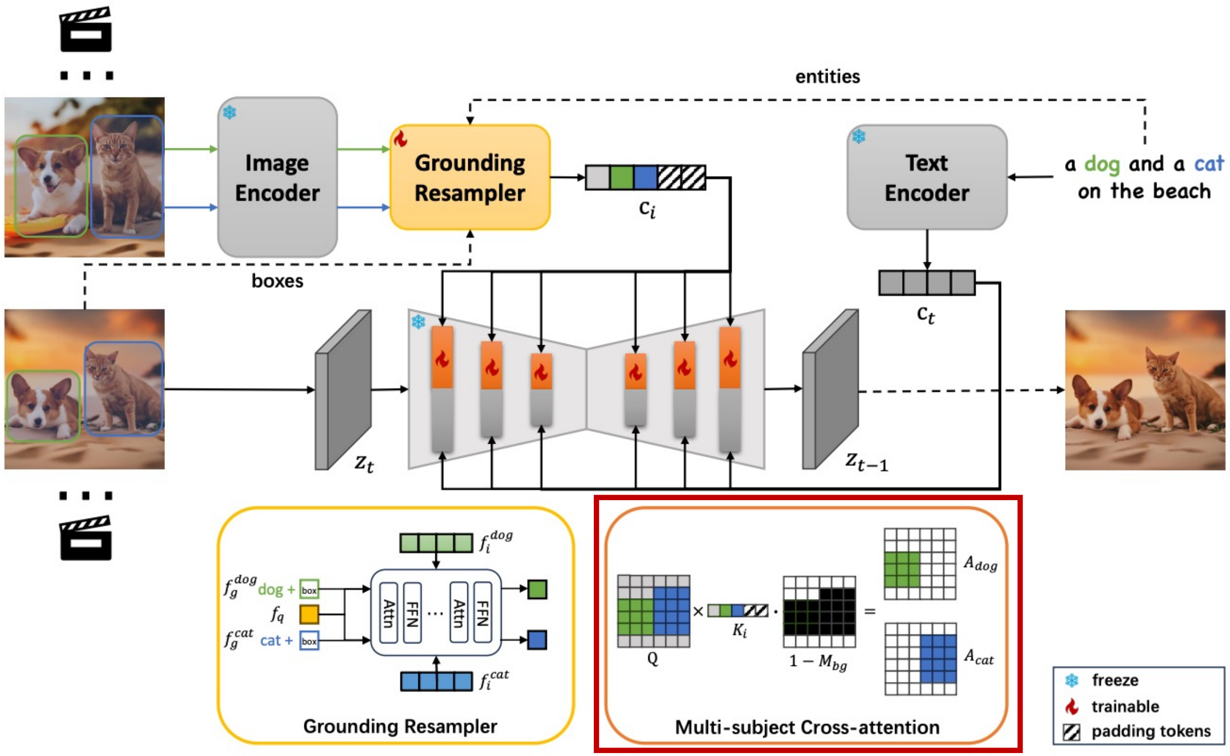
- Dataset construction for paired data
- **Grounding resampler for detailed image features**
  - Utilize a set of learnable tokens to distill pertinent information from image features

$$\mathrm{RSAttn} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\left(f_q\right)\mathbf{K}\left(\left[f_i, f_q\right]\right)}{\sqrt{d}}\right)\mathbf{V}\left(\left[f_i, f_q\right]\right)$$ where f_i is image embedding and f_q is learnable query



Main Idea: Separately extract image features of each subject with paired data

- Dataset construction for paired data
- Grounding resampler for detailed image features
- **Multi-subject cross-attention**
  - Attention mask to minimize discordance subject and background (or among subjects)



MS-Diffusion [Wang et al., 2024]

- **Generates images preserving each identities w/o being affected**

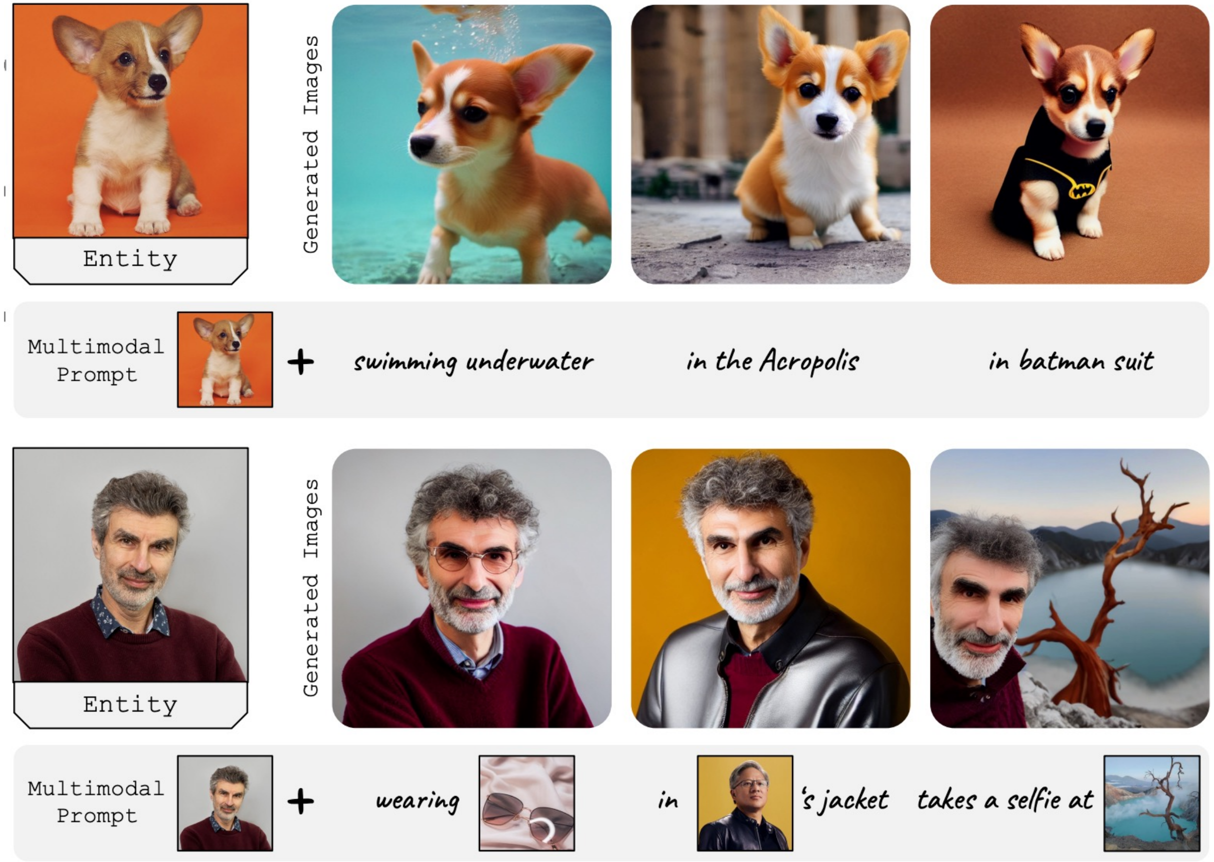
- **Generates images that adhere to layout conditions even with same categories**

- **Enables integrating different control conditions (e.g., depth, canny edge)**

KOSMOS-G: Generating Images in Context with Multimodal Large Language Models [Pan et al., 2024]

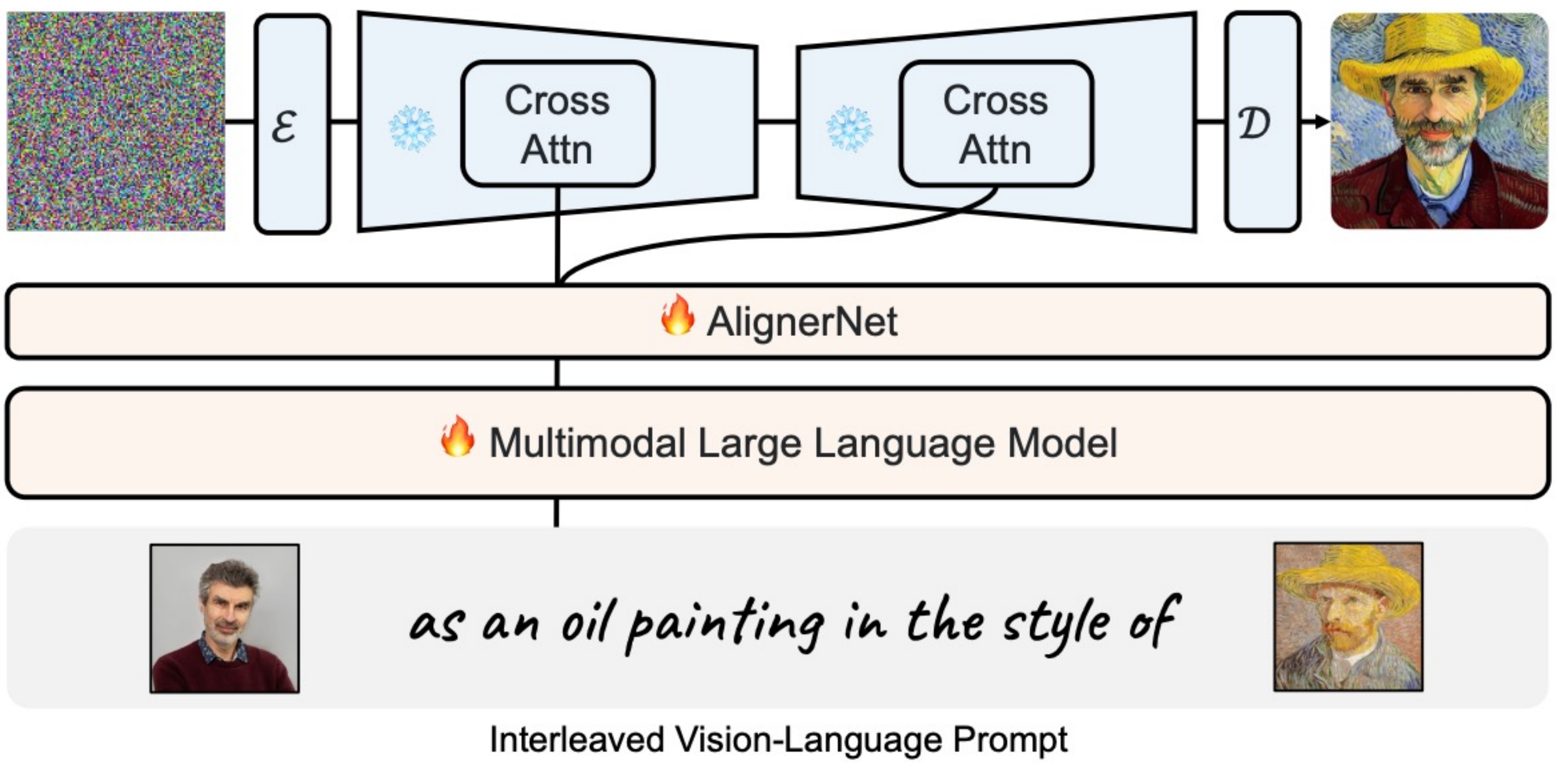
Motivation: Prior works cannot accept interleaved multi-image and text input

Contribution: Subject-driven generation leveraging MLLMs



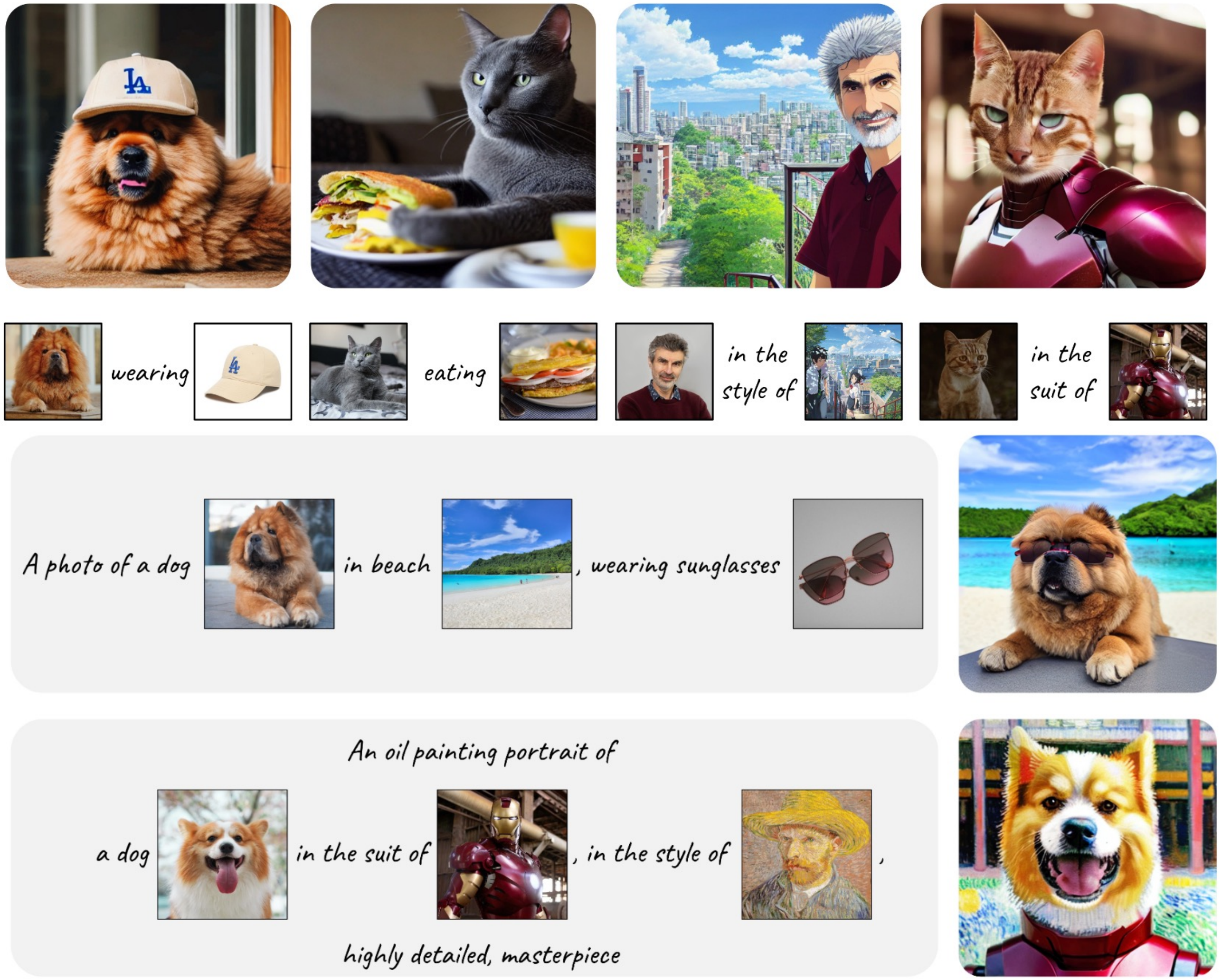
Main Idea: Interleaved multi-image and text input via MLLMs (align before instruct)

- **Multimodal language modeling:** pretrain MLLM on multimodal corpora, …
- **Image decoder aligning:** align output space to image decoder's input space
- **Instruction tuning:** fine-tune through a compositional generation task



KOSMOS-G [Pan et al., 2024]

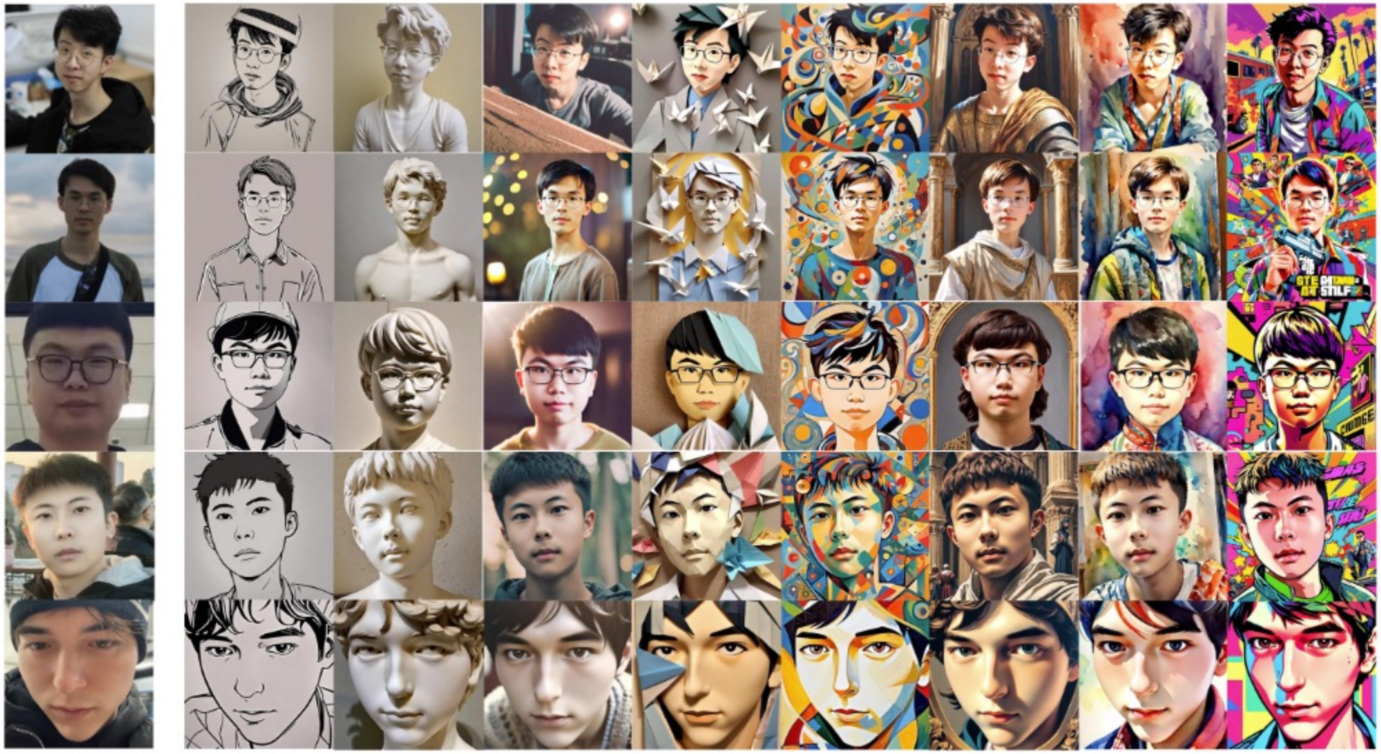
- **Enable image generation in various contexts (e.g., re-contextualization, stylization)**
  - w/ instruction based multi-image and text input



