TokenVerse: Versatile Multi-concept Personalization in Token Modulation Space

Daniel Garibi*, Shahar Yadin*, Roni Paiss, Omer Tov, Shiran Zada,
Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, Tali Dekel

Google DeepMind

SIGGRAPH 2025 best paper

STRUCT Group Seminar
Presenter: Yifan Li
2025.7.7

Outline

- Background
- Method
- Experiments
- Conclusion

Background: Problem Definition

- Unconditional Generation: lack of controllability
- Text-to-image Generation: lack of flexible personalization
- Customized Generation
  - Given personalized condition as input

Background: Problem Definition

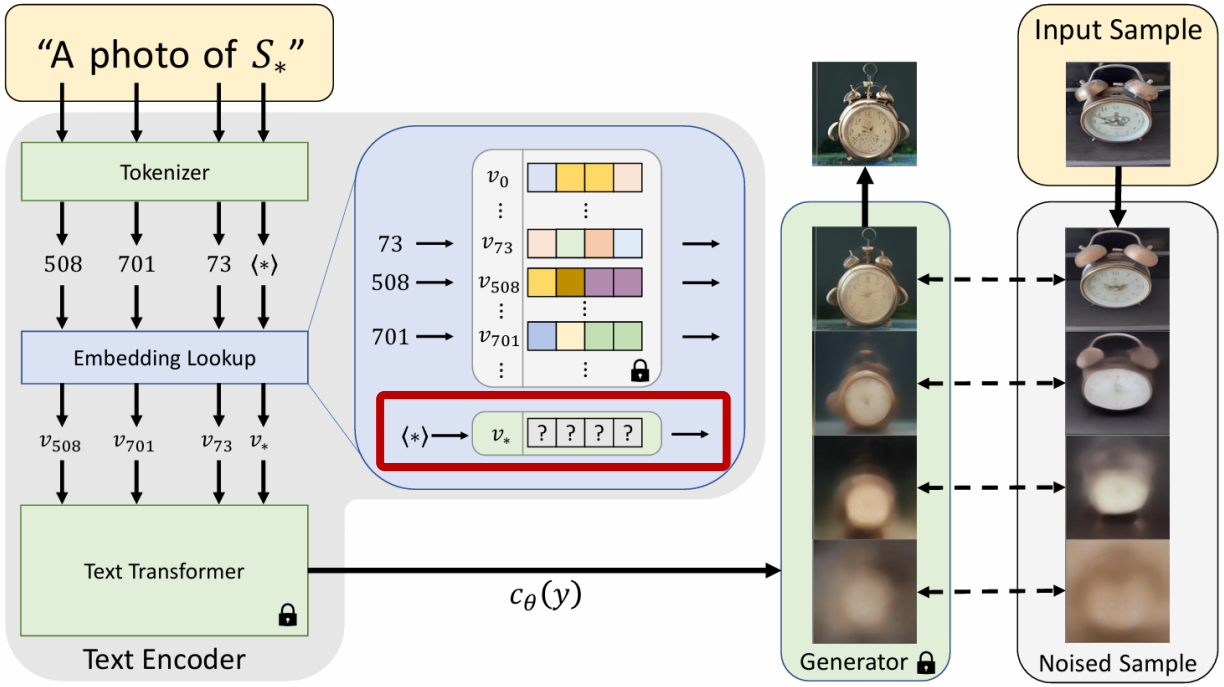
- How to leverage T2I models during personalization?
- Customized Generation based on T2I models
  - Optimize text embeddings
    - *Textual Inversion [ICLR'23]*
  - Finetune generative model
    - *DreamBooth [CVPR'23]*

Background: Textual Inversion

Optimize text embeddings

- Establish the correspondence between special text feature v_* and image

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

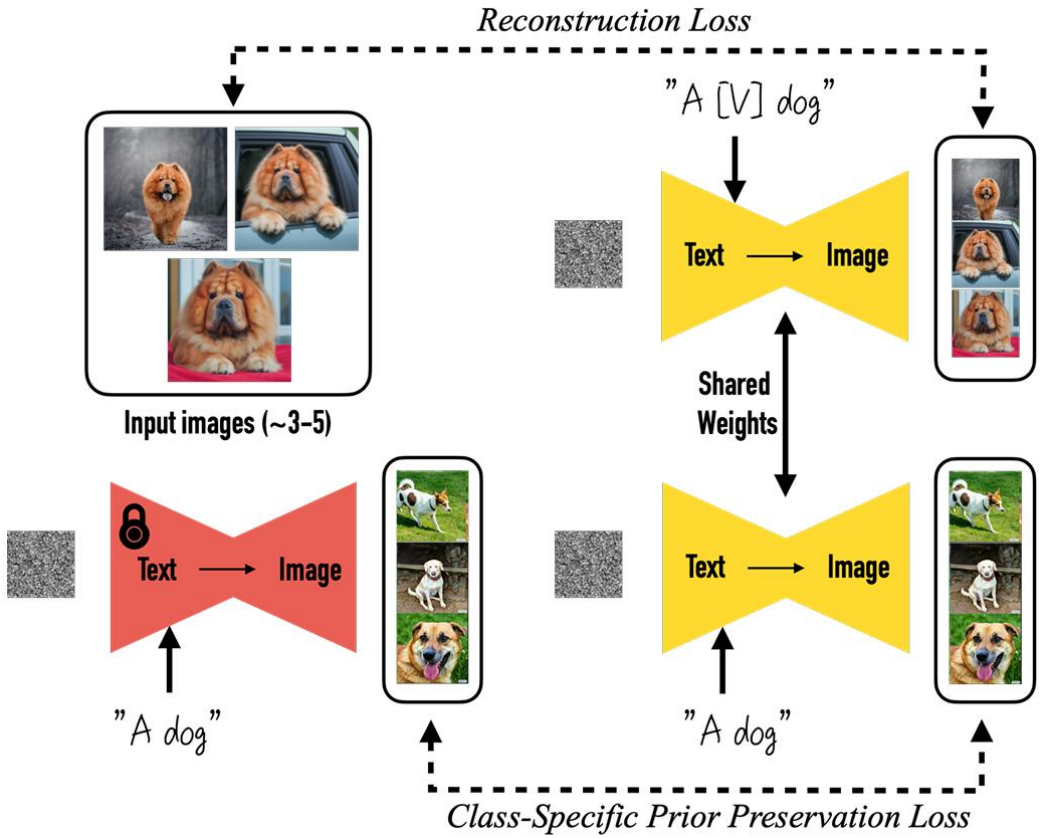


Rinon Gal, Yuval Alaluf, et al., "An image is worth one word: Personalizing text-to image generation using textual inversion". ICLR'23.

