Attention Interpolation of text-to-image Diffusion

NeurIPS 2024

Qiyuan He[1], Jinghao Wang[2], Ziwei Liu[2], Angela Yao[1]

[1]National University of Singapore, [2]Nanyang Technology University

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NUS National University of Singapore

01 Background on unconditional deep interpolation

- Target at generating image sequence given two generated images
- Mainly focus on interpolation along the latent space of initial seeds

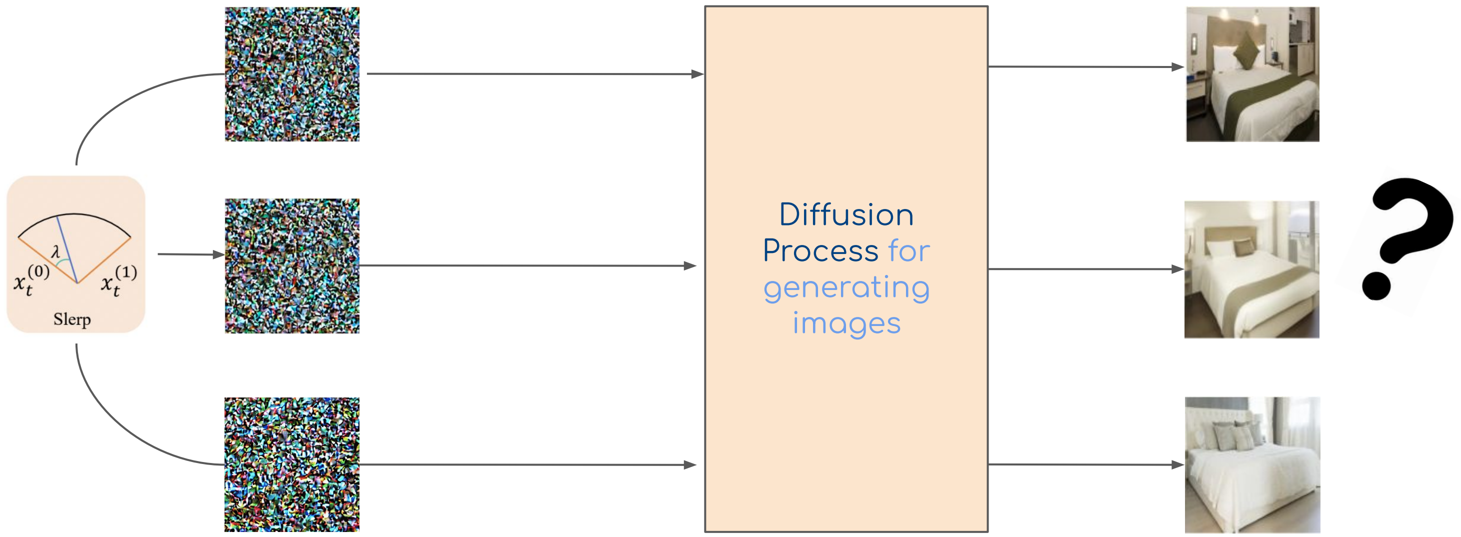


Image from: NOISEDIFFUSION: CORRECTING NOISE FOR IMAGE INTERPOLATION WITH DIFFUSION MODELS BEYOND SPHERICAL LINEAR INTERPOLATION

01 Conditional deep interpolation

- What if we want to interpolate between different conditions?

Interpolation between different initial seeds in latent space

How do we interpolate between different conditions?