InstantID : Zero-shot Identity-Preserving Generation in Seconds [Wang et al., 2024]

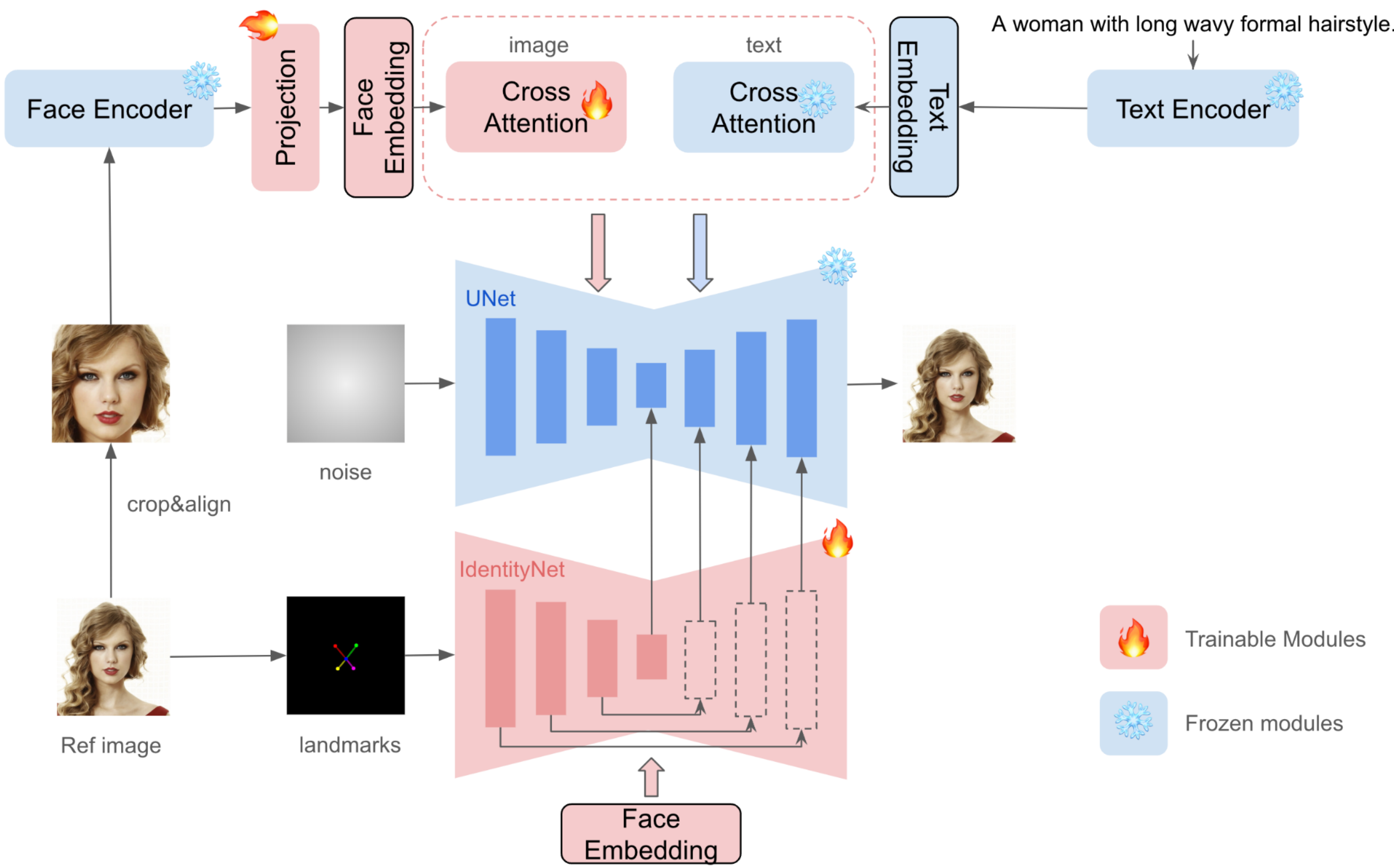
Motivation: Previous methods require extensive fine-tuning and lack compatibility with pre-trained models.

Contribution: Plug-and-play module for identity preserving generation especially on facial images.



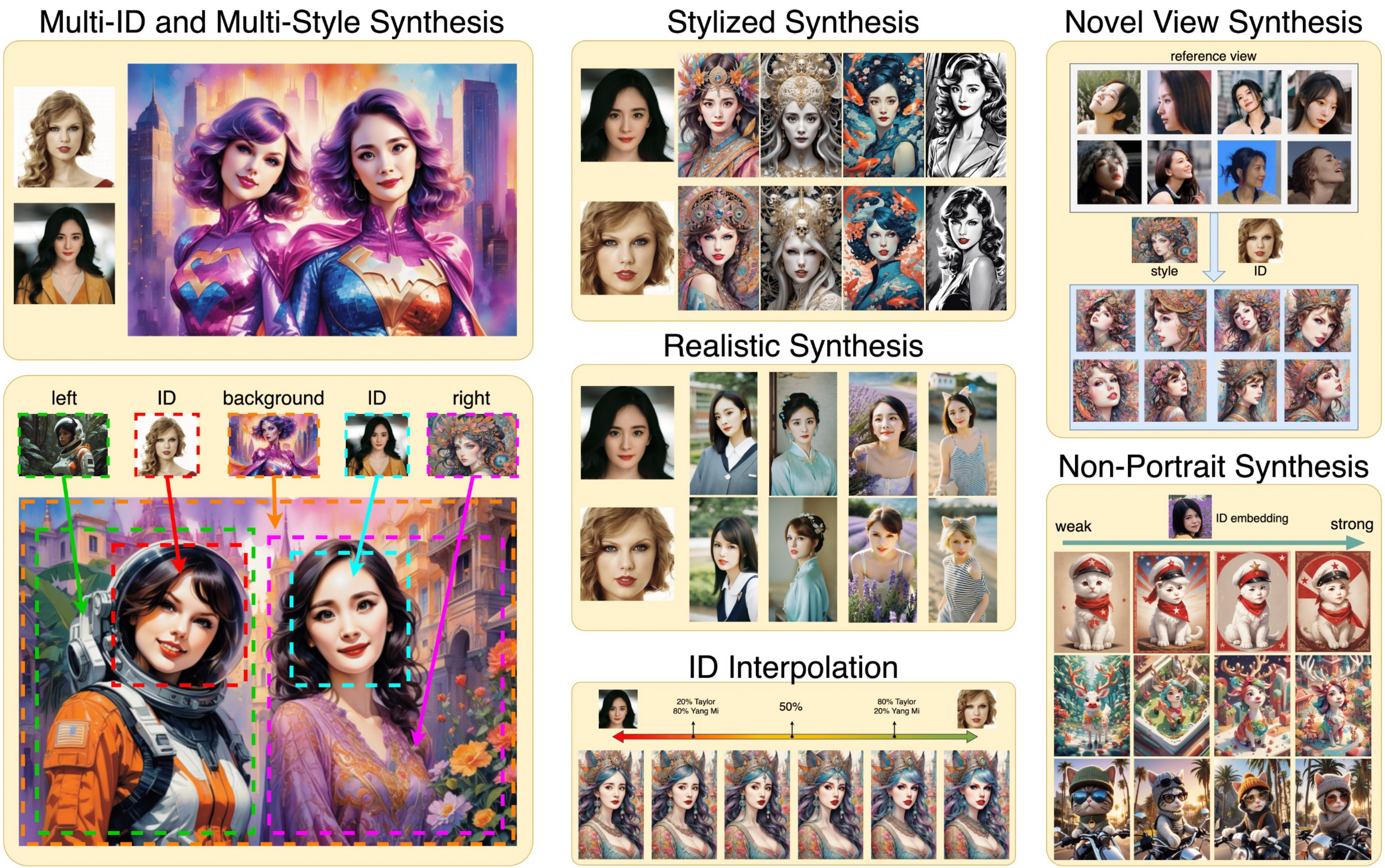
Main Idea: Introduce a variant of ControlNet for high-fidelity facial image generation.

- IdentityNet encodes details of reference facial images with spatial control.
- Decoupled cross-attention ensures text-based control over image generation.



InstantID [Wang et al., 2024]

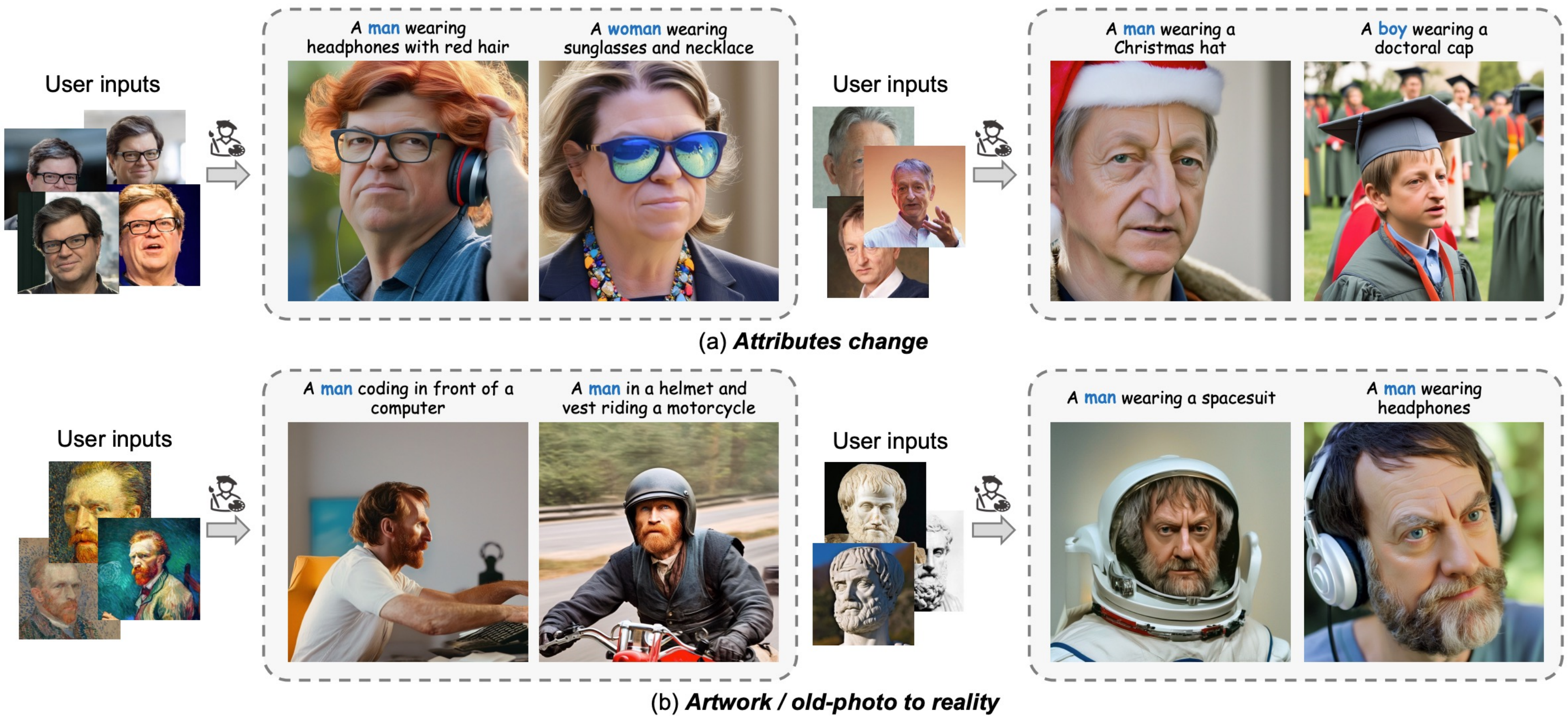
- **Enable diverse image generation with face images, faithfully preserving identities.**



PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding [Li et al., 2024]

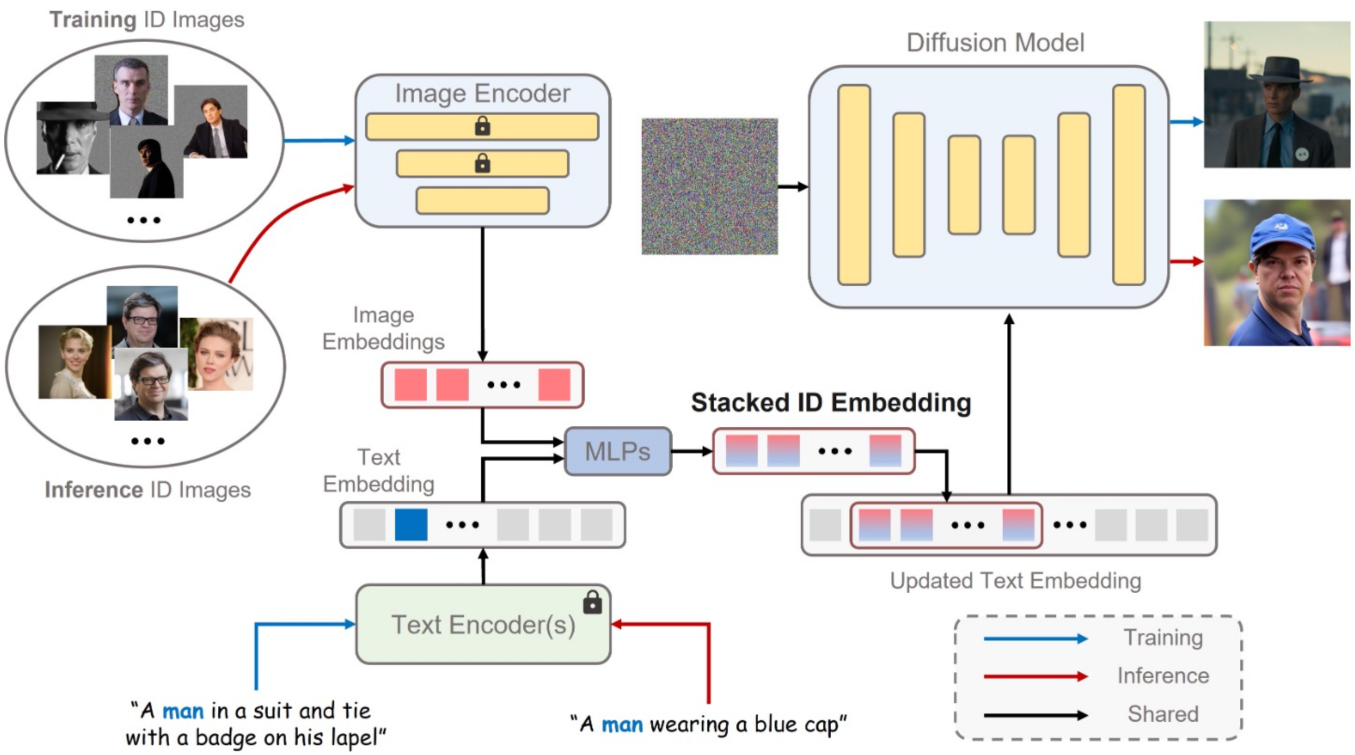
Motivation: Facial image generations lack of identity fidelity, text controllability an d efficiency.

Contribution: PhotoMaker ensures identity preservation, text prompt fidelity, and efficient personalized facial image generation.



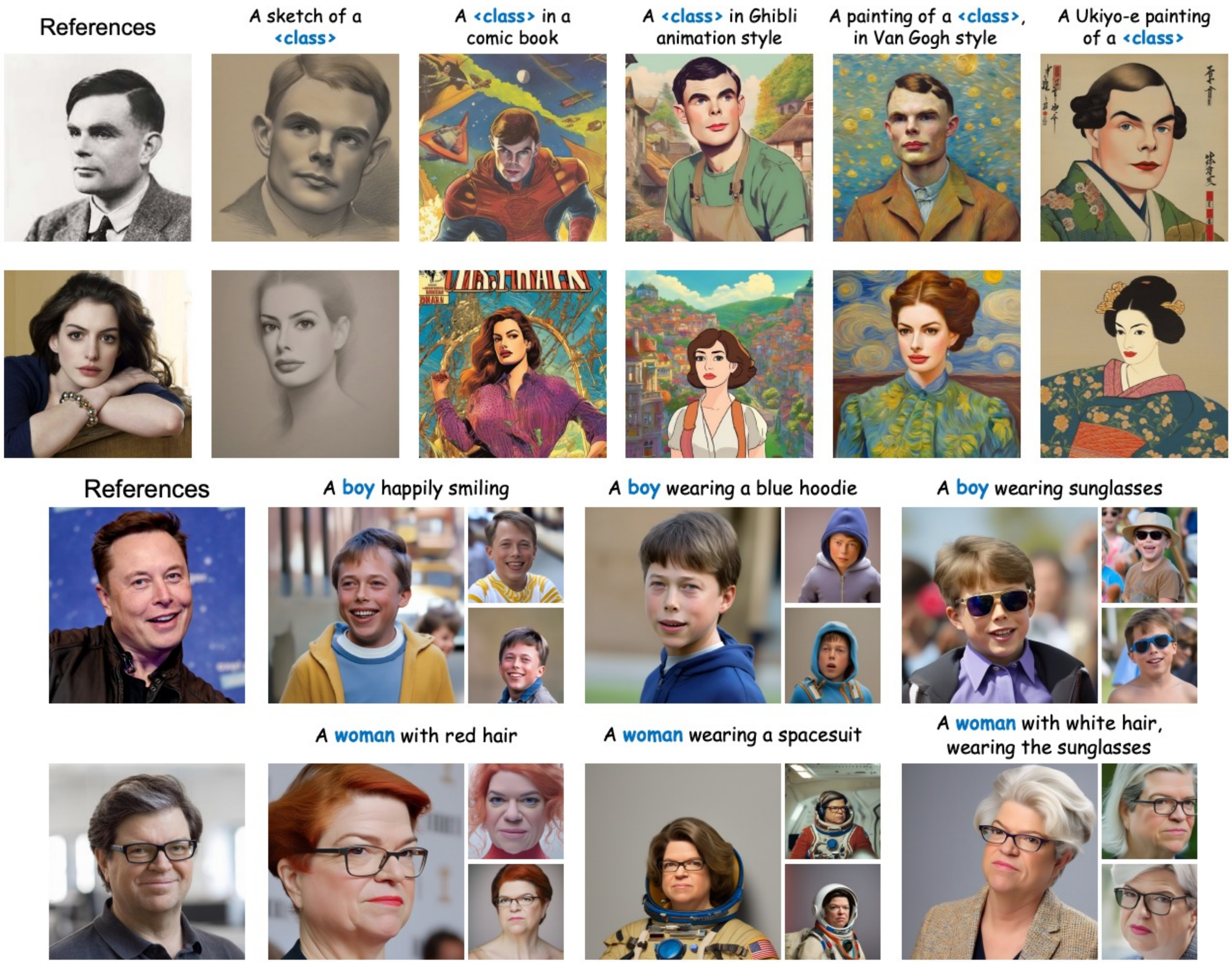
Main Idea: Construct a dataset for training and Exploit a few input images for high identity fidelity.