Background: DreamBooth

Finetune generative model

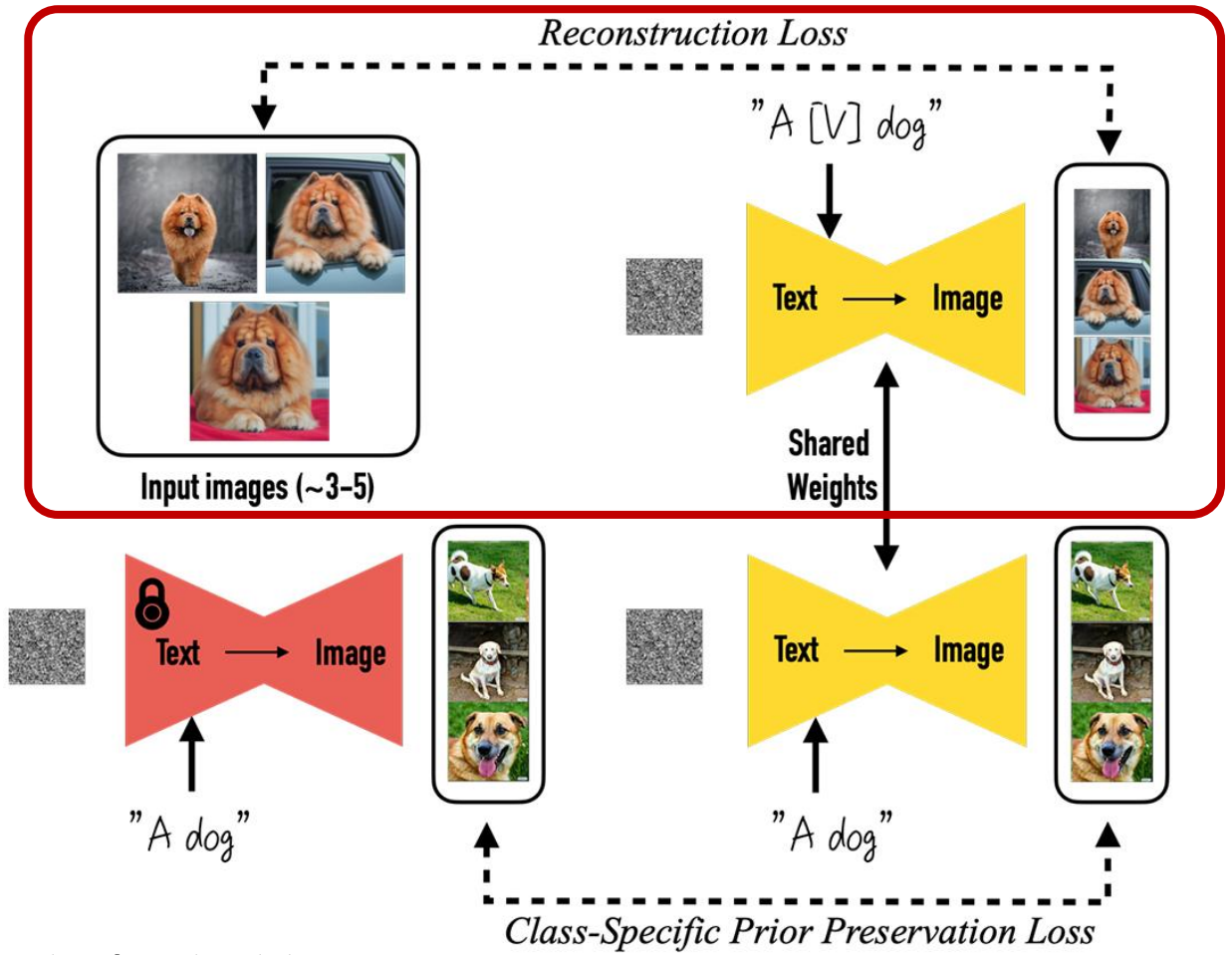
- Reconstruction: learn specific concept with a unique text token ‘[V]’
- Class-specific Prior Preservation: ensure normal generation ability



Background: DreamBooth

Reconstruction

- Embed personal concept within a specific token '[V]'
- Finetune the whole T2I model (VAE & U-Net)
  - High computational cost (5 min on a A100, SD)

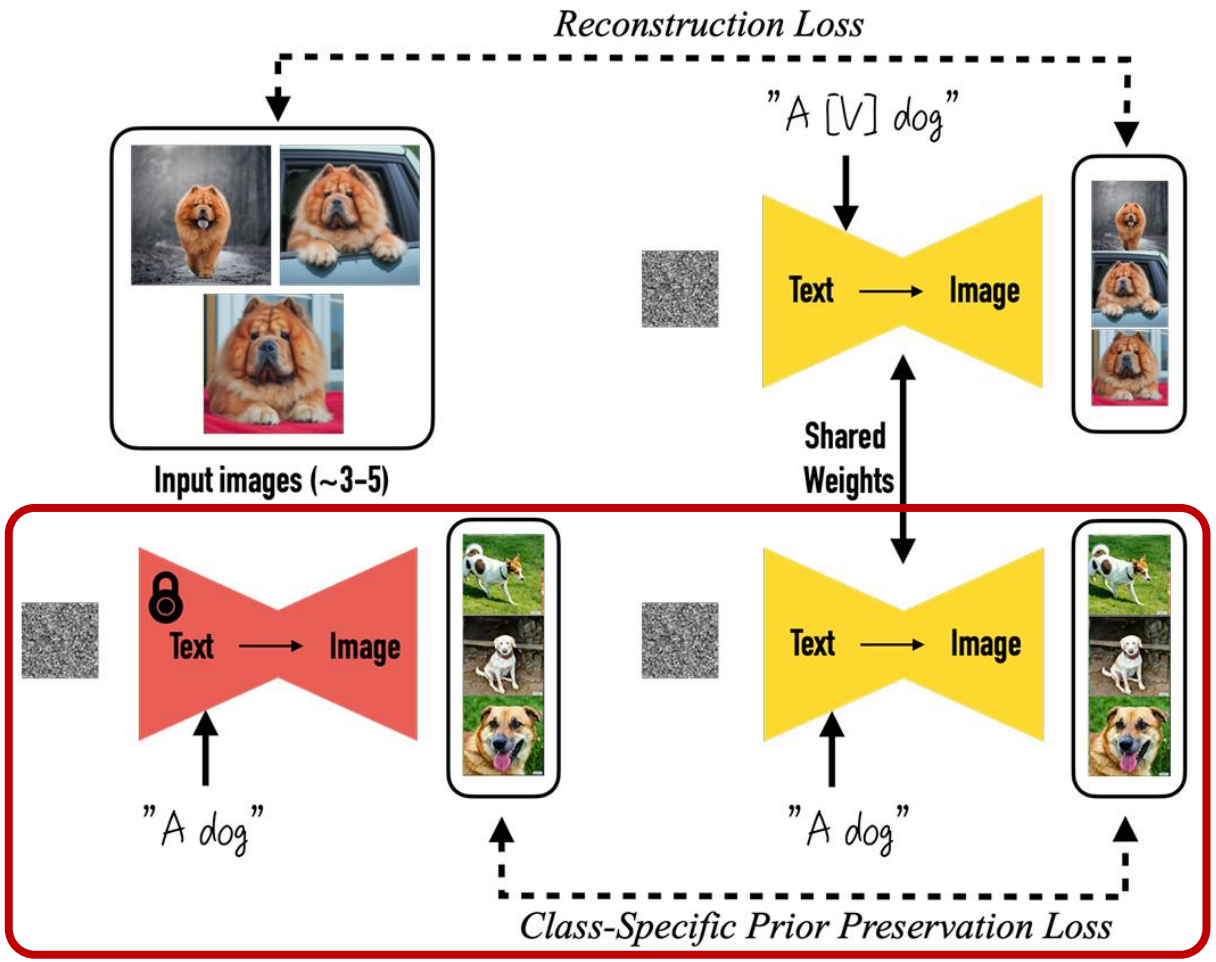


Nataniel Ruiz, Yuanzhen Li, et al., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". CVPR'23.

Background: DreamBooth

Class-specific Prior Preservation

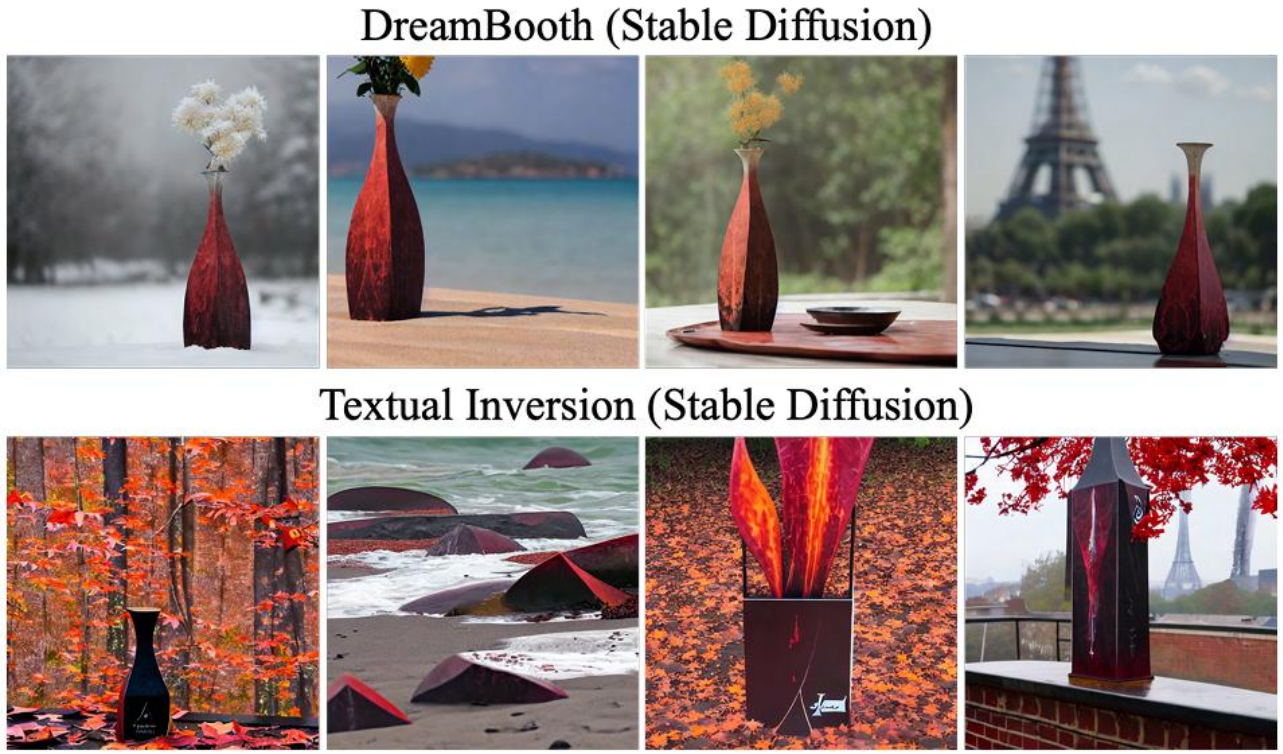
- Generate specific class-conditioned images as ground truth
- Avoid overfit to input concept images
- Encourage result diversity