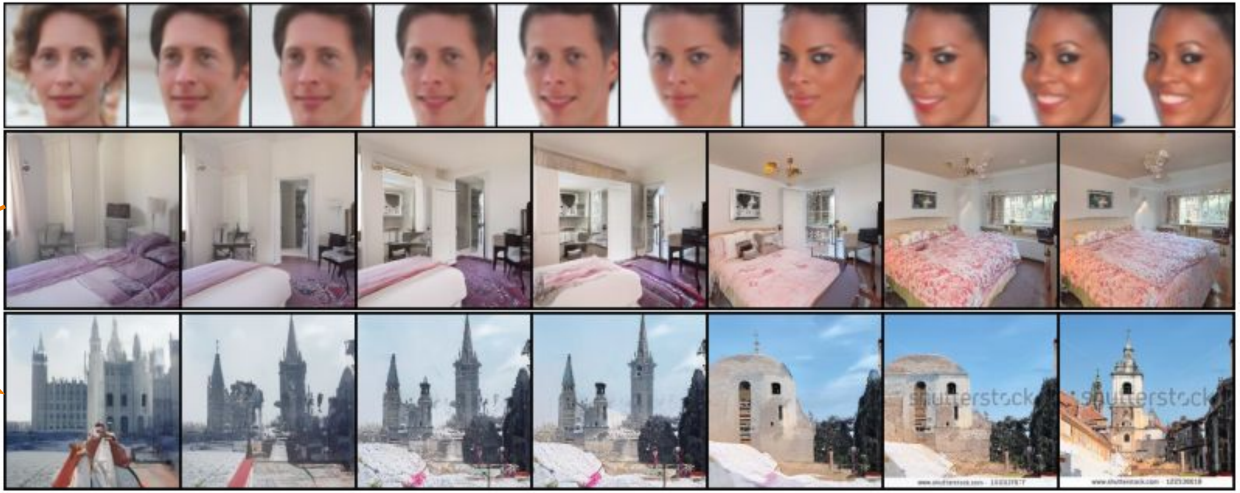
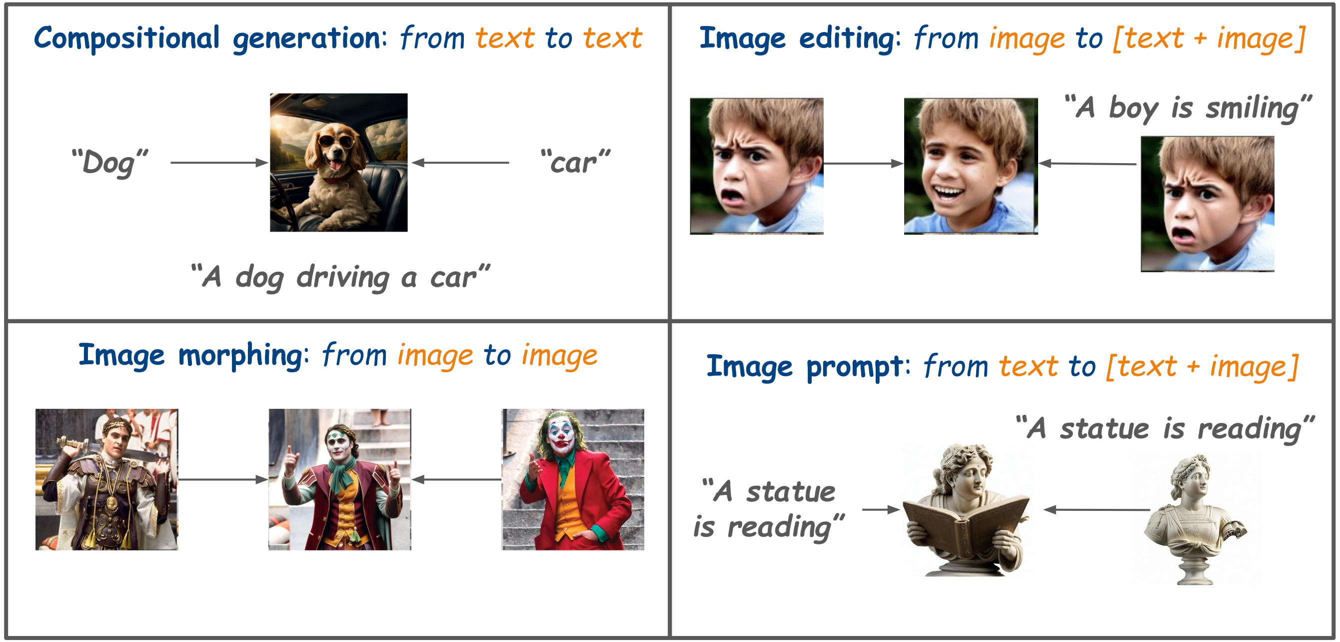


Image from: [DENOISING DIFFUSION IMPLICIT MODELS]