- Constructs a high-quality dataset through a meticulous data collection and filtering pipeline.
- Use a two-layer MLP to fuse ID features and class embeddings for an overall representation of human portrait.



PhotoMaker [Li et al., 2024]

- **Enable diverse image generation with face images, faithfully preserving identities.**
  - w/ text prompt for controllable image generation.



Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

Adding Conditional Control to Text-to-Image Diffusion Models [Zhang et al., 2023]

Motivation: Challenges in **additional control** on the text-to-image diffusion models

- Text prompt is not enough for matching mental imagery; need trial-and-error cycles
- Lack of data: Available data for a specific condition is small (e.g., human pose)

Contribution: End-to-end way that learns **conditional controls**

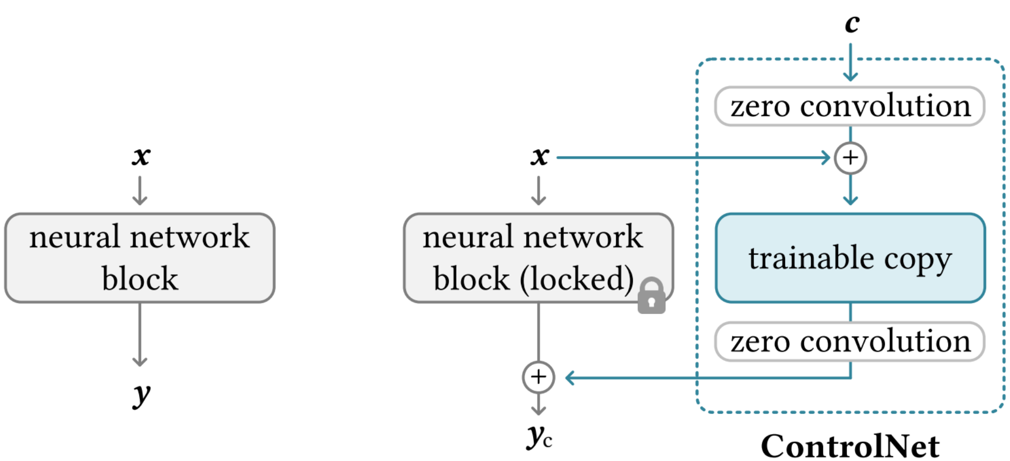
- while preserving the quality and capabilities of the large model

Main Idea: End-to-end neural network with **trainable copy** and **locked copy**

- Trainable copy: Cloning of the neural network block for task-specific dataset
- Locked copy: Preserve the capability of large-scale model

Effect of **zero convolution**:

- Reduce number of trainable parameters
- Elimination of harmful noise in training



Zero convolution
1×1 convolution layer
with zero weights and bias

$$\boldsymbol{y}_{\mathrm{c}} = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{\mathrm{z1}}); \Theta_{\mathrm{c}}); \Theta_{\mathrm{z2}})$$

: locked copy   : trainable copy

Main Idea: End-to-end neural network with **trainable copy** and **locked copy**

- Trainable copy: Cloning of the neural network block for task-specific dataset
- Locked copy: Preserve the capability of large-scale model

Effect of **zero convolution**:

- Reduce number of trainable parameters
- Elimination of harmful noise in training

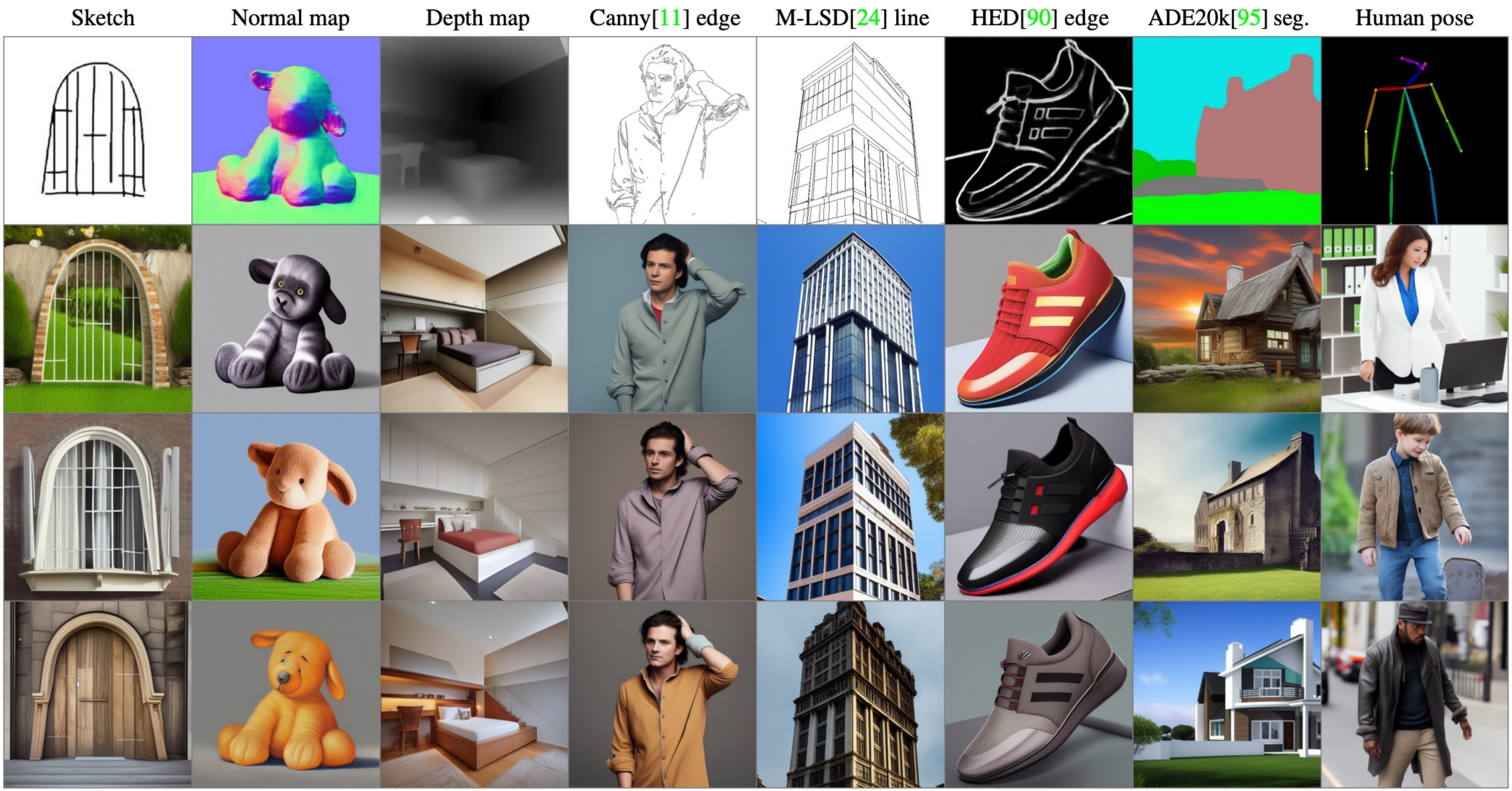
Training: **Fine-tune** the entire diffusion model with **ControlNet**

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f})) \|_2^2 \right]$$

: text prompt

: task-specific condition

ControlNet robustly interprets content semantics in diverse input conditioning



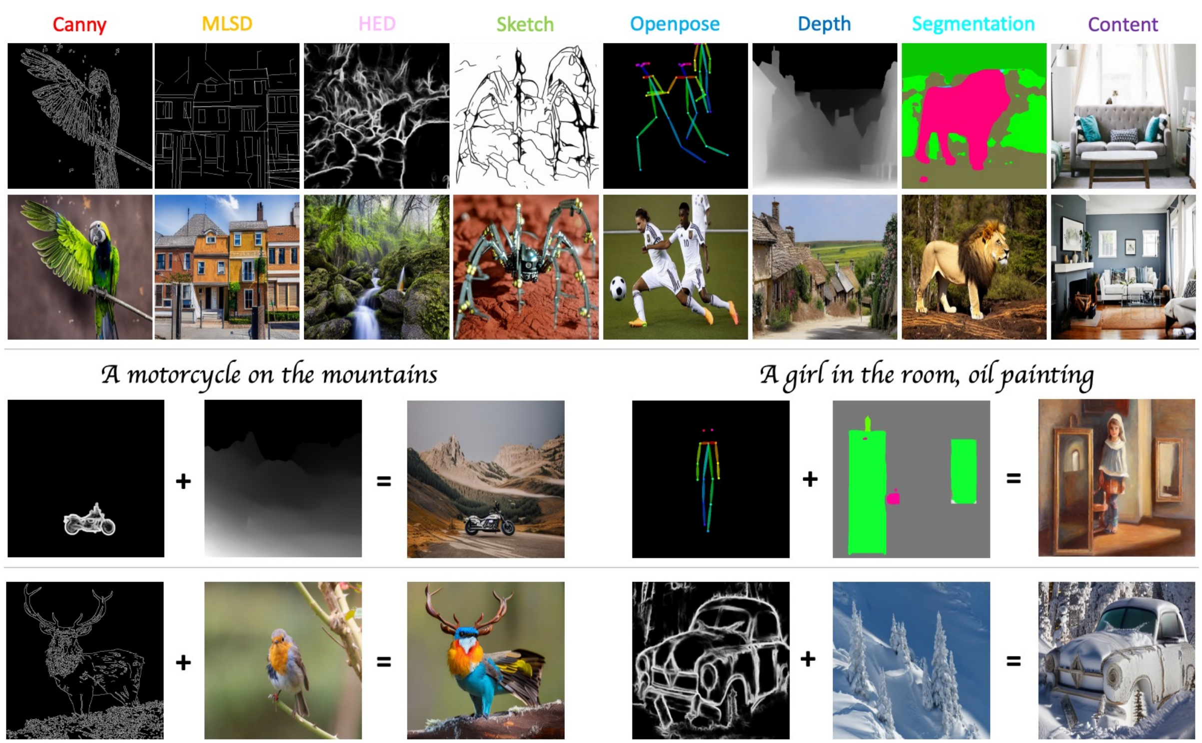
All-in-One Control to Text-to-Image Diffusion Models [Zhao et al., 2023]

Motivation: N ControlNets should be trained for N different conditions

- ControlNet only learn one kind of conditioning, requiring training each separately
- ControlNet can only accept one kind of conditioning at test-time

Contribution: Adapter-based generalizable ControlNet

- Learn any conditioning with same weights, and generate w/ more than 1 condition



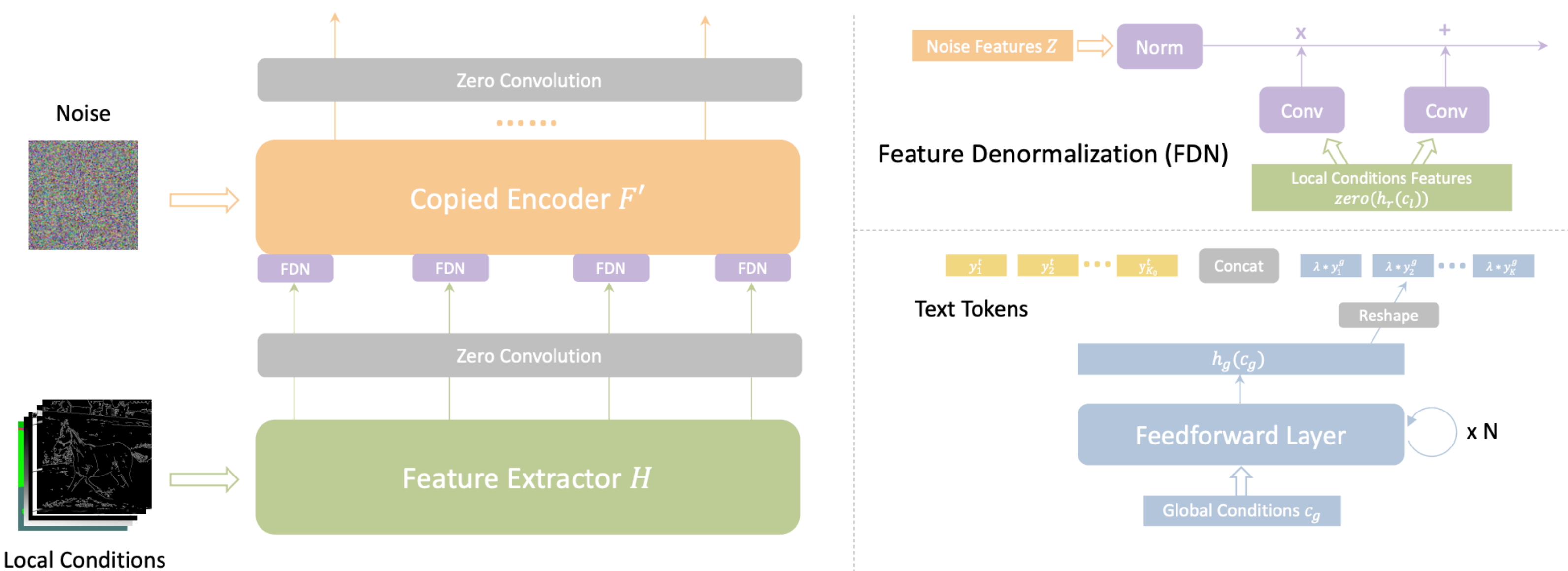
Uni-ControlNet [Zhao et al., 2023]

Main Idea: Use two adapters, 1) local and 2) global control adapter

- Local control adapter: Fine spatial control (e.g., edge maps, depth map)
- Global control adapter: CLIP image embedding

Difference from ControlNet:

- Local control adapter uses multi-resolution conditioning
- Composing local control: simply concatenating works very well



Local Control Adapter | Global Control Adapter

Main Idea: Use two adapters, 1) local and 2) global control adapter

- Local control adapter: Fine spatial control (e.g., edge maps, depth map)
- Global control adapter: CLIP image embedding

Difference from ControlNet:

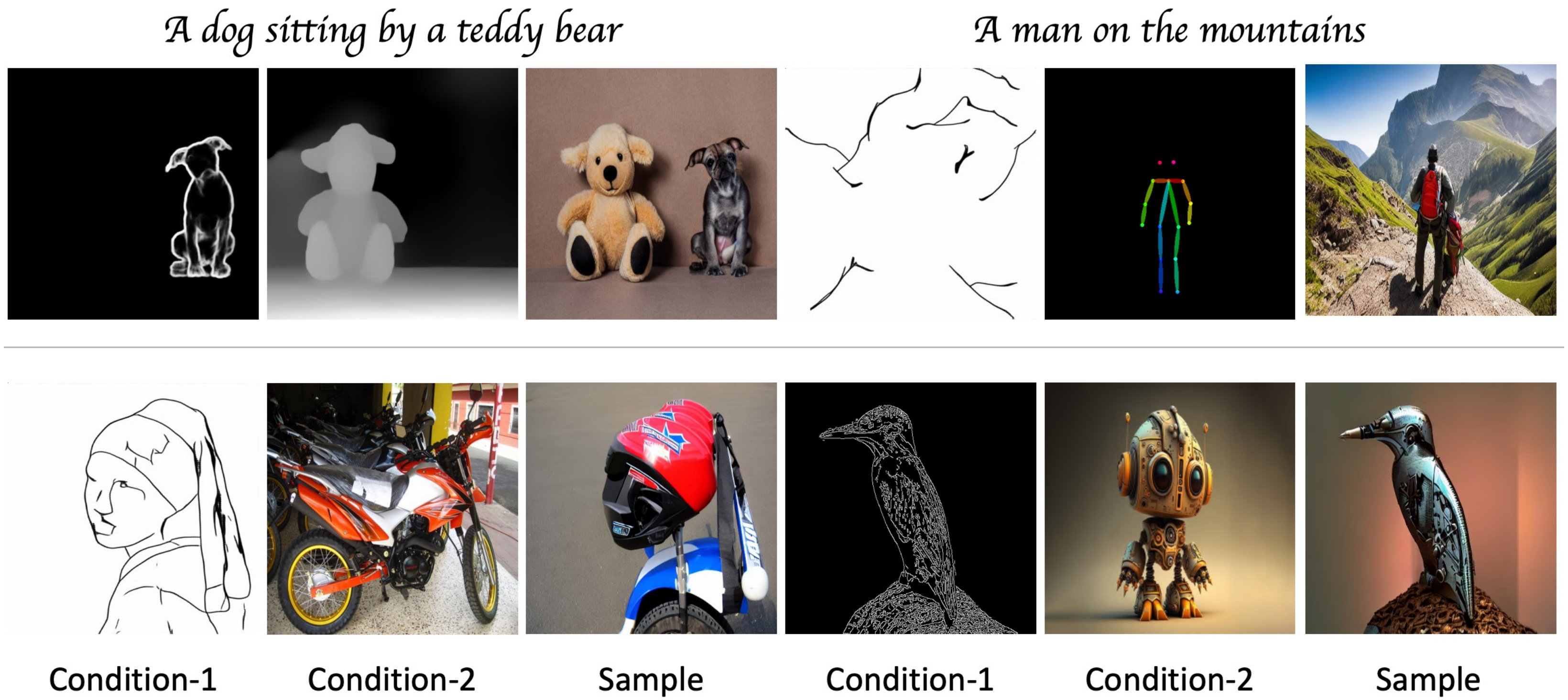
- Local control adapter uses multi-resolution conditioning
- Composing local control: simply concatenating works very well

Training strategy:

- When using both local and global control adapter, global guidance can dominate
  - This leads to insufficient local adapter training
- **Solution:** Drop each condition with some probability in each training step

Uni-ControlNet [Zhao et al., 2023]

Uni-ControlNet effectively generalizes ControlNet to be able to learn multiple number of conditioning with same weights, and to accept multiple conditioning at test-time



Recaptioning, Planning, and Generating with Multimodal LLMs [Yang et al., 2024]