Background: DreamBooth

DreamBooth achieve better results compared with *Textual Inversion*

Input images



Conceptualize results

Background

Single-concept personalization

- Struggle to disentangle **non-object** concepts
- Struggle to disentangle **multiple concepts** within one image

Multi-concept personalization

- Object, **accessories, materials, pose, lighting**, …
- Disentanglement of concepts
- Composition of concepts

Background

Multi-concept personalization

- Disentanglement of concepts
  - Inverse content back to text embeddings
  - Train a specific LoRA to overfit input images
- Composition of concepts
  - Replace original words with special text embeddings and corresponding spatial guidance (bbox or mask)
  - Fuse multiple LoRAs via optimization

Background: Inspiration Tree

Decompose concepts in a hierarchical tree structure

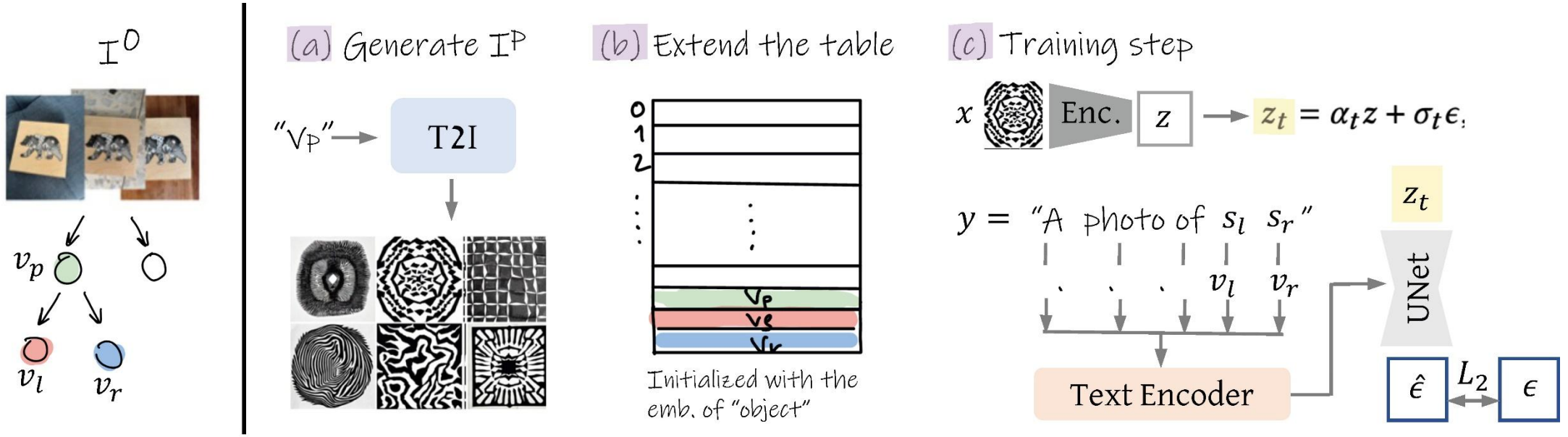
- Build binary tree from top to bottom
- Iteratively add two new nodes at a time

Yael Vinker et al.,
"Concept Decomposition for Visual Exploration and Inspiration". SIGGRAPH Asia 2023 best paper

Background: Inspiration Tree

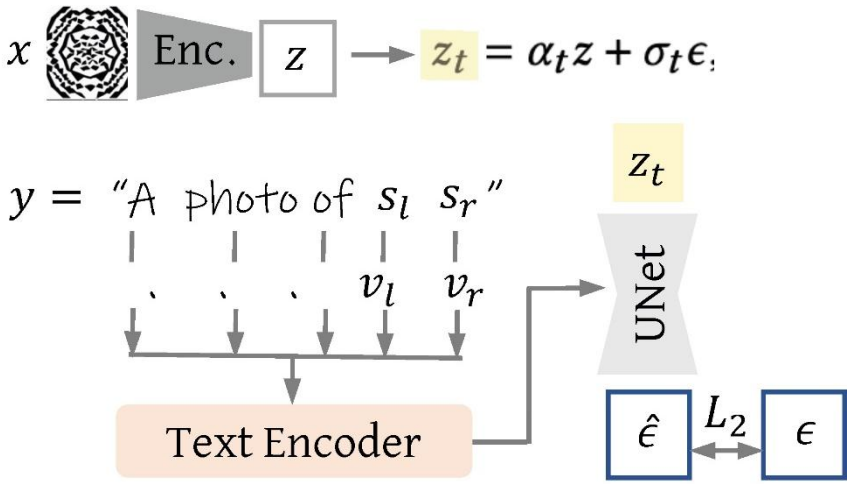
Main philosophy: Divide and conquer

- Divide pairs of child node to represent distinct concepts
- Maintain reasonable semantics in each nodes



Background: Inspiration Tree

- **Binary Reconstruction**
  - Each pair of children nodes together should encapsulate the concept depicted by their parent node



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

Background: Inspiration Tree

- **Coherency**
  - Each individual node should depict a coherent concept which is distinct from its sibling