Motivation: T2I models poorly handle **lengthy, complex prompts** with multiple objects

Contribution: **Planning-based training-free** T2I generating/editing framework

- MLLM splits prompt into smaller sub-prompts for region-wise generation

SDXL

DALL-E 3

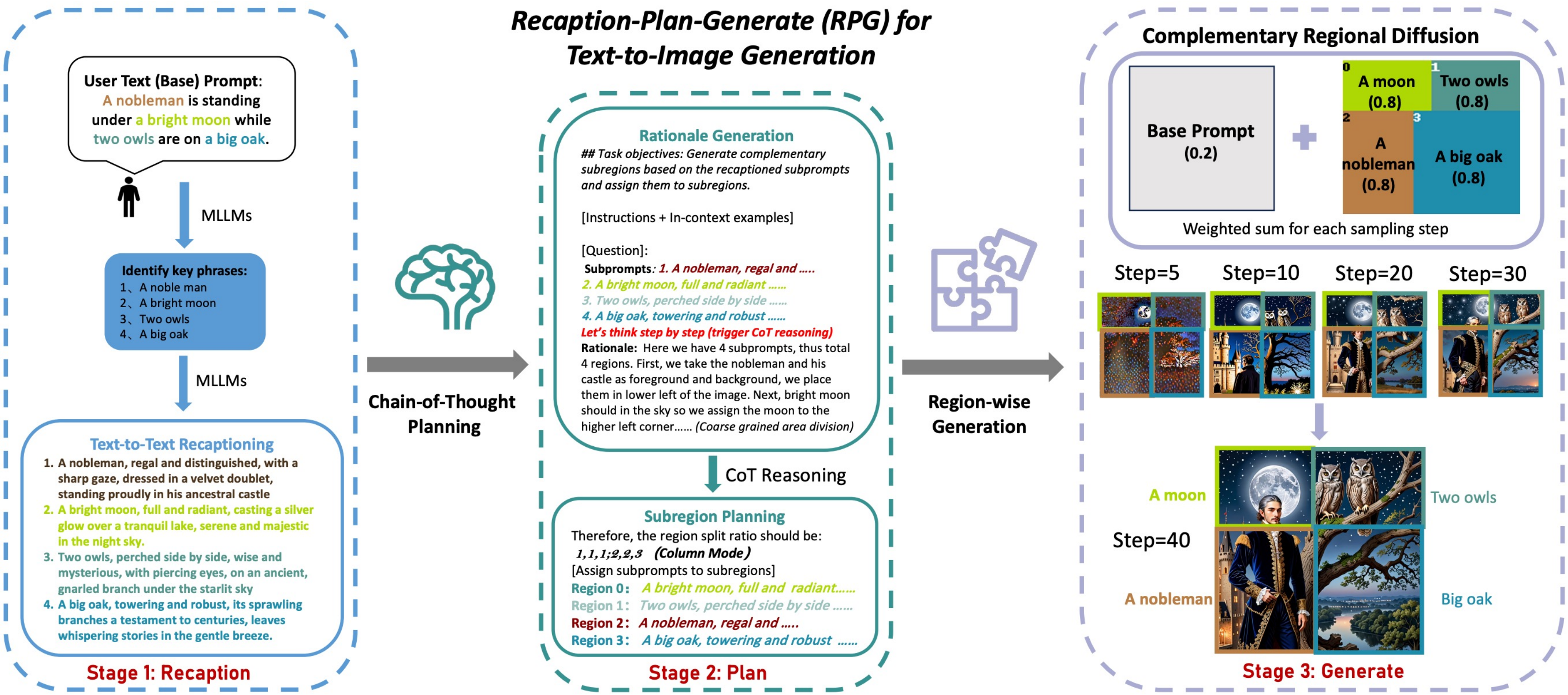
RPG (Ours)

Prompt: A **green twintail** girl in **orange dress** is sitting on the sofa while **a messy desk** in under a big window on the left, while a **lively aquarium is on the top right of the sofa**, realistic style.

RPG-Master [Yang et al., 2024]

Main Idea: Divide prompt into small regions using MLLM, and combine regions

- Subprompt generation: MLLM splits a given complex prompt into key pieces
- Complementary regional diffusion: Prompt-weighted denoising for each region

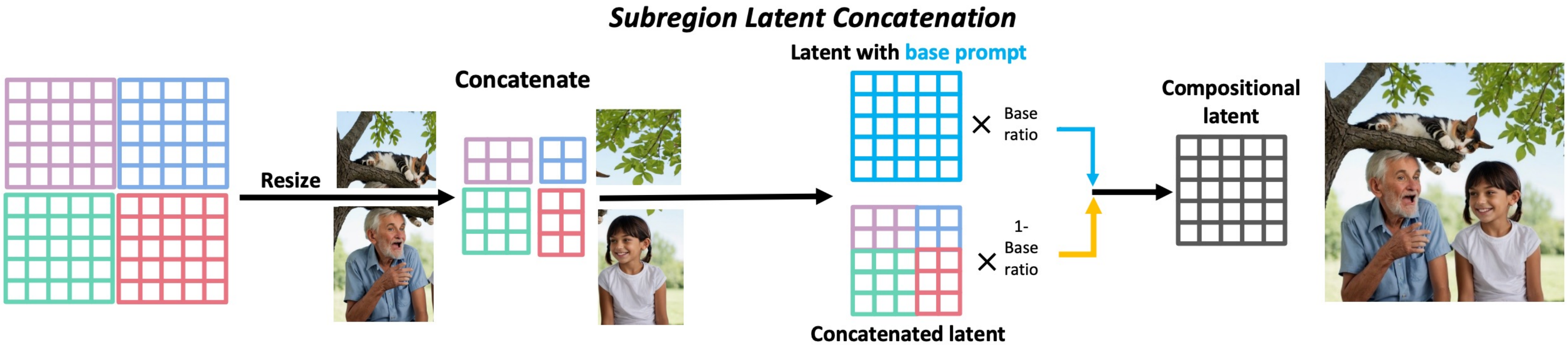


Main Idea: Divide prompt into small regions using MLLM, and combine regions

- Subprompt generation: MLLM splits a given complex prompt into key pieces
- Complementary regional diffusion: Prompt-weighted denoising for each region

Sampling using complementary regional diffusion:

- Resize each latent, concatenate, combine with latent from base prompt
- Denoise using the compositional latent



Main Idea: Divide prompt into small regions using MLLM, and combine regions

- Subprompt generation: MLLM splits a given complex prompt into key pieces
- Complementary regional diffusion: Prompt-weighted denoising for each region

Sampling using complementary regional diffusion:

- Resize each latent, concatenate, combine with latent from base prompt
- Denoise using the compositional latent

Example of region division: Spatial ratios planned by MLLMs

Example

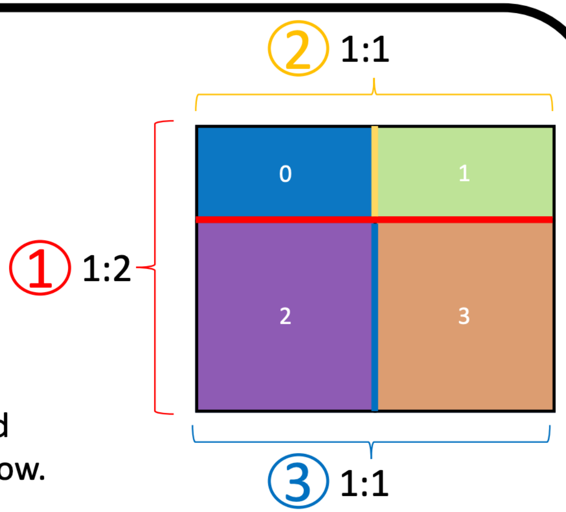
Split ratio: 1,1,1; 2,1,1

1. First, we split the image into two rows where horizontal split ratio is 1:2

2. Second, we split the first row into two columns, where vertical split ratio is 1:1, and we get region 0 and region 1

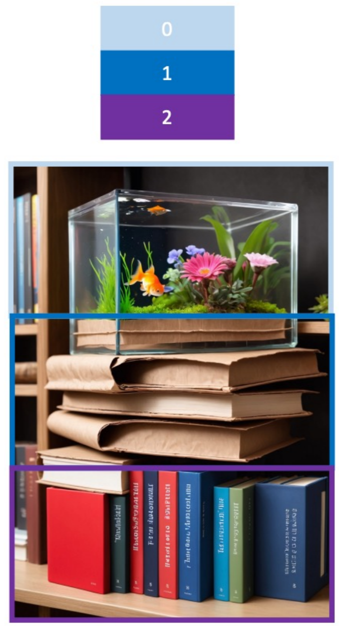
3. Finally, we split the second row into two columns, where vertical split ratio is also 1:1, and we get **region 2** and **region 3**

Rules: We start splitting from the top left corner of the image, and the region is numbered from the top left region to right, row by row.



RPG-Master generates images containing multiple objects with different attributes and relationships flawlessly, powered by LLM-based spatial planning

A large bookshelf with five floors and six compartments with lively aquarium on the top left,and plant in terrarium on the top right,the Books with ancient kraft paper covers in the second and third floors,newly printed books including red and blue books in the fourth and fifth floors.



Base prompt : A large bookshelf with three floors and six compartments with books, lively aquarium, and plant in terrarium

Region 0: small lively aquarium with goldfish and sea weed in the compartment of the bookshelf, delicate flowers in the terrarium in the compartment of the bookshelf

Region 1: Books with ancient kraft paper covers in the compartment of the bookshelf

Region 2: Some new books including red covers and blue covers in the compartment of the bookshelf

Base prompt : A large bookshelf with three floors and six compartments with books, lively aquarium, and plant in terrarium

Region 0: small lively aquarium with goldfish and sea weed in the compartment of the bookshelf,

Region 1: delicate flowers in the terrarium in the compartment of the bookshelf

Region 2: Books with ancient kraft paper covers in the compartment of the bookshelf

Region 3: Books with ancient kraft paper covers in the compartment of the bookshelf

Region 4: Some new books including red covers and blue covers in the compartment of the bookshelf

Region 4: Some new books including red covers and blue covers in the compartment of the bookshelf

Diffusion Self-Guidance for Controllable Image Generation [Epstein et al., 2023]

Motivation: Text prompts are not sufficient to specify spatial relationships of objects

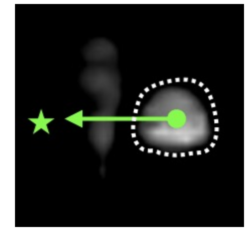
Contribution: Zero-shot controllable generation by manipulating attention maps

- Object position, size, shape can be modified by changing attention maps
- Training-free method

Main Idea: Attention map control while sampling for spatially controlled generation

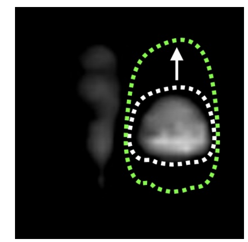
- Object position: Modify the centroid of the attention channel

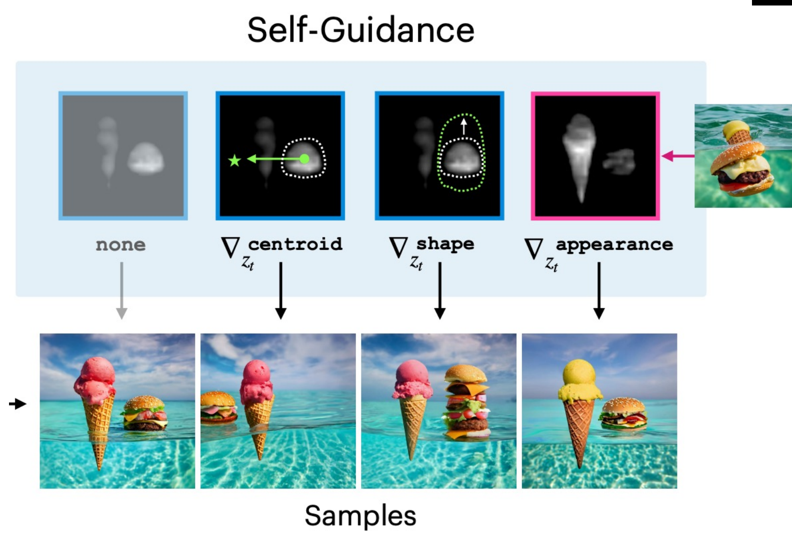
$$\texttt{centroid}\,(k) = \frac{1}{\sum_{h,w} \mathcal{A}_{h,w,k}} \begin{bmatrix} \sum_{h,w} w \cdot \mathcal{A}_{h,w,k} \\ \sum_{h,w} h \cdot \mathcal{A}_{h,w,k} \end{bmatrix}$$



- Object size: Modify the sum of the attention channel

$$\texttt{size}\,(k) = \frac{1}{HW} \sum_{h,w} \mathcal{A}_{h,w,k}$$

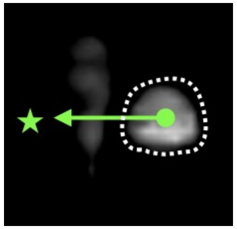




Main Idea: Attention map control while sampling for spatially controlled generation

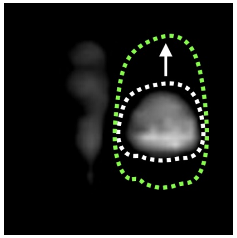
- Object position: Modify the centroid of the attention channel

$$\texttt{centroid}\,(k) = \frac{1}{\sum_{h,w} \mathcal{A}_{h,w,k}} \begin{bmatrix} \sum_{h,w} w \cdot \mathcal{A}_{h,w,k} \\ \sum_{h,w} h \cdot \mathcal{A}_{h,w,k} \end{bmatrix}$$



- Object size: Modify the sum of the attention channel

$$\texttt{size}\,(k) = \frac{1}{HW} \sum_{h,w} \mathcal{A}_{h,w,k}$$



- These equations do not necessarily modify just one object. Can we control one object specifically?

These equations do not necessarily modify just one object. Can we control one object specifically?

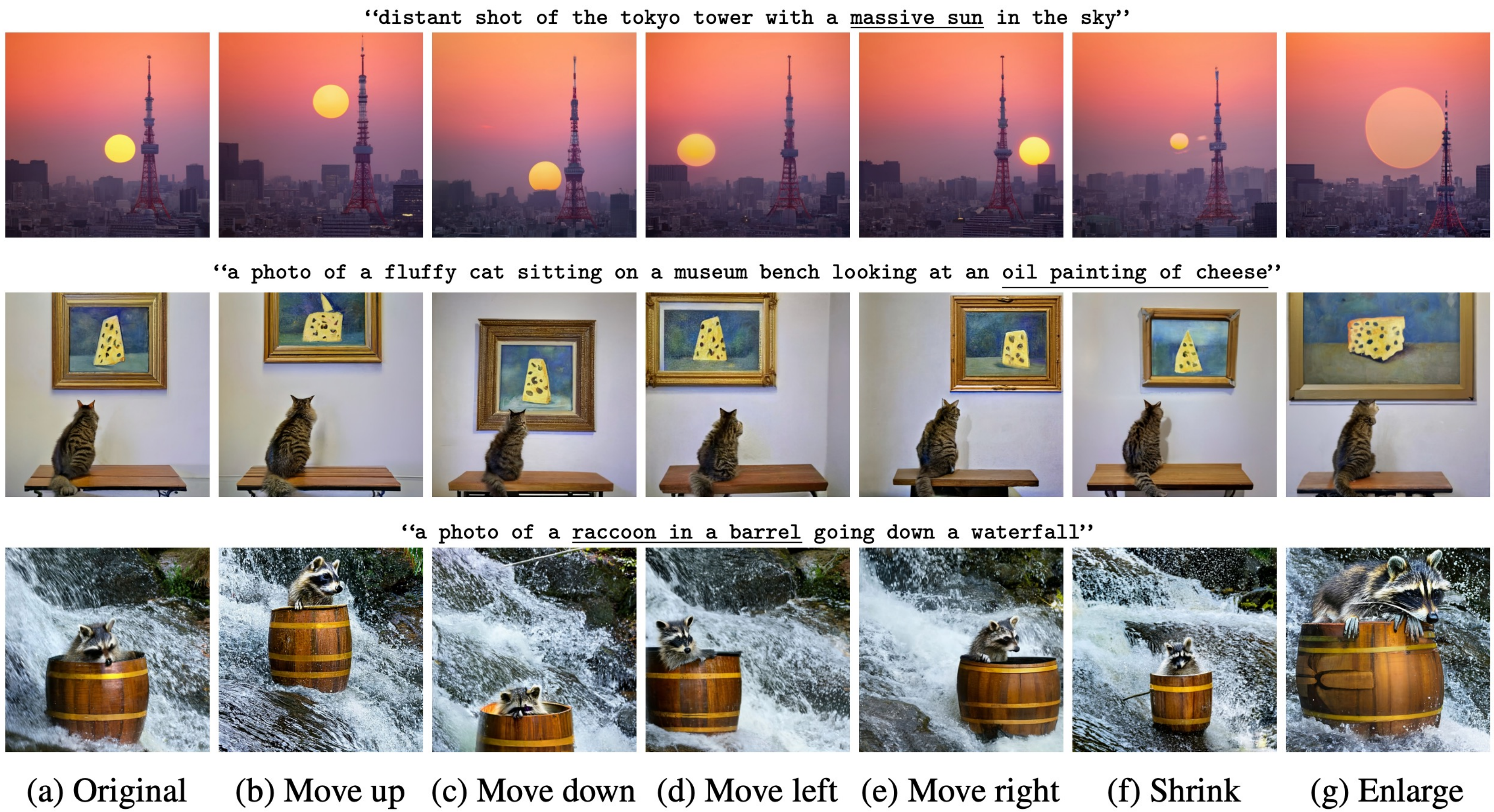
Solution: Just fix all other objects using these equations and change the desired one

$$g = w_0 \overbrace{\frac{1}{|O|-1} \sum_{o \neq o_k \in O} \frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} \|\texttt{shape}_{i,t,\text{orig}}(o) - \texttt{shape}_{i,t}(o)\|_1}^{\text{Fix all other object shapes}}$$

$$+ w_1 \overbrace{\frac{1}{|O|} \sum_{o \in O} \|\texttt{appearance}_{t,\text{orig}}(o) - \texttt{appearance}_t(o)\|_1}^{\text{Fix all appearances}}$$

$$+ w_2 \overbrace{\frac{1}{|\mathcal{A}|} \sum_{i=0}^{|\mathcal{A}|} \|\mathcal{T}\left(\texttt{shape}_{i,t,\text{orig}}(o_k)\right) - \texttt{shape}_{i,t}(o_k)\|_1}^{\text{Guide } o_k\text{'s shape to translated original shape}}$$

Self-guidance is able to provide significant control over the spatial aspect of generation simply by directly modifying the attention channel of the internal representations of a diffusion model



Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

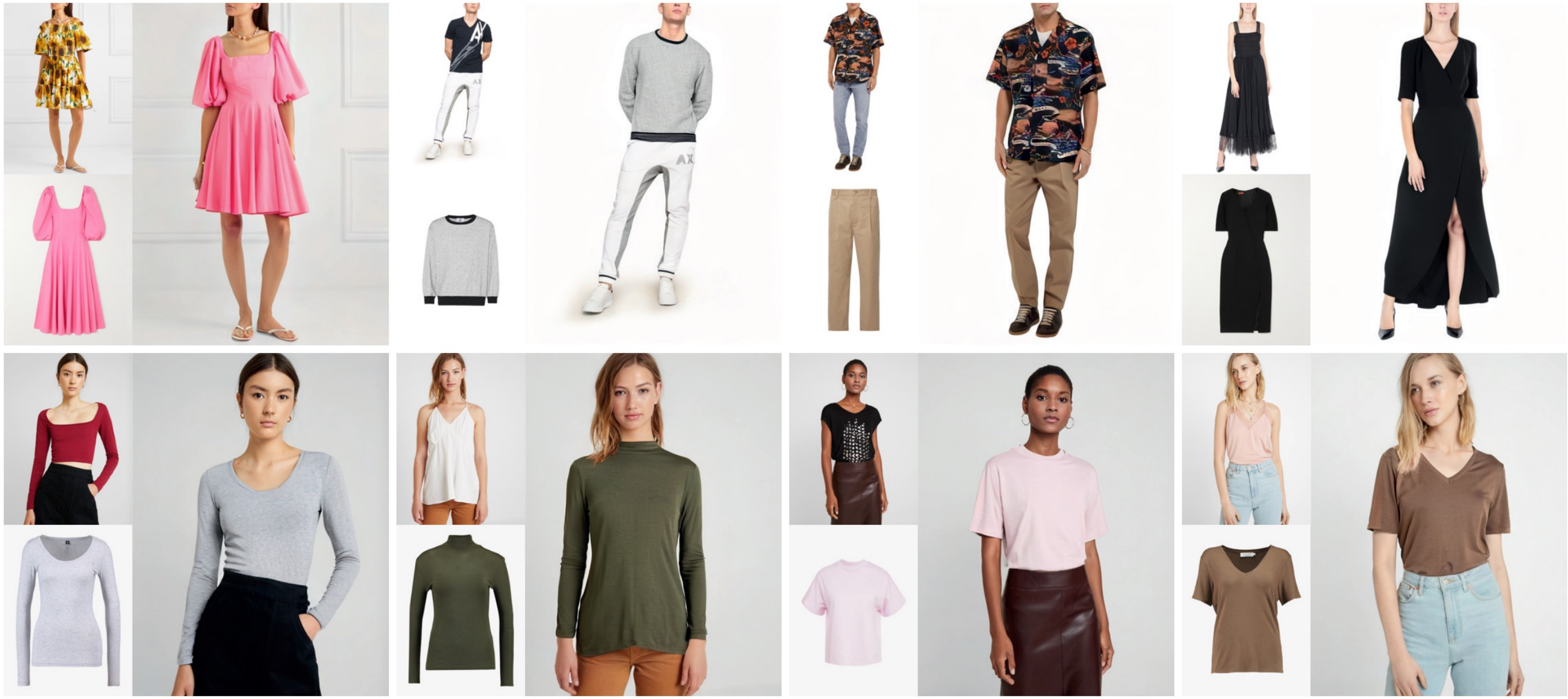
- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

Latent Diffusion Textual-Inversion Enhanced Virtual Try-On [Morelli et al., 2023]

Motivation: **Leverage diffusion models** to generate natural try-on images.

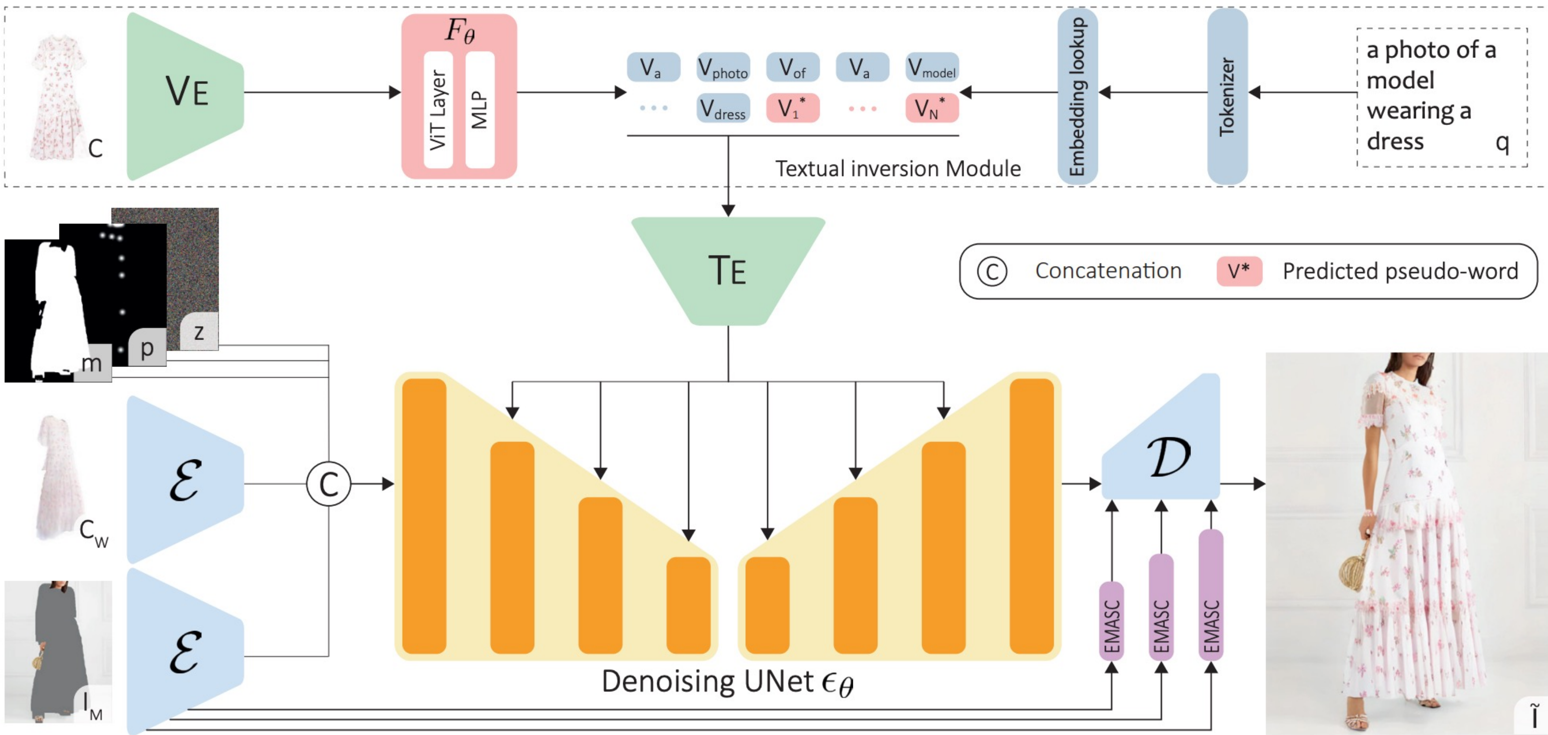
- Prior works employ **GAN**, which fails to produce realistic images.
- Consider virtual try-on task as an exemplar-based image inpainting.

Contribution: First work to utilize **diffusion models** for **virtual try-on** task.

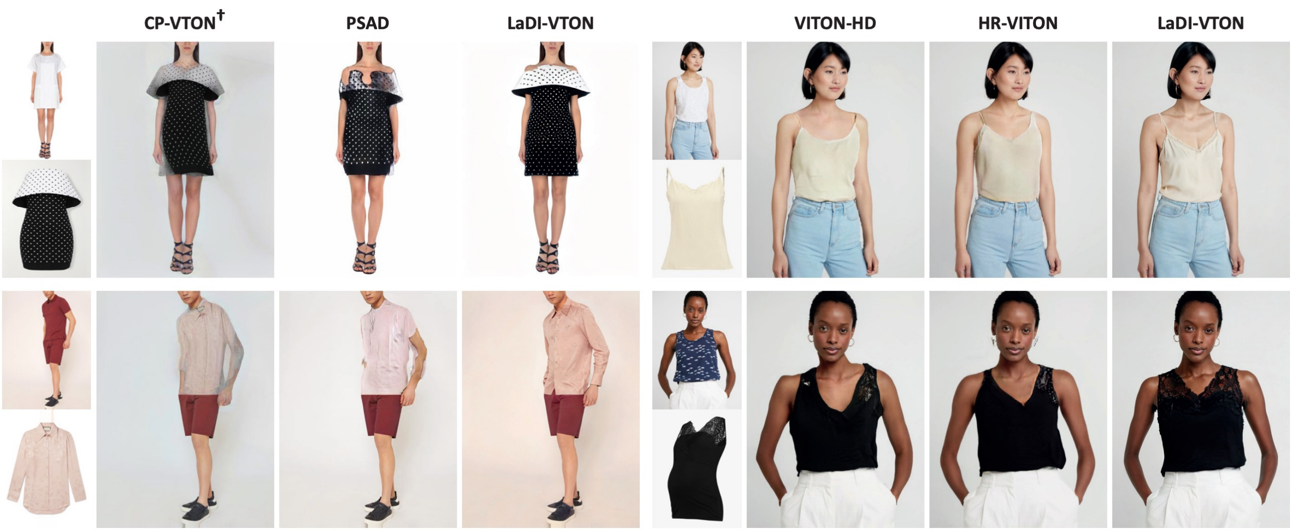


Main Idea: Utilize **CLIP embedding** space for **garment conditioning**.

- Frozen image encoder (CLIP) to extract garment features.
- Utilize text prompt (e.g., a photo of a model wearing {category}) to exploit T2I prior.
- Introduce small module to prevent distortion outside the mask region.



- Generates **natural images** comparing to **GAN-based** models.
- However, it **fails** to preserve **fine-details of garments,** since it use CLIP embeddings for garment conditioning.



<Qualitative results>

<Limitation>

Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On [Kim et al., 2023]

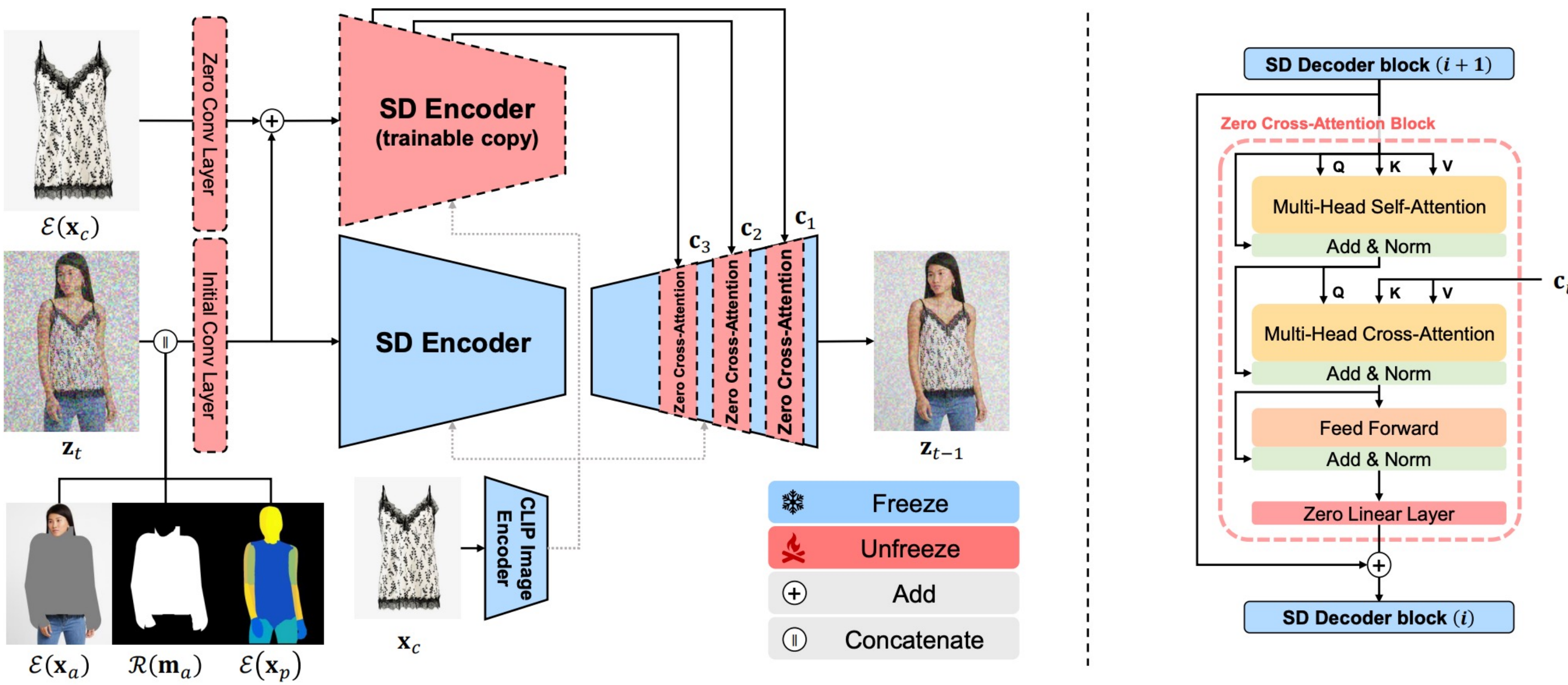
Motivation: Improve **garment encoding** with **controlnet-style encoder.**

- CLIP embedding is too coarse to fully encode garments.
- Introduce controlnet-style encoder for finer conditioning of garments.

Contribution: Improved garment encoding, which is a main challenge for diffusion-based Virtual Try-on.

Main Idea: Utilize **Controlnet-style SD Encoder** for **finer encodings** for garments.

- **SD Encoder** to encode garment with **more details**.
- Introduce an **auxiliary loss** to **prevent attention** from mapping to **multiple regions.**
- **Data augmentation** to enhance generalization capabilities.



StableVITON [KIM et al., 2023]

- Generates **natural images** with **preserving fine-details**.
- However, it still struggles to preserve **fine-details of garments,** especially with **in-the-wild** images**.**

<Qualitative results>

<Limitation>

IDM-VTON [Choi et al., 2024]

Improving Diffusion Models for Authentic Virtual Try-on in the Wild [Choi et al., 2024]

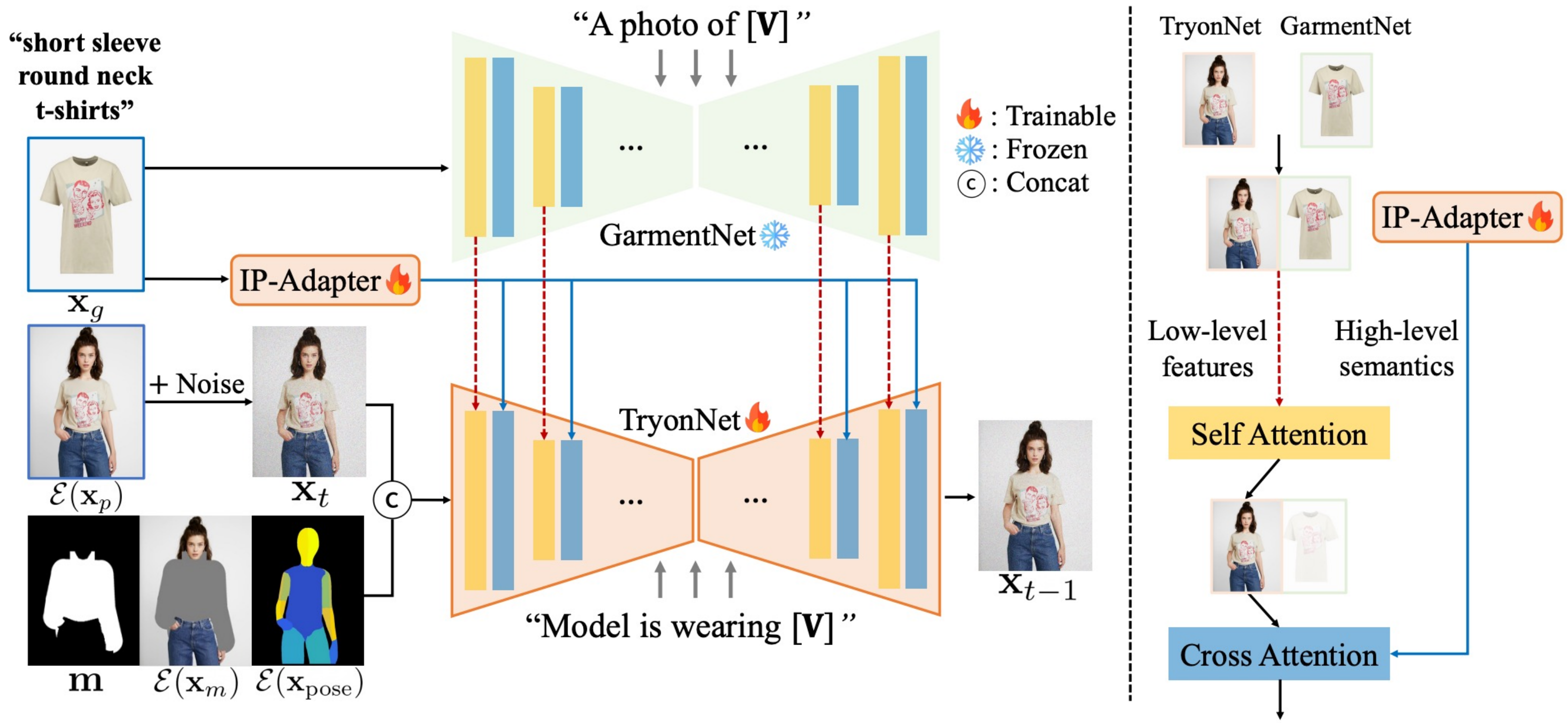
Motivation: Virtual Try-on for **in-the-wild** images, which is **more practical scenarios**.