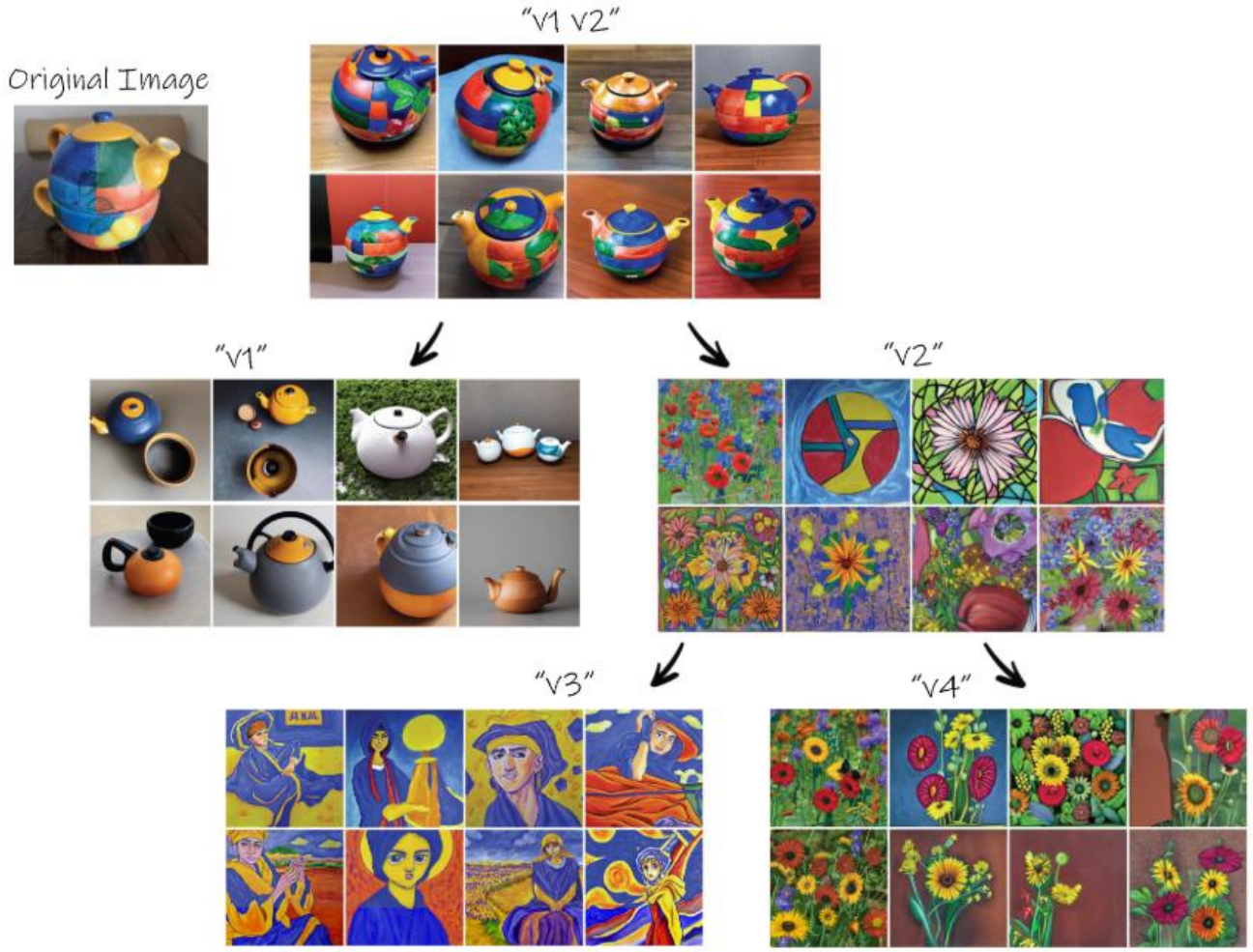
Background: Inspiration Tree

- **Coherency**
  - Each individual node should depict a coherent concept which is distinct from its sibling
  - Formally, measure coherency by cosine similarity of CLIP image embeddings: $\mathcal{C}(I^a, I^b) = mean_{I_i^a \in I^a, I_j^b \in I^b, I_i^a \neq I_j^b}(sim(CLIP(I_i^a), CLIP(I_j^b)))$
  - Maximize left/right child node intra-similarity
    Minimize left/right child node inter-similarity:

$$\{v_l^*, v_r^*\} = \underset{\{v_l^i, v_r^i\} \in V_s}{\arg\max} \left[C_l^i + C_r^i + (min(C_l^i, C_r^i) - \mathcal{C}(I^{v_l^i}, I^{v_r^i}))\right]$$

Background: Inspiration Tree

- Strong Representation Extraction Ability



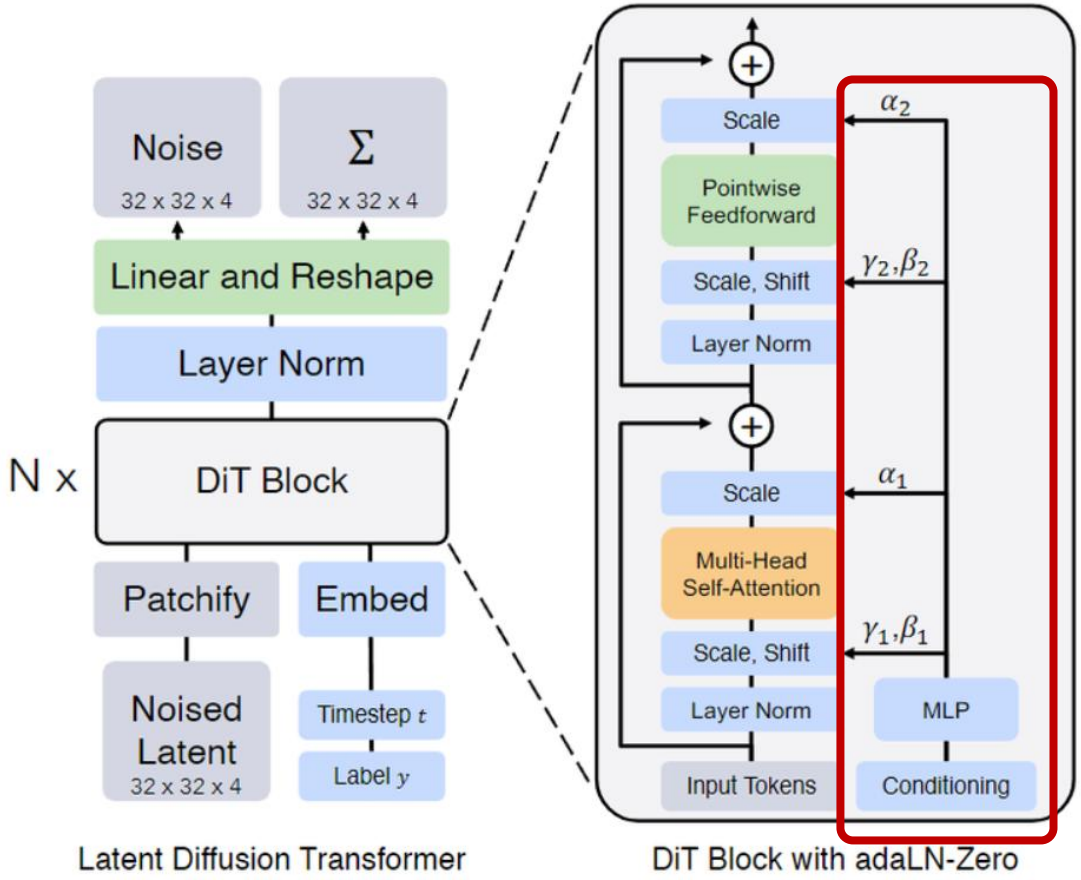
Outline

- Background
- Method
- Experiments
- Conclusion

Method: Preliminary

Diffusion Transformer (DiT)

- Jointly processes text and image tokens with self-attention
- Scalable arch.



Global condition guidance

Method: Preliminary

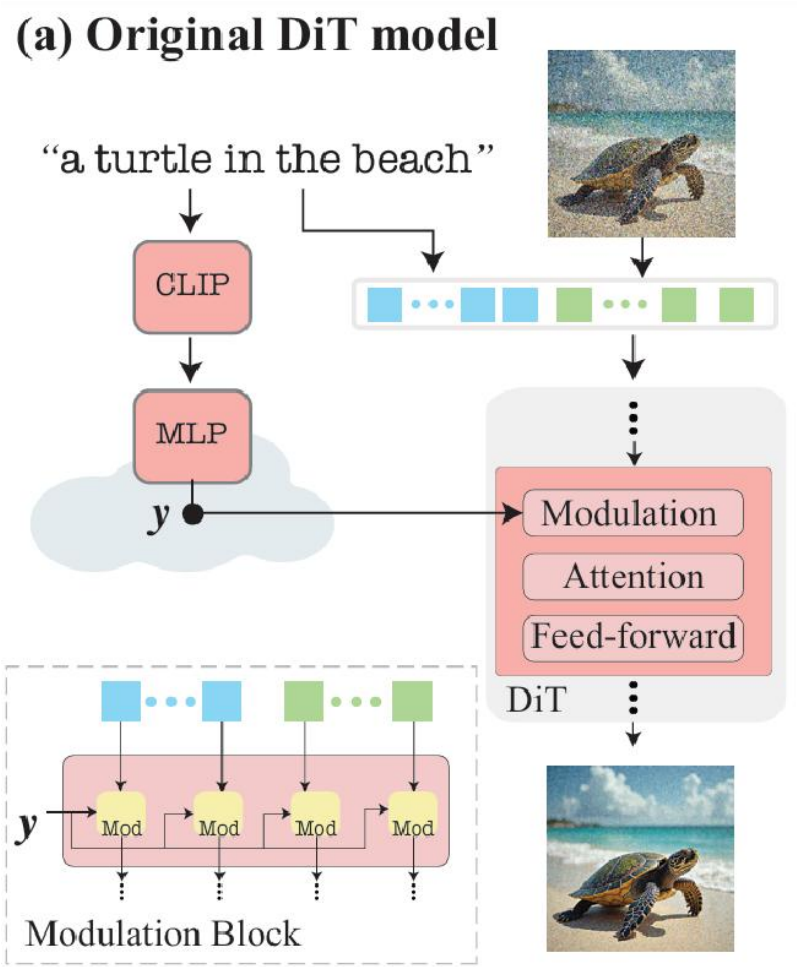
Modulation mechanism in DiT

- Merge timestep and pooled CLIP text embedding with MLP

$$y = \mathrm{MLP}(t, \mathrm{CLIP}(p))$$

- As a global modulation signal, y is then split to channel-wise scale and shift parameters

$$f_{mod} = Scale(y) \cdot f_{ori} + Shift(y)$$



Method

Naïve Solution: modulate control signal on global-level

$$\Delta_{\text{attribute}} = \text{MLP}(t, e_{\text{attribute}}) - \text{MLP}(t, e_{\text{neutral}})$$