- Decompose garment encoding with high and low-level features of garment.
- Customize network for particular garment by users.

Contribution: Authentic virtual try-on for **in-the-wild** scenarios.

Main Idea: Decompose **garment encoding** with **high** and **low-level features** of garment.

- **Parallel unet encoder** for **low-level** features.
- IP-Adapter for **high-level** semantics.
- Effective & efficient **fine-tuning network** for customizing it on particular garment.



IDM-VTON [Choi et al., 2024]

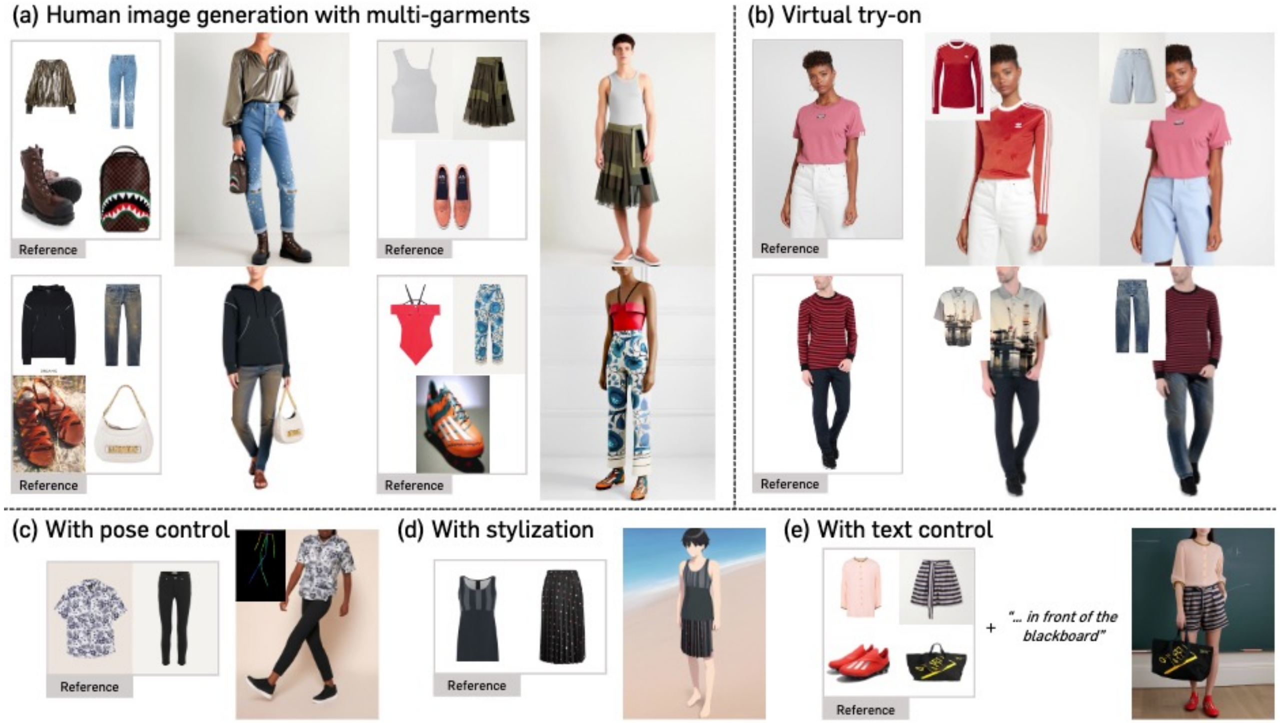
- Generates **natural images** with **preserving fine-details**.
- Demonstrates strong performance in **challenging in-the-wild** scenarios.

Controllable Human Image Generation with Personalized Multi-Garments [Choi et al., 2024]

Motivation: Collecting **paired data** of ***multiple*** **references** is **challenging**.

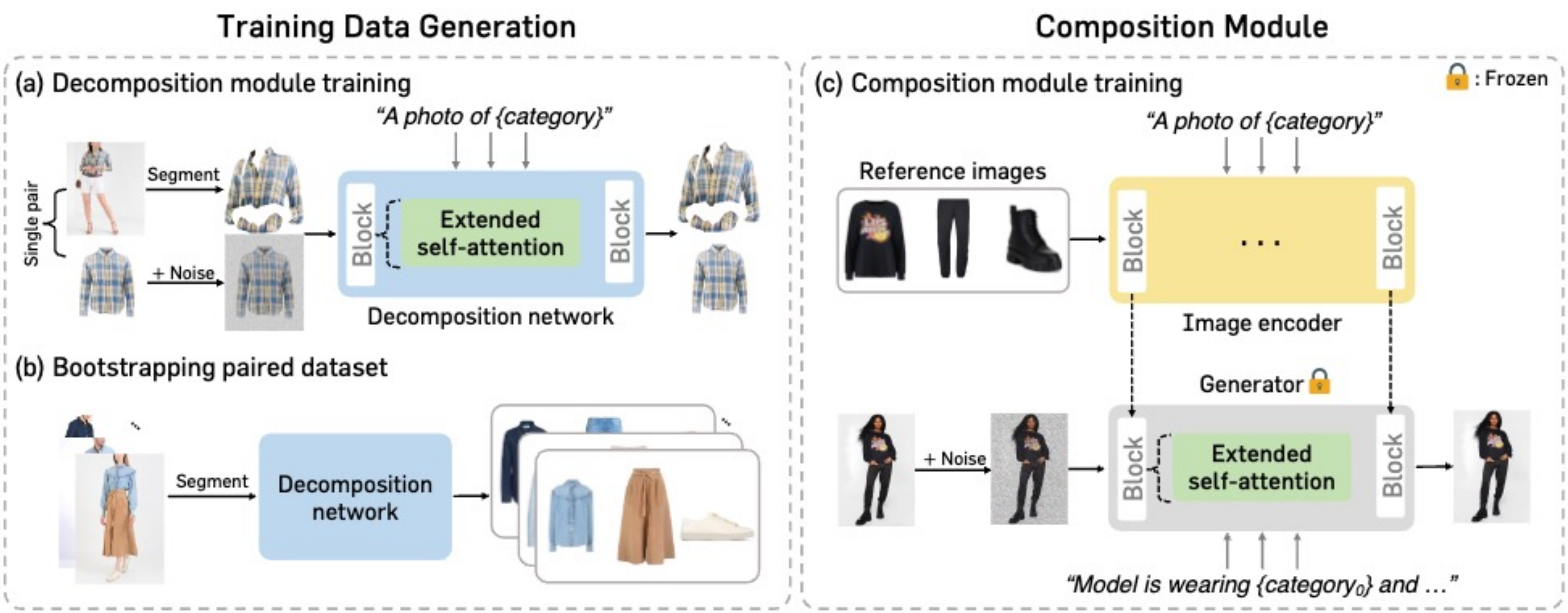
- Introduce a **synthetic paired data generation pipeline** for **multiple reference**.
- **Dual denoising path** for **composing** multiple reference garment.

Contribution: Synthetic data generation, effective for **controllable generation.**



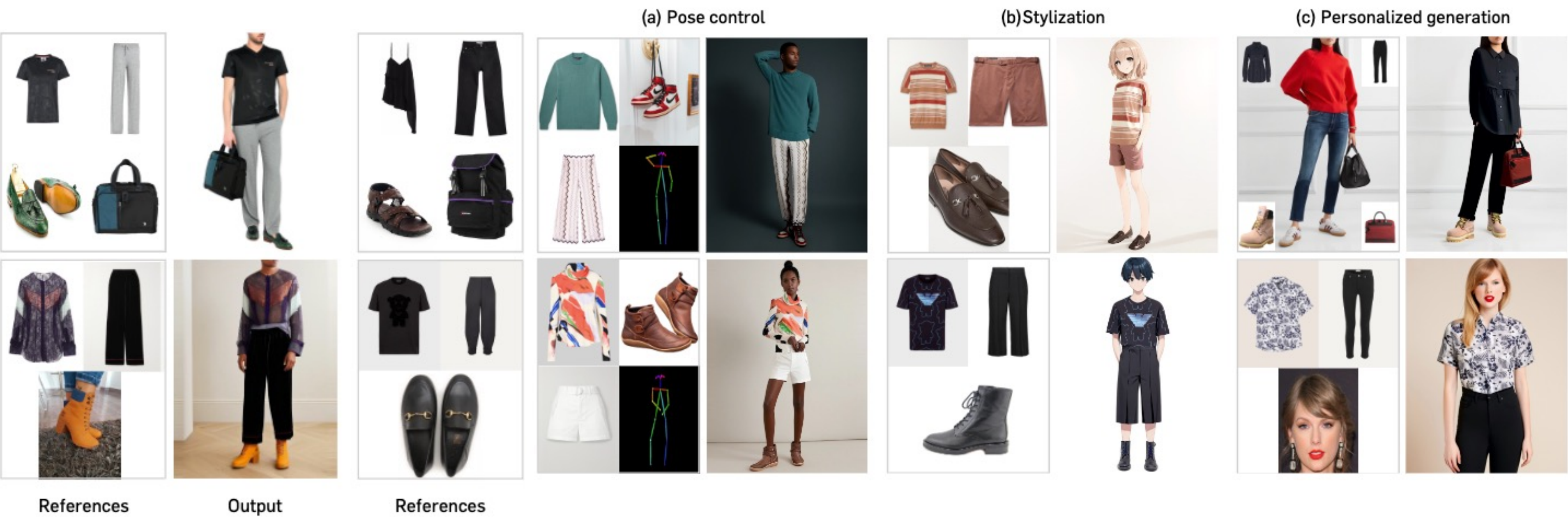
Main Idea: Synthesize **multiple paired data** by leveraging **single paired data**, which is easy to collect.

- **Decomposition** network, **mapping segmented garment** to **garment in product view**.
- Bootstrapping multiple paired data.
- **Filtering strategy** for high-quality data.
- **Composition module** for generating human images with multiple-garments.



BootComp [Choi et al., 2024]

- Generates **human images** with **multiple garments**.
- Diverse applications such as **pose control**, **stylization**, **virtual try-on**.



Due to the existence of large-scale pretrained T2I models, many following works focused on extending the capability beyond image generation

From now on, we explore recent topics in leveraging T2I models for

- Image editing (or image-to-image translation) using text
- Personalization
- Controllable generation
- Virtual try-on
- Text-to-3D generation

DreamFusion: Text-to-3D using 2D diffusion [Poole et al., 2023]

- Recent, Text-to-image (T2I) diffusion models have shown impressive capabilities
    - Synthesizing high-quality, realistic, diverse images with the text given as input

- How can we **utilize T2I diffusion models to 3D synthesis** without 3D training data?
- How can we **use DMs as a critic to optimize** the underlying 3D representation?

- Poole et al. (2023): **Score Distillation Sampling (SDS)**
    - Probabilistic density distillation enabling the use of a 2D diffusion models for priors

- **DreamFusion**: Optimize NeRF using T2I diffusion models with SDS
    - Optimize NeRF $g(\theta)$, that look like images $\mathbf{x}$ when rendered from random angles
    - The optimized NeRF yields good images appropriate for given text prompt
    - Does not require 3D training data and no modification to the image diffusion models

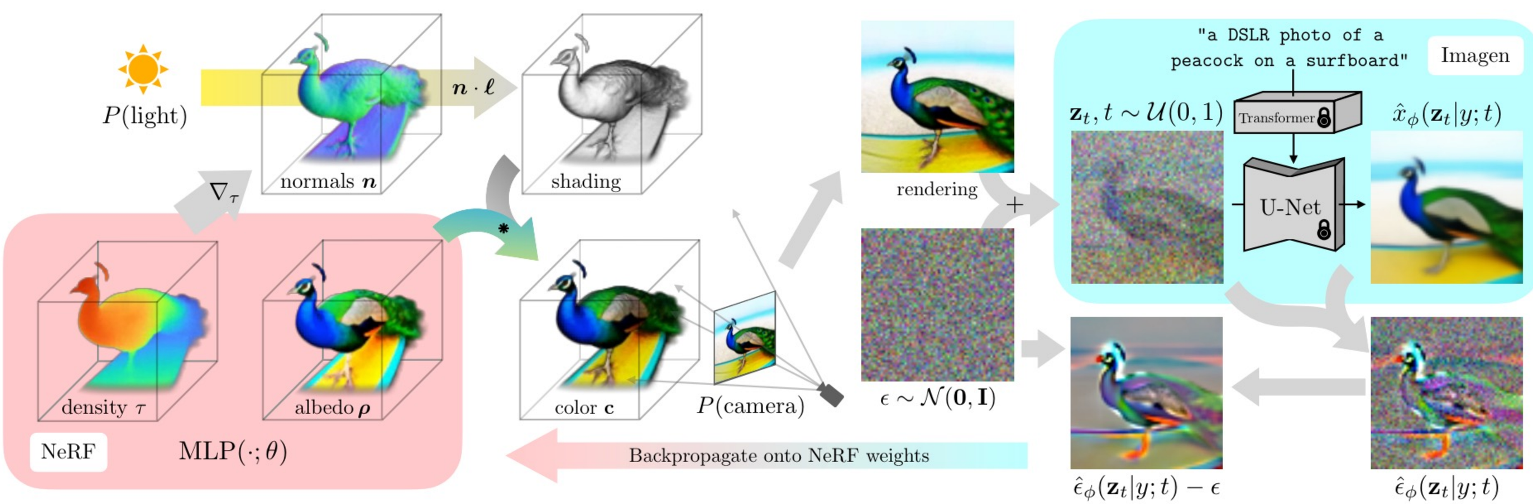
* source : Poole et al., DreamFusion: Text-to-3D using 2D diffusion, ICLR 2023

DreamFusion: Text-to-3D using 2D diffusion [Poole et al., 2023]

- **How does DreamFusion create 3D assets from text descriptions?**
  1. Initialization:
     NeRF is randomly initialized and trained from scratch for each caption
  2. NeRF parameter updates:
     DreamFusion diffuses the rendering and reconstructs it with a (frozen) Imagen

$$\underbrace{\hat{\epsilon}_\phi(\mathbf{z}_t|y;t)}_{\text{prediction of injected noise}} - \underbrace{\epsilon}_{\text{injected noise}}$$

  - Subtracting the injected noise produces a low variance update direction
  - Backpropagated through the rendering process to update the NeRF MLP parameters



* source : Poole et al., DreamFusion: Text-to-3D using 2D diffusion, ICLR 2023

- **Score distillation sampling enables sampling in parameter space, not pixel space**
  - create 3D models that look like good images when rendered from random angles

1. Training objective of diffusion models is as follows:

$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\alpha_t\mathbf{x} + \sigma_t\epsilon; t) - \epsilon\|_2^2\right]$$

2. Minimize the diffusion model training loss w.r.t a generated data point $\mathbf{x} = g(\theta)$

$$\theta^* = \text{argmin}_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta))$$

3. Gradient of the training objective becomes:

$$\nabla_\theta \mathcal{L}_{\text{Diff}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{t,\epsilon}\left[w(t)\underbrace{(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)}_{\text{Noise Residual}}\underbrace{\frac{\partial\hat{\epsilon}_\phi(\mathbf{z}_t; y, t)}{\mathbf{z}_t}}_{\text{U-Net Jacobian}}\underbrace{\frac{\partial\mathbf{x}}{\partial\theta}}_{\text{Generator Jacobian}}\right]$$

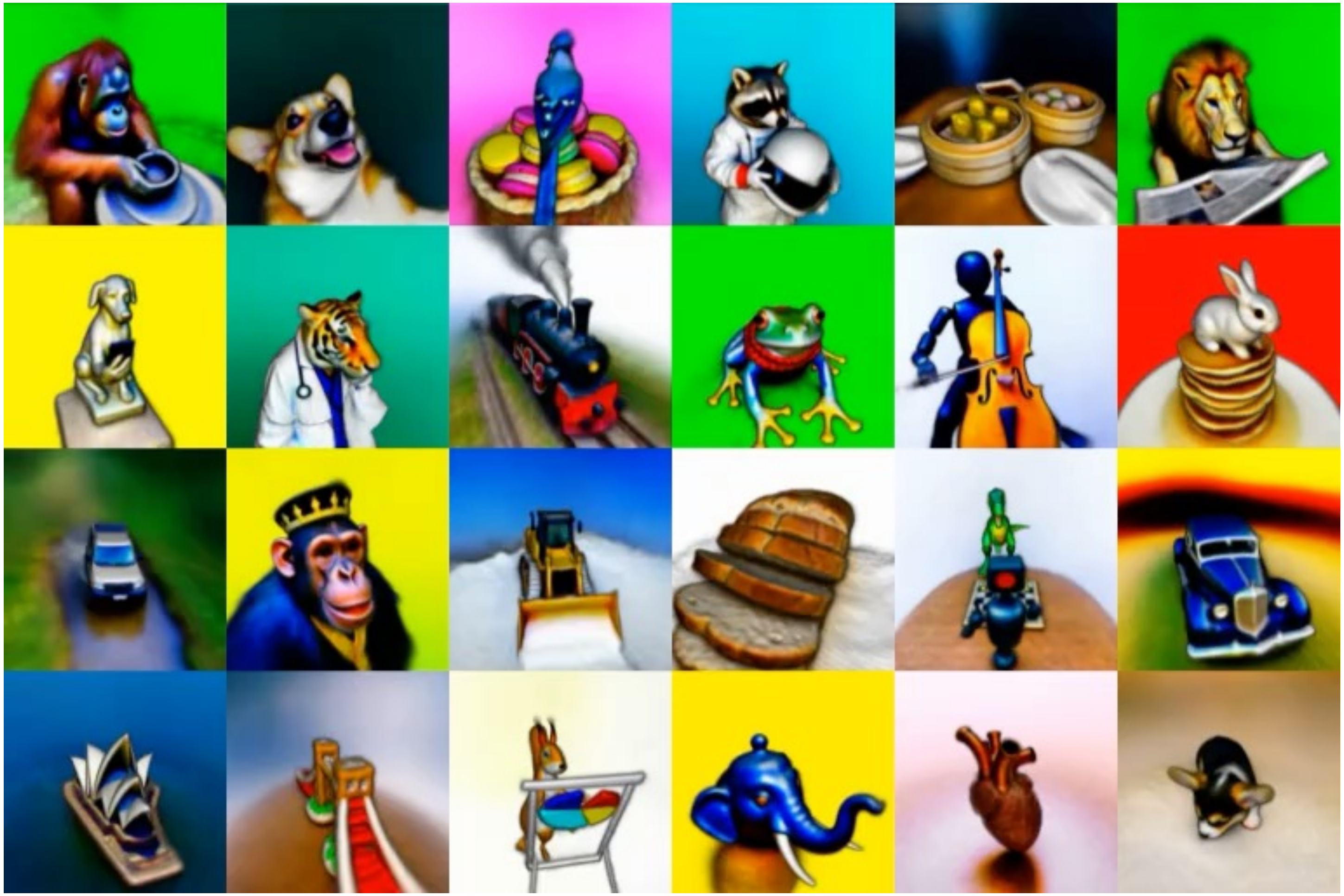
4. Score Distillation Sampling

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon)\frac{\partial\mathbf{x}}{\partial\theta}\right]$$