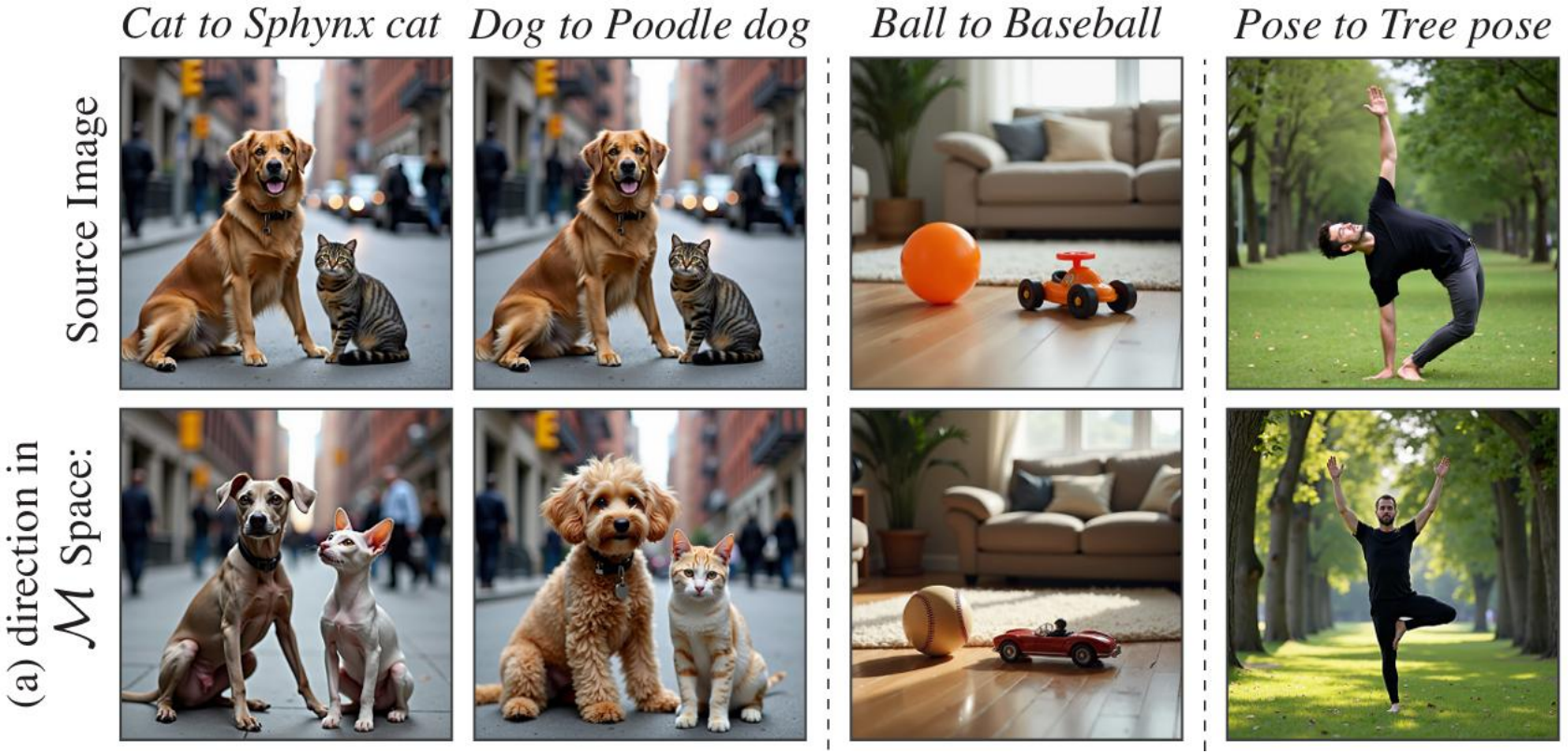
$$y = y + w\Delta_{\text{attribute}}$$

- $e_{neutral}$: pooled embedding of original description (‘*A dog*’)
- $e_{attribute}$: pooled embedding with some attribute added (‘*A poodle dog*’)

Method

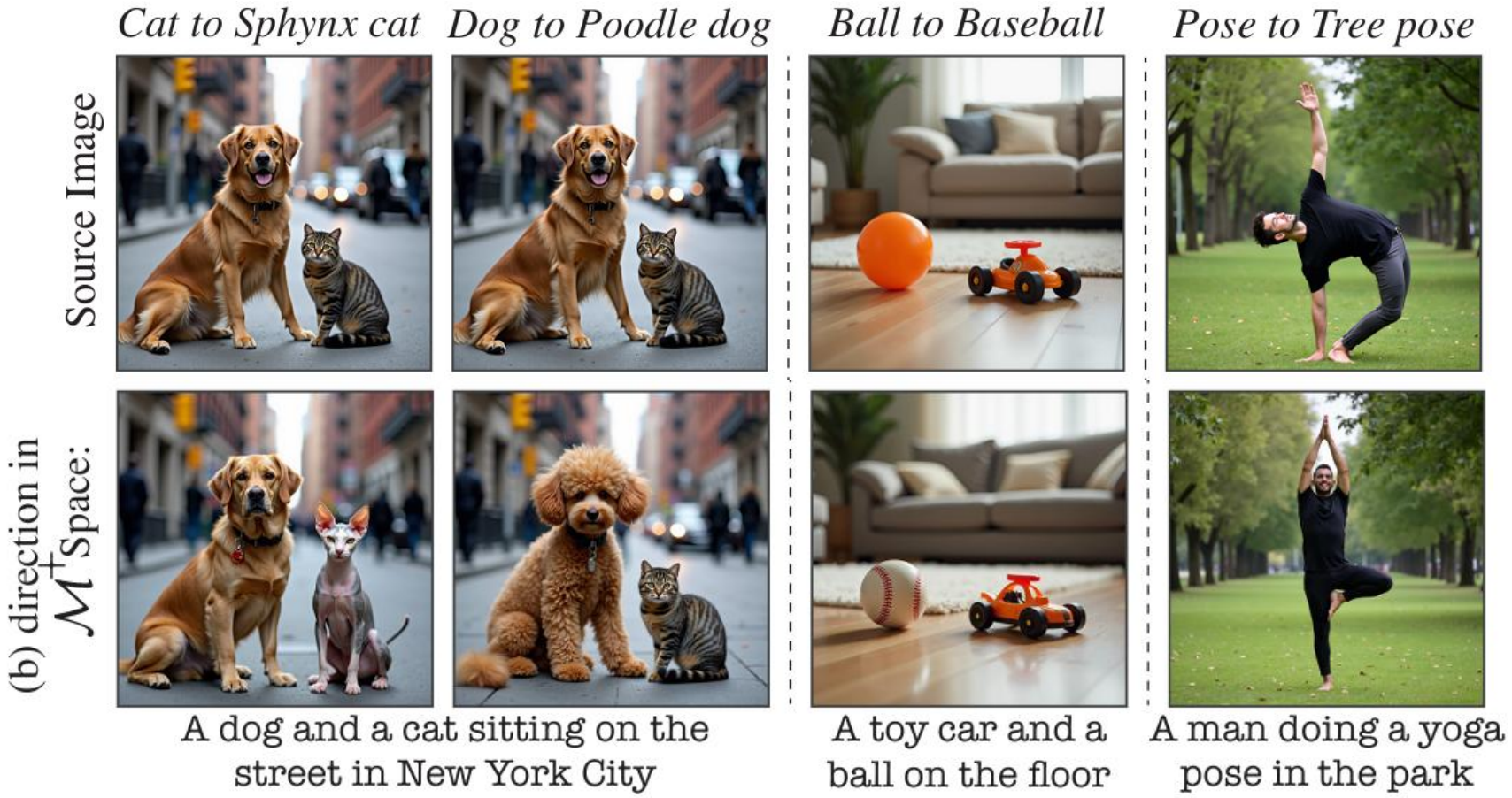
Naïve Solution: modulate control signal on global-level

- Drawbacks: inaccurate control with non-local edit



Method

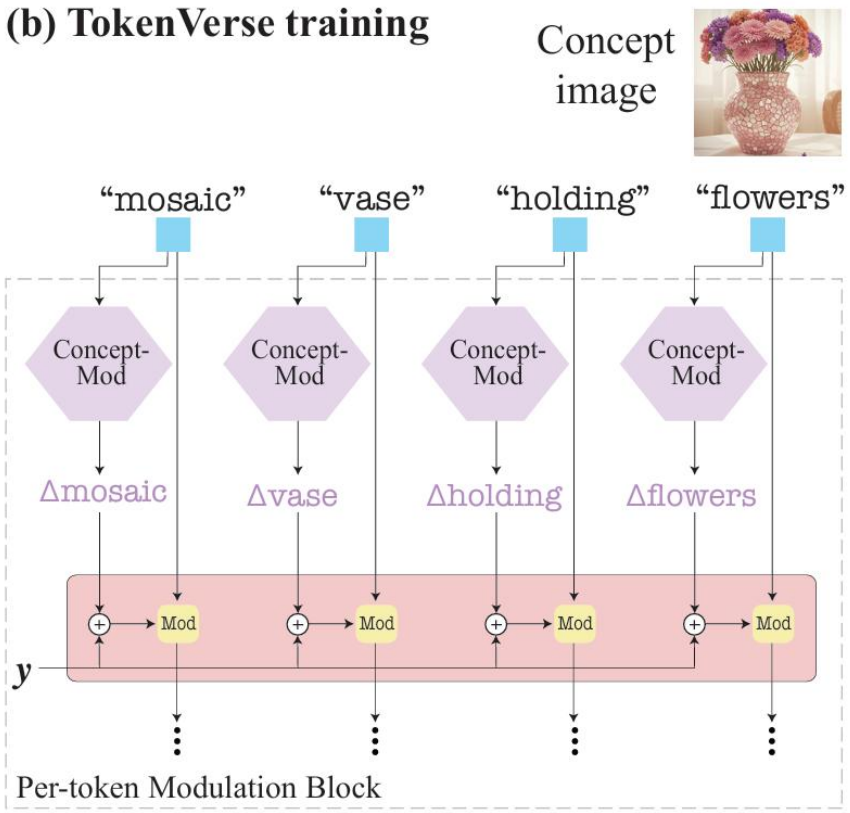
Improved Solution: per-token modulation



Method

Improved Solution: per-token modulation

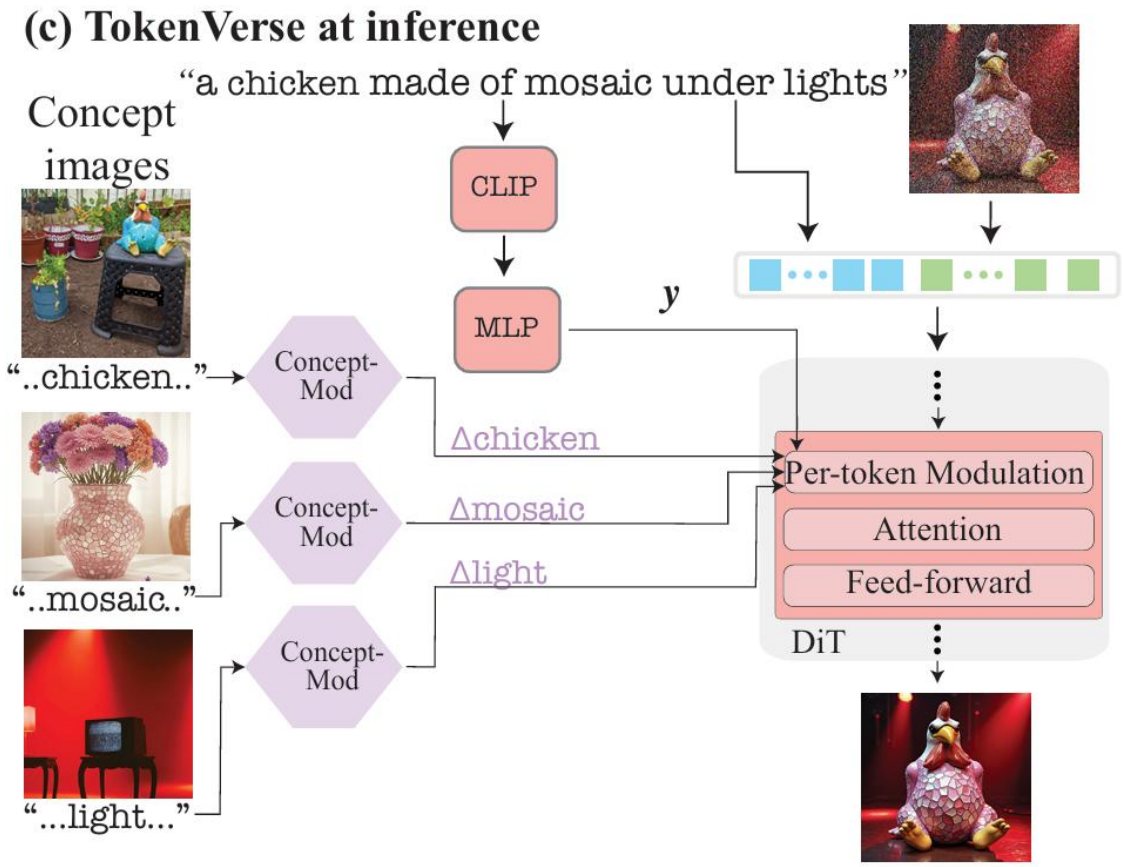
- Learns a modulation vector offset Δ for each text token



Method

Improved Solution: per-token modulation

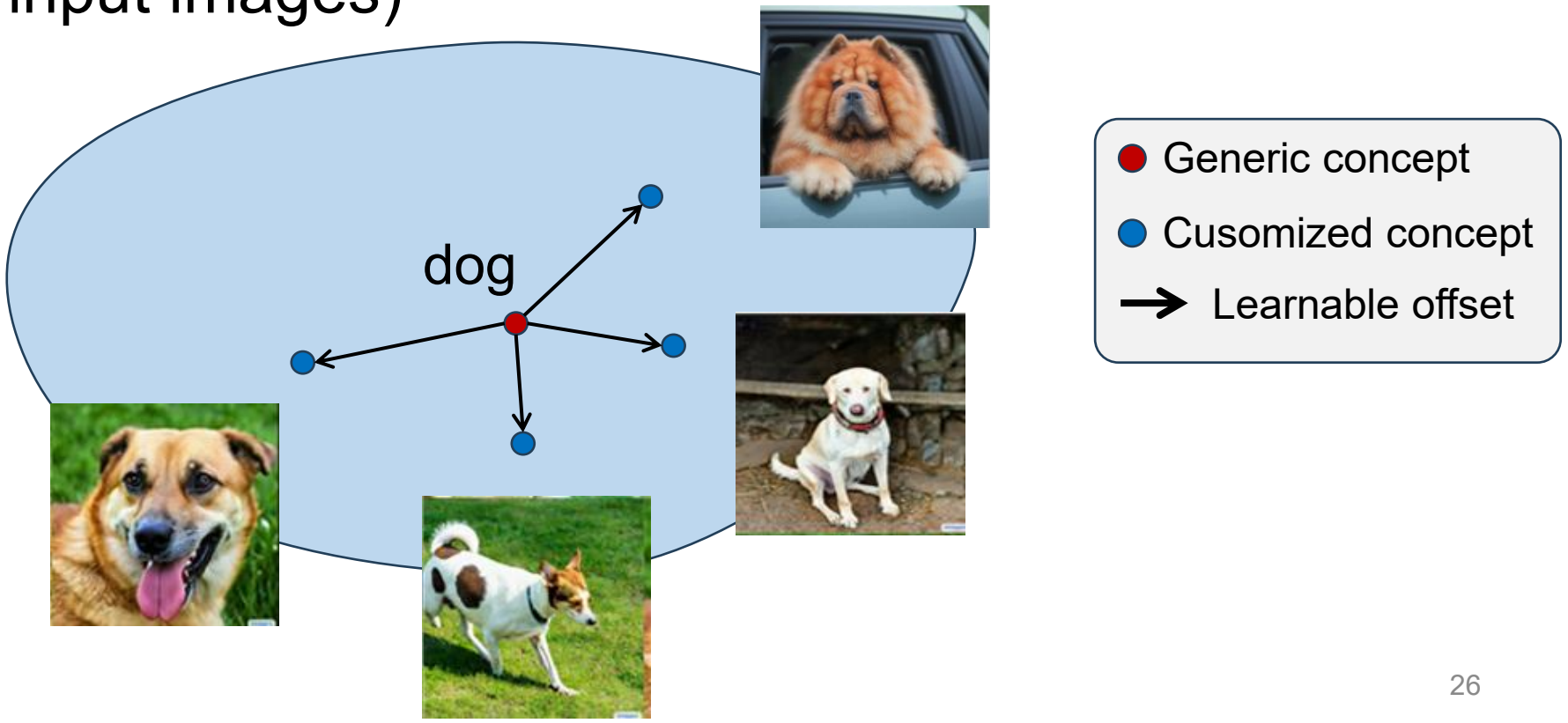
- Combines each concept through learned per-token offsets