* source : Poole et al., DreamFusion: Text-to-3D using 2D diffusion, ICLR 2023

DreamFusion: Text-to-3D using 2D diffusion [Poole et al., 2023]

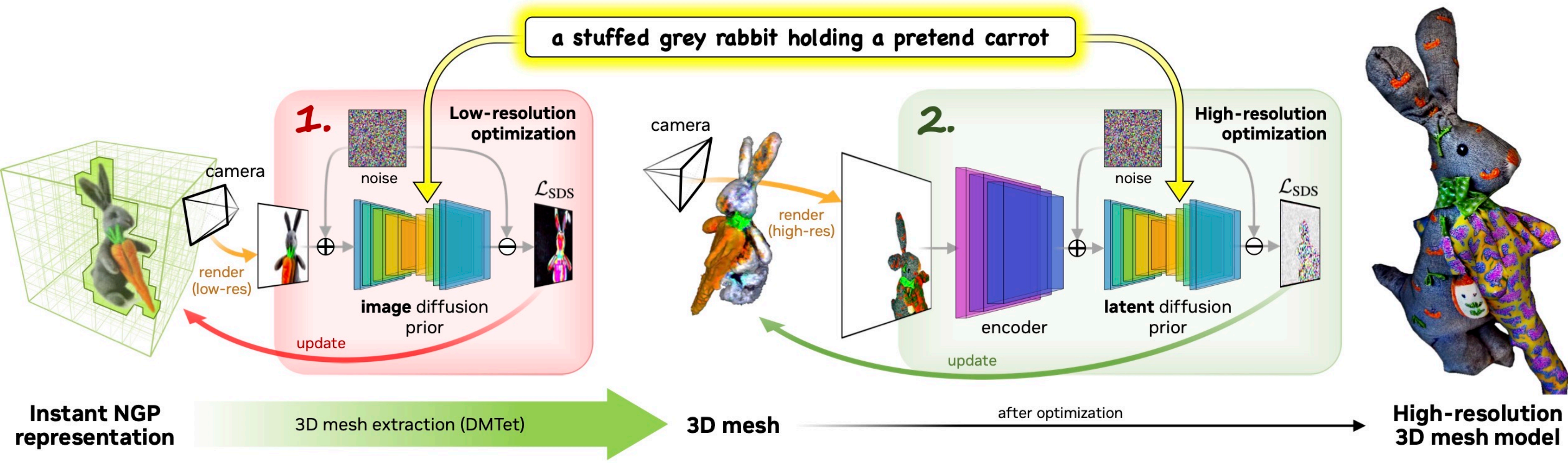
- DreamFusion generates coherent 3D scenes from a variety of text prompts



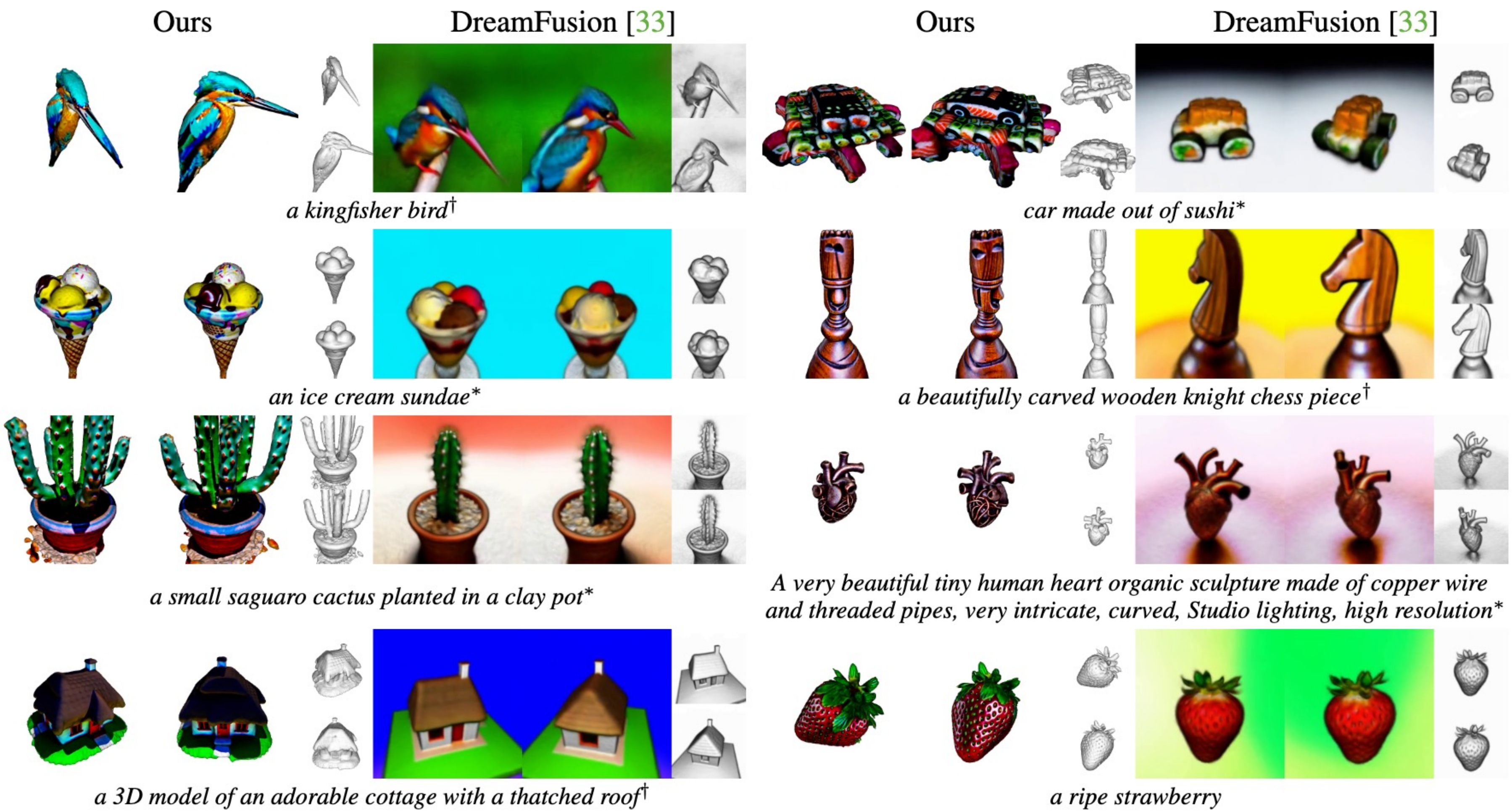
* source : Poole et al., DreamFusion: Text-to-3D using 2D diffusion, ICLR 2023

Magic3D [Lin et al., '23]

- DreamFusion is of low resolution (e.g., 64x64)
- Magic3D upscale text-to-3D model by two stage coarse-to-fine optimization
  - Stage 1. generate low-resolution NeRF using SDS
  - Stage 2. export to 3D mesh and use high-res. LDM for high-resolution 3D mesh



Comparison with DreamFusion



ProlificDreamer [Wang et al. '23]

- SDS suffers from over-saturated image because of high guidance scale
- ProlificDreamer resolves this problem by using variational score distillation (VSD)
  - ProlificDreamer generates high-quality text-to-3D model

Michelangelo style statue of dog reading news on a cellphone. | A pineapple. | A chimpanzee dressed like Henry VIII king of England. | An elephant skull.

Variational Score Distillation (VSD)

- VSD uses Wasserstein gradient flow of variational inference problem

$$q^* = \arg\min_q D_{\mathrm{KL}}(q \| p) \qquad \frac{dx_t}{dt} = \nabla \log p(x_t) - \nabla \log q_t(x_t)$$

- Since we do not know the score of rendered noisy images, it trains additional diffusion model on rendered images

$$\min_{\phi} \sum_{i=1}^{n} \mathbb{E}_{t\sim\mathcal{U}(0,1),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),c\sim p(c)} \left[\| \boldsymbol{\epsilon}_\phi(\alpha_t \underbrace{\boldsymbol{g}(\theta^{(i)}, c)}_{\text{rendered images}} + \sigma_t \boldsymbol{\epsilon}, t, c, y) - \boldsymbol{\epsilon} \|_2^2 \right]$$

- The final VSD update is given as follows:

$$\nabla_\theta \mathcal{L}_{\mathrm{VSD}}(\theta) \triangleq \mathbb{E}_{t,\boldsymbol{\epsilon},c} \left[\omega(t) \left(\boldsymbol{\epsilon}_{\mathrm{pretrain}}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, c, y) \right) \frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta} \right]$$

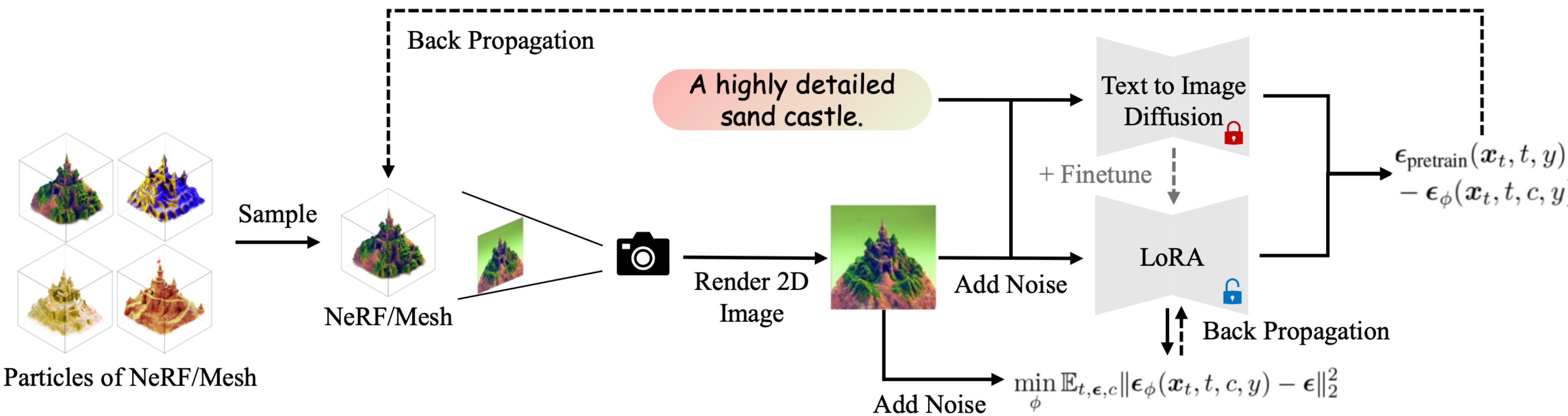
ProlificDreamer implementation

- Two-stage alternating update
  - Update NeRF (or 3D mesh) using VSD gradient update

$$\nabla_\theta \mathcal{L}_{\text{VSD}}(\theta) \triangleq \mathbb{E}_{t,\boldsymbol{\epsilon},c}\left[\omega(t)\left(\boldsymbol{\epsilon}_{\text{pretrain}}(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon}_\phi(\boldsymbol{x}_t, t, c, y)\right)\frac{\partial \boldsymbol{g}(\theta, c)}{\partial \theta}\right]$$

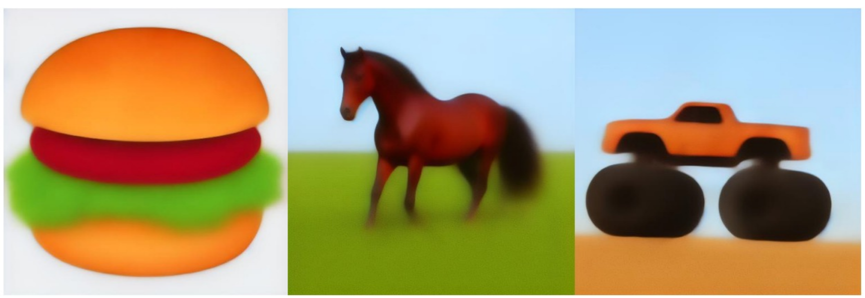
  - Update diffusion model using LoRA

$$\min_\phi \sum_{i=1}^{n} \mathbb{E}_{t\sim\mathcal{U}(0,1),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),c\sim p(c)}\left[\|\boldsymbol{\epsilon}_\phi(\alpha_t \boldsymbol{g}(\theta^{(i)}, c) + \sigma_t\boldsymbol{\epsilon}, t, c, y) - \boldsymbol{\epsilon}\|_2^2\right]$$

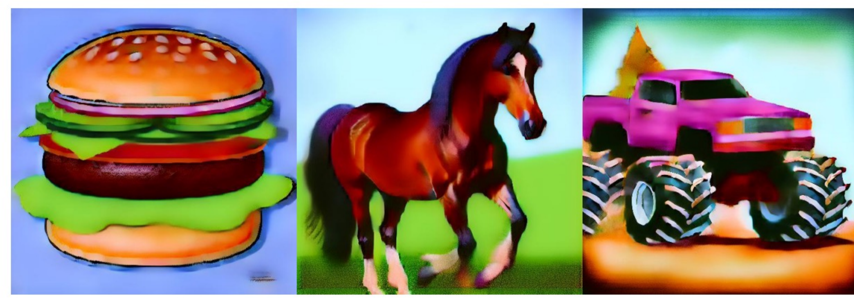


VSD allows low guidance scale in generation

- Sharp image generation with low classifier-free guidance scale, unlike SDS



(a) SDS [33] (CFG = 7.5)

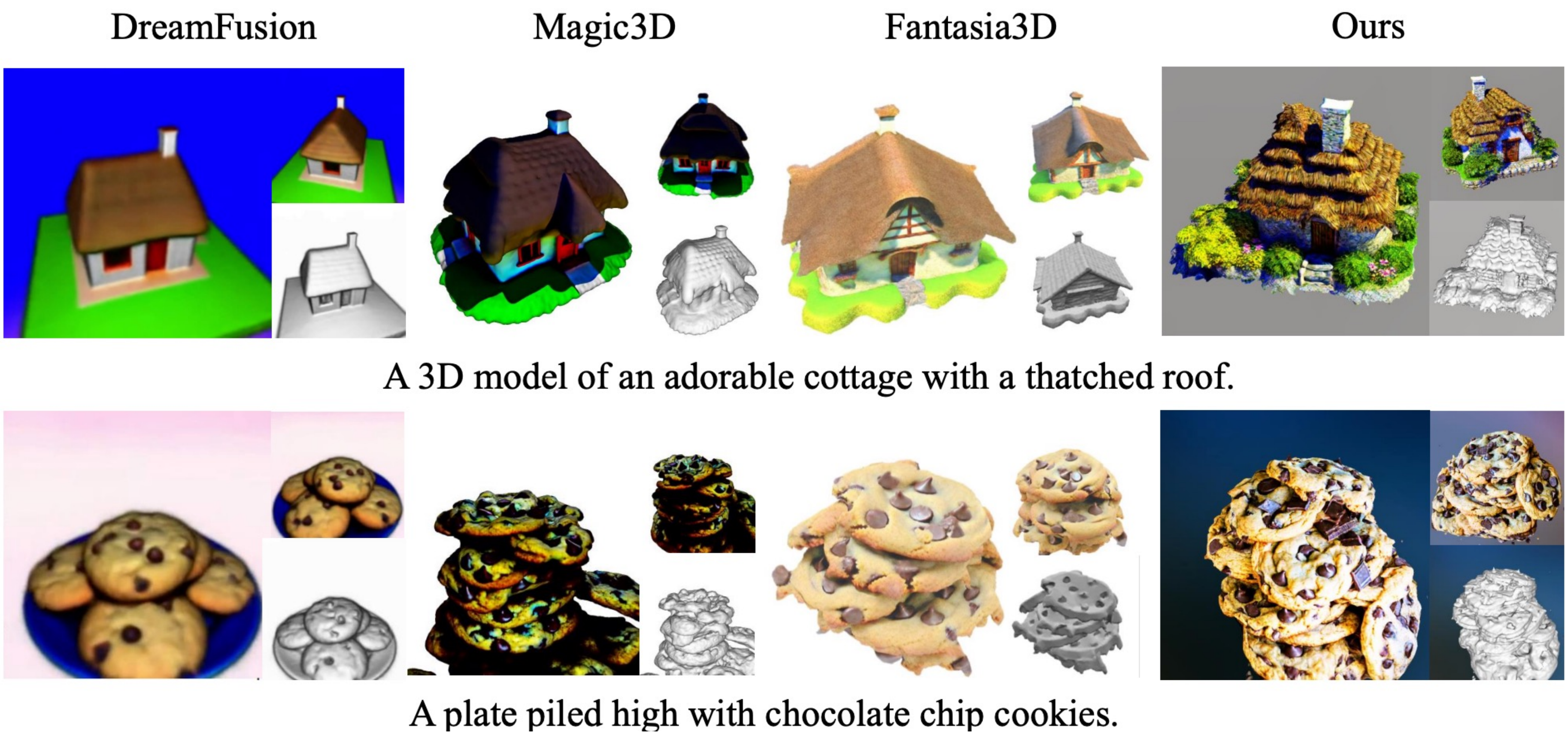


(b) SDS [33] (CFG = 100)

(c) Ancestral sampling [27] (CFG = 7.5)

(d) VSD (CFG = 7.5, **ours**)

Comparison with baseline



DreamFlow [Lee et al. ‘23]

- Score distillation methods suffer from content-shifting problem due to random timestep sampling during update
- In contrast, diffusion model samples with decreasing timestep schedule
- DreamFlow proposes to approximate probability flow for 3D optimization
    - This improves convergence speed and quality