Method

Improved Solution: per-token residual learning

- Offset: transfer a **generic concept** to its **customized version** (indicated by input images)



Method

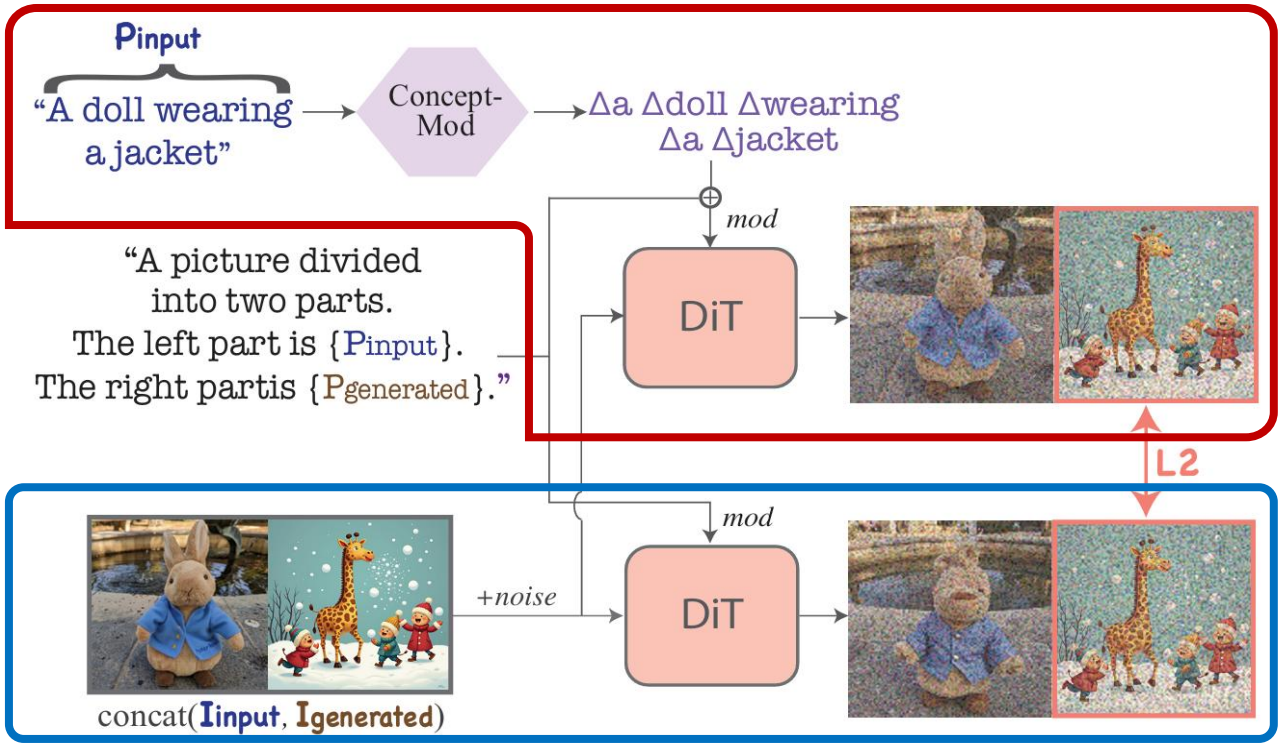
Per-block optimization

- Coarse-to-fine training
  - Stage-1: high noise level (800~1000), aims at coarse concept alignment
  - Stage-2: refine directions with lower noise levels (0~800)
- Per-token, per-block
  - Train MLP that outputs a vector per transformer block
  - **Instead of** only adding offset on text tokens
    (which is equal to textual inversion)

Method

Concept isolation loss

- steer the optimization such that the optimized directions do not affect concepts that do not appear in the concept image



Residual modulation generation

Consistent Regularization

Original generation

Outline

- Background
- Method
- Experiments
- Conclusion

Experiments

- Multiple object concepts disentanglement & composition

a **dog** wearing a hat and a **necklace**

a cat wearing **glasses** and a **shirt**

a man wearing a **shirt**

a laying **dog** wearing **glasses**

a **dog** wearing a **shirt** and **necklace** having a picnic

a **dog** wearing a **shirt**, **glasses** flying in the sky tied to balloons

a **dog** with a **shirt**, **glasses** on a rollercoaster

Experiments

- Multiple complex concepts disentanglement & composition

a **mosaic** vase, holding flowers

a mountain under a **light**.

a **doll** wearing a jacket

a **boat** is floating on the water

a **light** over paris

a **doll** on a **boat** made of **mosaic**

a **doll** surfing on a surfboard made of **mosaic** under a **light**

Experiments

- Multiple complex concepts disentanglement & composition

a **doll** inside a bucket

a **rocking horse** on the floor

a **doll** sitting on **bench** in the **garden**

a woman standing in a **fog** holding lamp

a **doll** riding **rocking horse** in a **fog**

a **doll** riding **rocking horse** in the **garden**

a **doll** on a **bench** in the park, **fog** around

Experiments

- Multiple complex concepts disentanglement & composition

a **doll** inside a bucket on a cart

a dog made of **plastic**

a rocking horse on a **carpet**

a tv on a table in under a **light**

a **doll** in a bucket made of **plastic** on a **carpet** under a **light**

Experiments

- Extreme multiple concepts personalization

Concept Images

a **man** leaning against a wall

a man wearing a **shirt**

a dog wearing a **hat** and a necklace

a cat wearing **glasses** and a shirt

a dog wearing a hat and a **necklace**

a **backpack** hanging on a chair

a doll sitting on a **bench** in the garden.

a mirror on a wall and a **table**

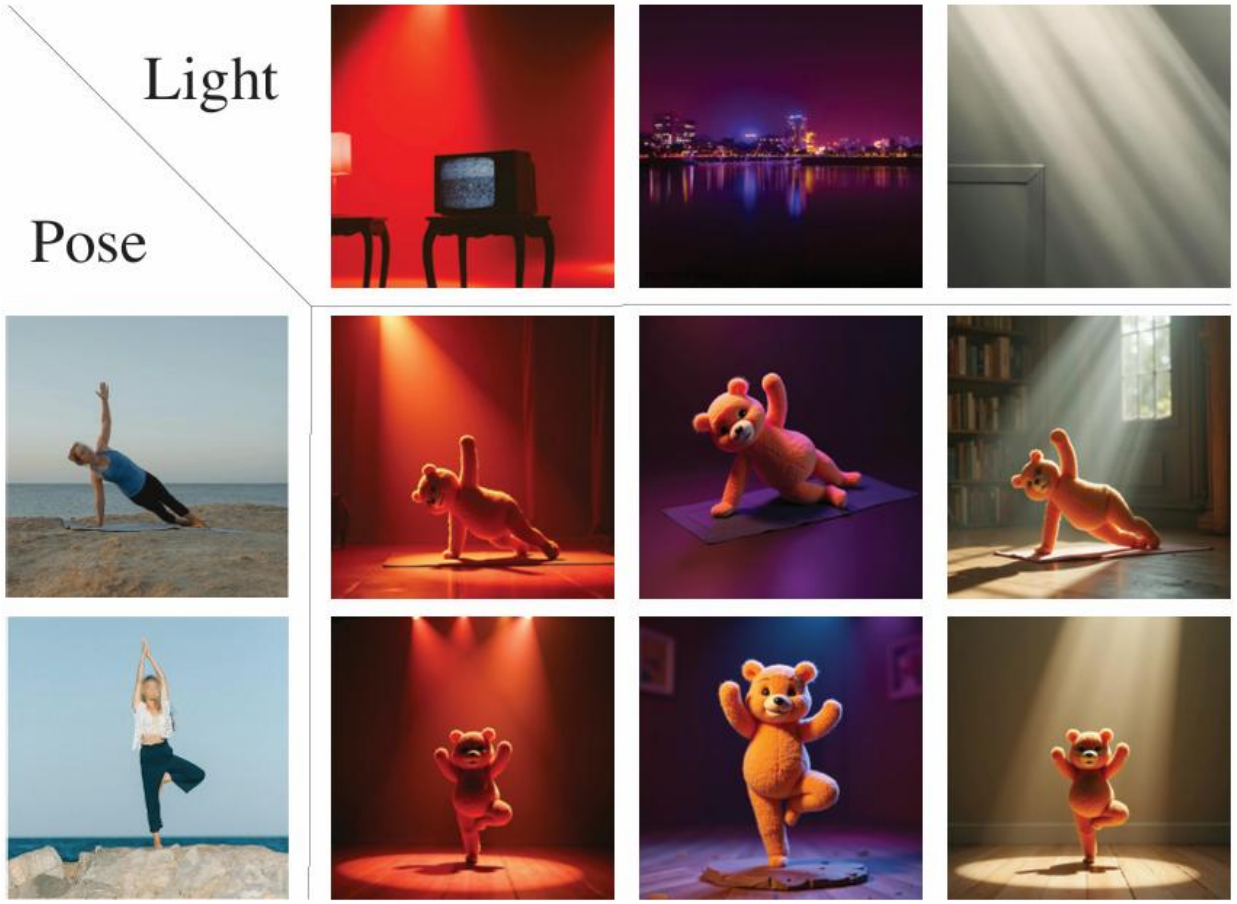
a woman standing in a **fog**

Generated Image

a **man** wearing **shirt**, **hat**, **glasses**, **necklace** holding a **backpack**, sitting on a **bench** near a **table** in front of a **fog**.

Experiments

- Concepts beyond objects (Background Relight)

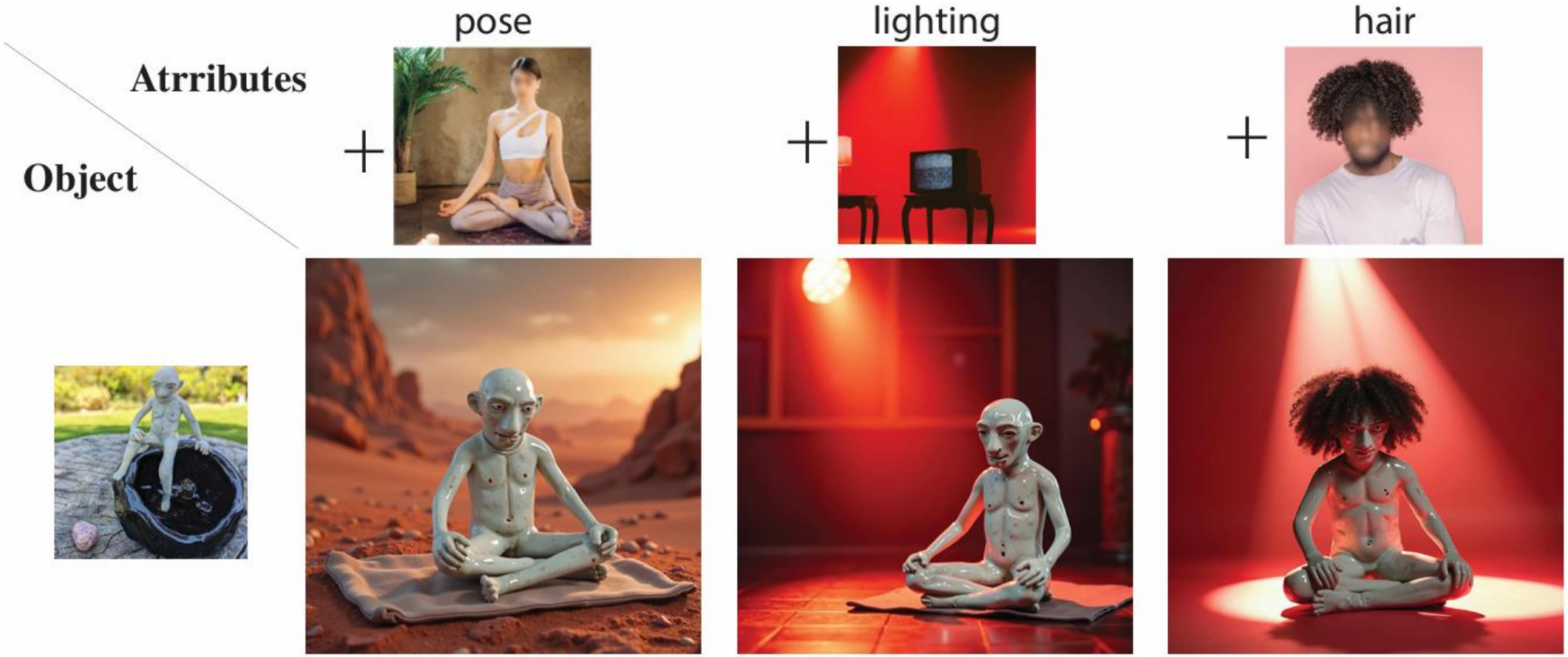


Experiments

- Progressive composition of concepts

Experiments

- Progressive composition of concepts



Experiments

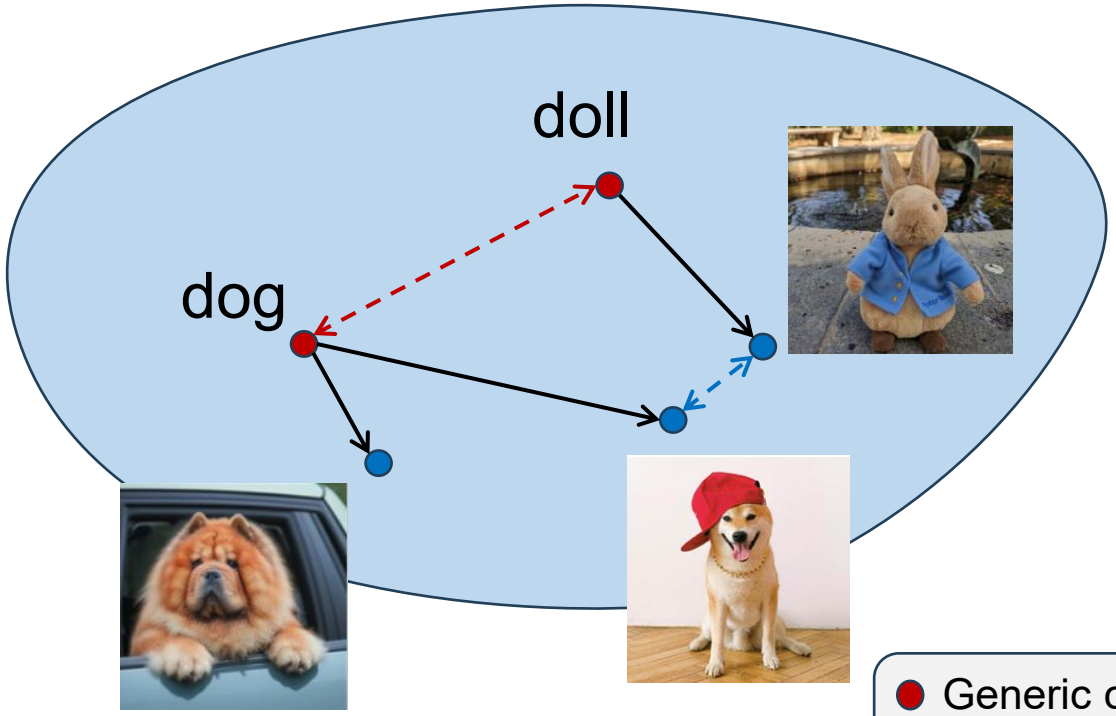
- Ablation Study

Experiments: Limitations

- Sensitive to initial caption
  - “sheep” → “doll” brings a huge difference

Experiments: Limitations

- Hybrid results when concepts are similar



a **dog** wearing a hat and a necklace

a **doll** wearing a jacket

a **doll** and a **dog** in the garden

Generic concept

Cusomized concept

Learnable offset

Experiments: Limitation

- Optional Mitigation:
  - Joint training on both concepts

a **doll** wearing a jacket next to
a **dog** wearing a hat and a necklace

Training

a **doll** and a **dog** in the garden

Inference

Outline

- Background
- Method
- Experiments
- Conclusion

Conclusion

- A versatile multiple complex concept generation method
- Insightful innovations based on DiT, which jointly process image and text tokens
- Fancy visualized results, rich application scenarios
- Fluent paper writing, complete experiment and limitation analysis

Thanks for listening!

Experiments: Details

- Backbone: Flux-dev, 58 DiT blocks, 3072 middle dimension
- Concept isolation loss: randomly sample from a fixed set of 25 pairs of captions and generated images