A tiger dressed as a doctor

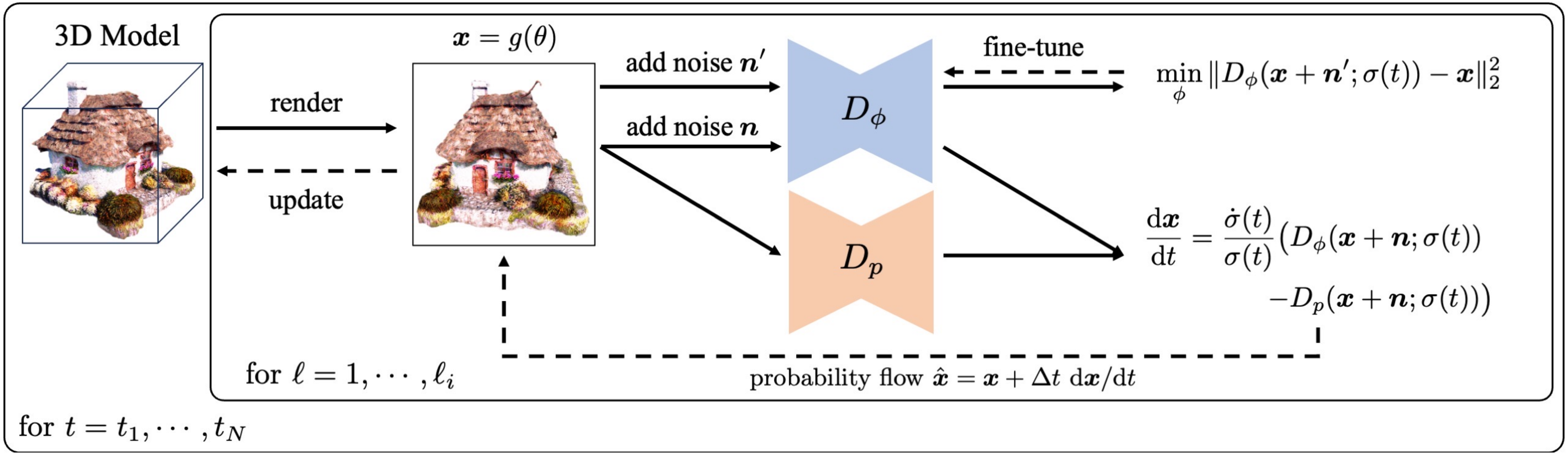

Corgi wearing a top hat


A squirrel made out of fruit


A baby dragon hatching out of stone egg

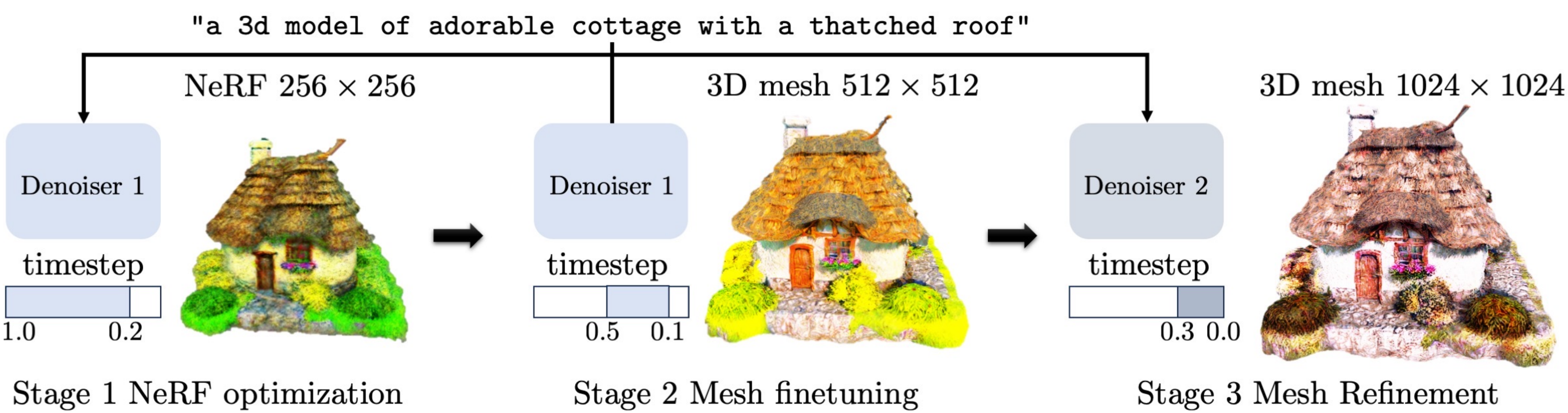
Approximate Probability Flow ODE

- Use predetermined timestep schedule (same as diffusion model) and update the 3D model with probability flow generated by pretrained diffusion model
- Amortized update (i.e., update multiple views at once) for 3D consistency



Coarse-to-fine optimization

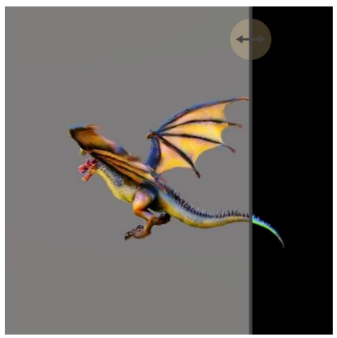
- Similar to Magic3D, DreamFlow use coarse-to-fine optimization
- Three stage update
  - Stage 1. NeRF optimization with large timesteps (res. 256x256)
  - Stage 2. 3D mesh fine-tuning with mid-timesteps (res. 512x512)
  - Stage 3. 3D mesh refinement with SDXL refiner (res. 1024x1024)



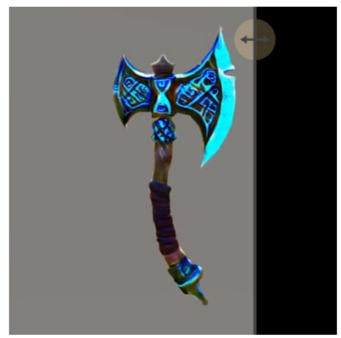
Comparison with baselines

MULTI-VIEW DIFFUSION FOR 3D GENERATION [Shi et al. '23]

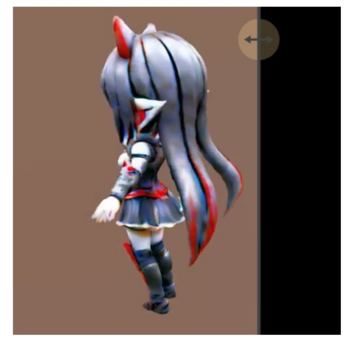
- Existing **2D-lifting** methods **suffer** from **multi-view inconsistencies**, while 3D generative models lack generalizability due to limited data.
- Proposed a multi-view diffusion model, **fine-tuning 2D diffusion** model with **multi-view awareness**.



Flying Dragon, highly detailed, breathing fire

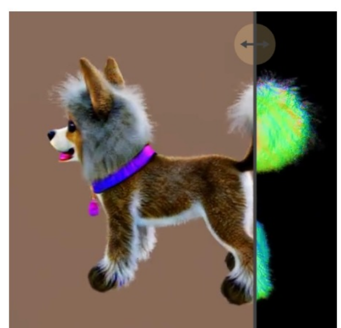


Viking axe, fantasy, weapon, blender, 8k, HD

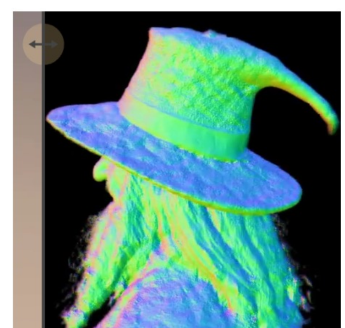


mecha vampire girl chibi

higly detailed, majestic royal tall ship, ...



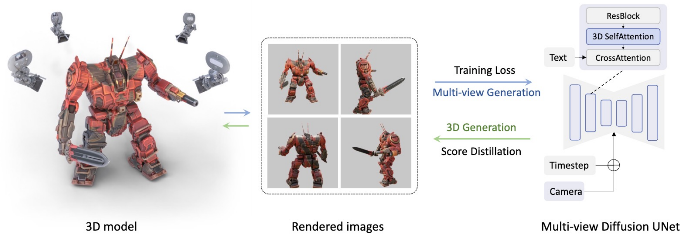
a cute fluffy dog, 4K, HD, raw



Gandalf smiling, white hair, ...

Multi-view Consistent Image Generation

- Extends **2D self-attention** into **3D** by connecting all views within the same attention layer.



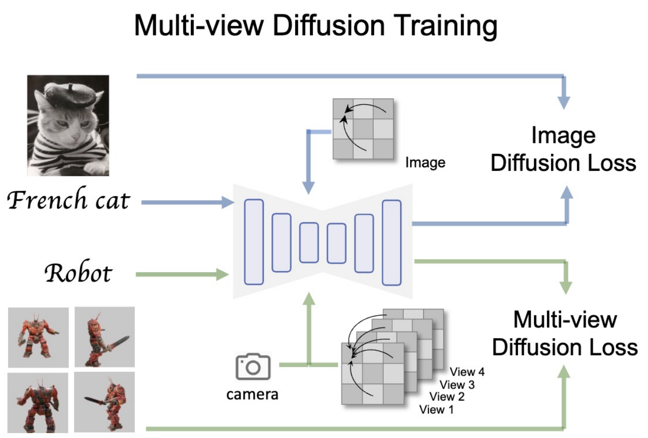
Camera Embeddings

- Encodes **camera parameters** (e.g., position, orientation) into the model for viewpoint awareness.

Training loss function

- **Balances** multi-view consistency and generalizability by integrating **3D-rendered datasets** and large-scale **2D datasets**.

$$\mathcal{L}_{MV}(\theta, \mathcal{X}, \mathcal{X}_{mv}) = \mathbb{E}_{\mathbf{x},y,\mathbf{c},t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t; y, \mathbf{c}, t)\|_2^2\right]$$

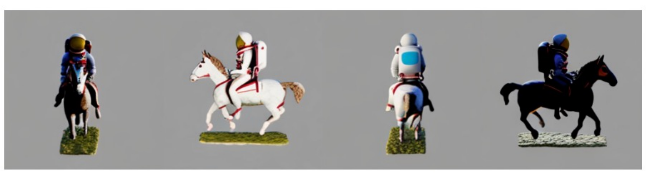


Text-to-3D generation.

- Multi-View Diffusion Prior
  - Uses a multi-view diffusion model to guide Score Distillation Sampling **(SDS)** for consistent 3D object generation.

- Improved Efficiency and Quality
  - Enhances geometry and texture quality using advanced loss techniques like **x0-reconstruction** and **CFG rescaling**.

$$\mathcal{L}_{SDS}(\phi, \mathbf{x} = g(\phi)) = \mathbb{E}_{t,\mathbf{c},\epsilon}\left[\|\mathbf{x} - \hat{\mathbf{x}}_0\|_2^2\right].$$



An astronaut riding a horse

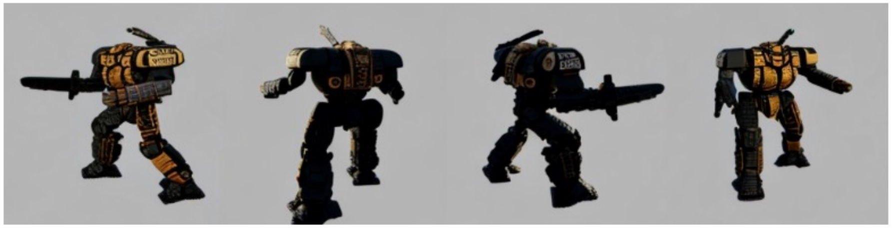
A bald eagle carved out of wood

A bull dog wearing a black pirate hat

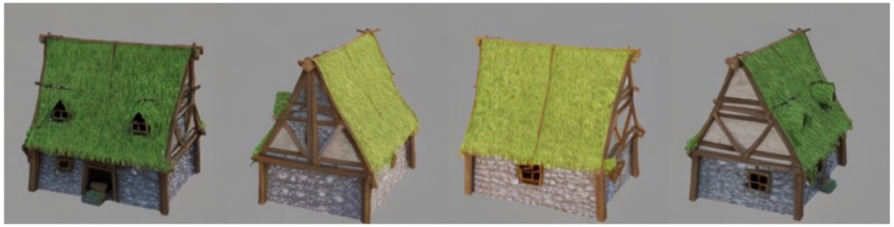
a DSLR photo of a ghost eating a hamburger

MVDream generates multi-view consistent and high-quality 3D representations, following text prompts.

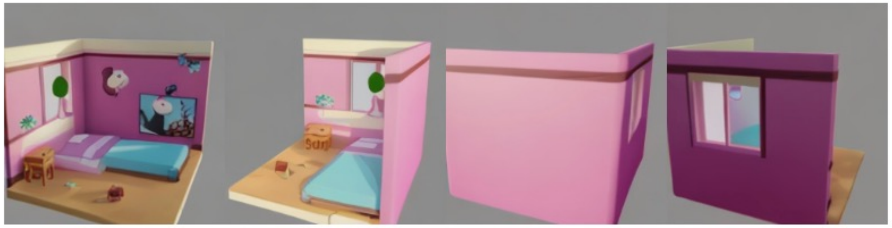
Zombie bust, terror, 123dsculpt, bust, zombie



Battletech Zeus with a sword!, tabletop, miniature, battletech, miniatures, wargames, 3d asset

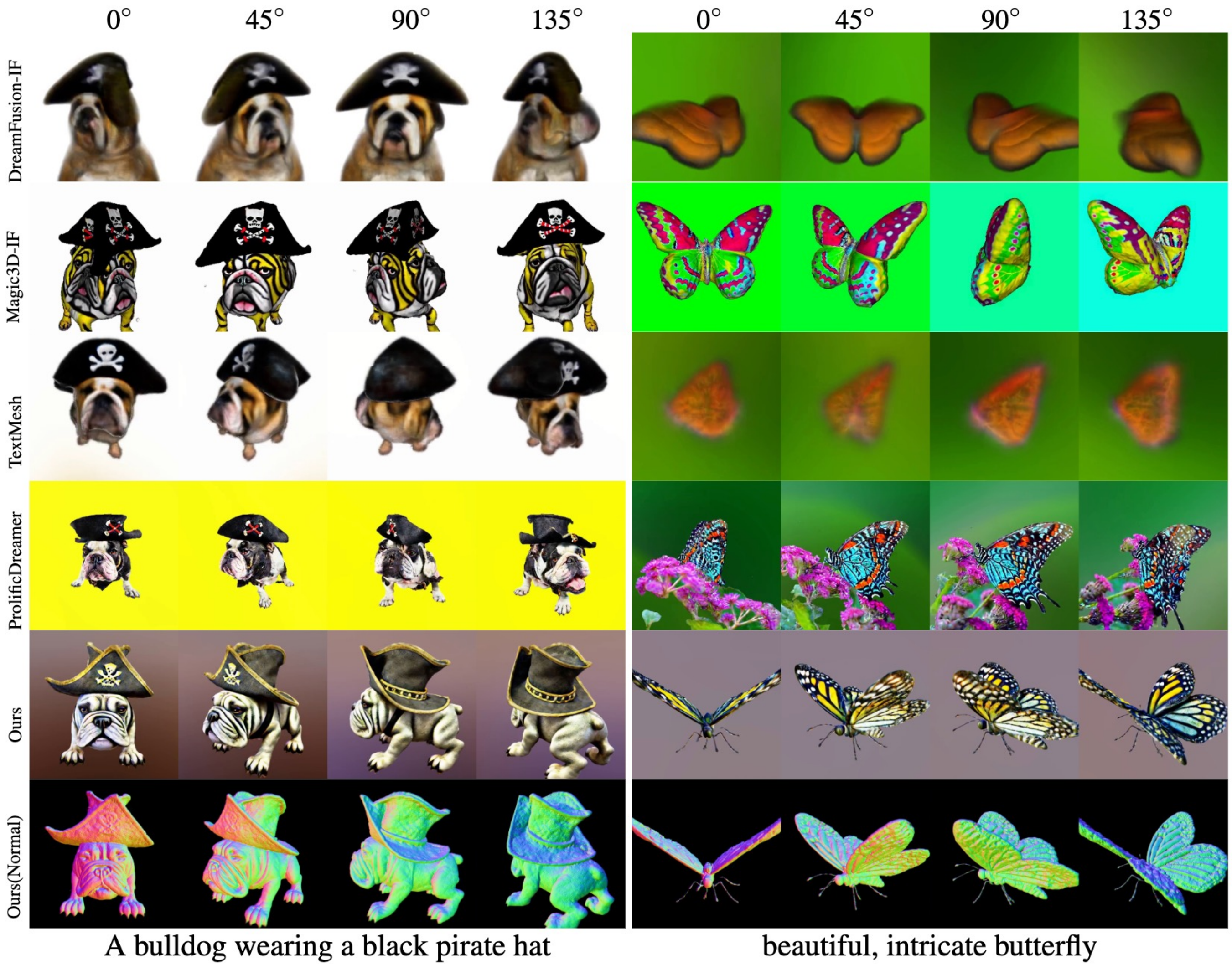


Medieval House, grass, medieval, vines, farm, middle-age, medieval-house, stone, house, home, wood, medieval-decor, 3d asset



Isometric Slowpoke Themed Bedroom, fanart, pokemon, bedroom, assignment, isometric, pokemon3d, isometric-room, room-low-poly, 3d asset

Comparison with baselines.



References

[Hertz et al., 2022] Prompt-to-Prompt Image Editing with Cross Attention Control, ICLR 2023

[Brooks et al., 2022] InstructPix2Pix: Learning to Follow Image Editing Instructions, CVPR 2023

[Gal et al., 2022] An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

[Ruiz et al., 2022] DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, CVPR 2023

[Lee et al., 2024] Direct Consistency Optimization for Robust Customization of Text-to-Image Diffusion Models, NeurIPS 2024

[Hertz et al., 2024] Style Aligned Image Generation via Shared Attention, CVPR 2024

[Ye et al., 2024] IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, arXiv 2023

[Wang et al., 2024] Multi-subject Zero-shot Image Personalization with Layout Guidance, arXiv 2024

[Pan et al., 2024] KOSMOS-G: Generating Images in Context with Multimodal Large Language Models, ICLR 2024

[Wang et al., 2024] Multi-subject Zero-shot Image Personalization with Layout Guidance, arXiv 2024

[Wang et al., 2024] InstantID: Zero-shot Identity Preserving Generations in Seconds, arXiv 2024

[Wang et al., 2024] PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embeddings, CVPR 2024

[Zhang et al., 2023] Adding Conditional Control to Text-to-Image Diffusion Models, ICCV 2023

[Poole et al., 2022] DreamFusion: Text-to-3D using 2D Diffusion, ICLR 2023