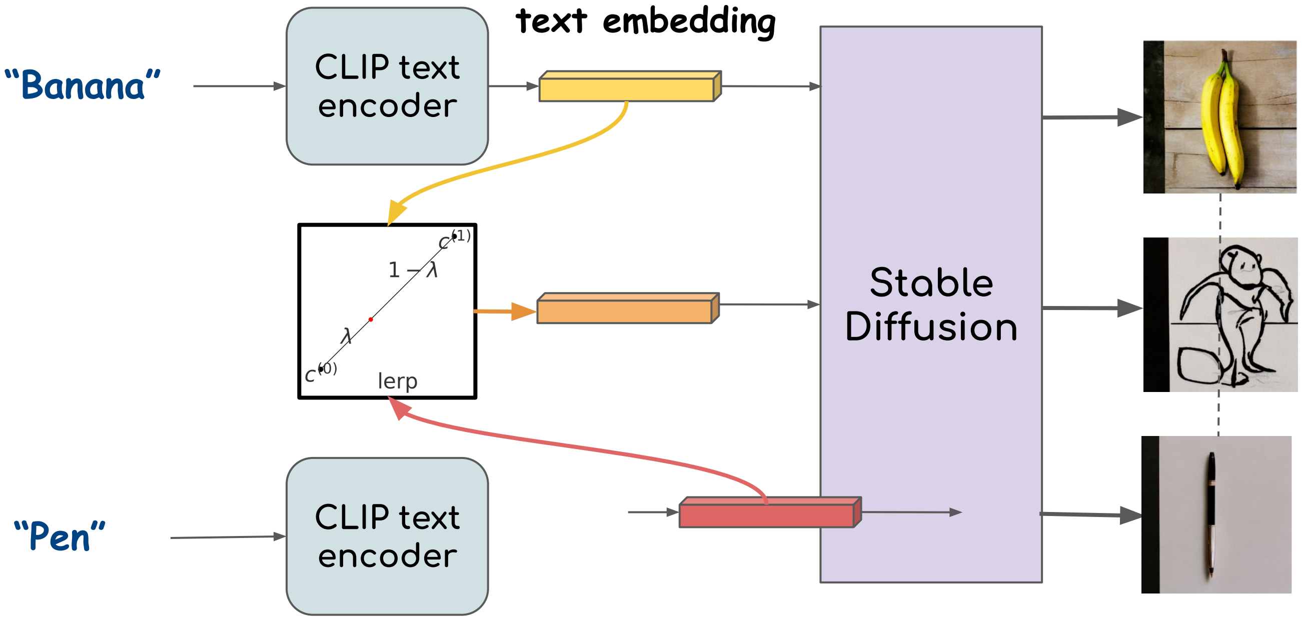
02 Motivation for conditional deep interpolation

- Deep interpolation different two conditions is ***important*** in many applications
- Rarely explored as independent research question



02 Failure on text embedding interpolation

- Text embedding interpolation is implicitly used in many methods for application
- but **not roundly evaluated**
- The interpolated image does not look "***nice***"?



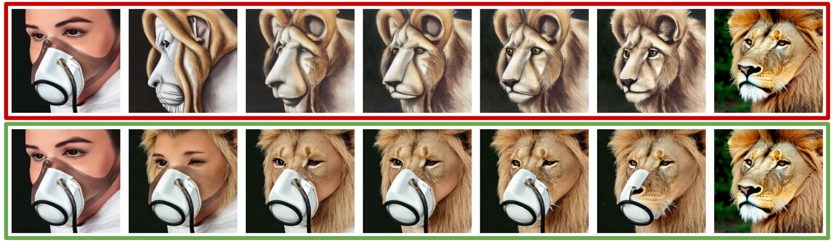
02 Evaluation on the quality of deep interpolation

- This is not a good interpolation intuitively, but what is exactly a “***good***” or “***bad***” interpolation?

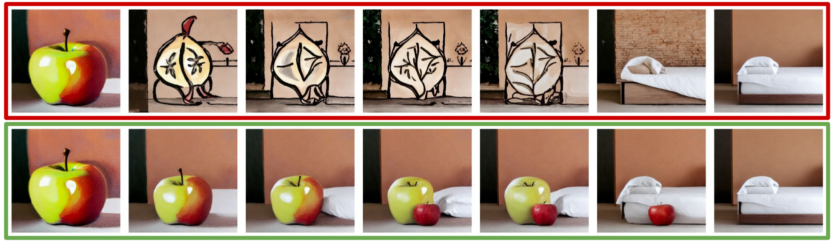
02 Evaluation on the quality of deep interpolation

- Perceptual consistency
  - How direct the transition is?
  - Learned Perceptual Image Patch Similarity (LPIPS) model [1] for perceptual distance

- Perceptual smoothness
  - How smooth the transition is?
  - LPIPS model [1] for perceptual distance



- Fidelity
  - What about the quality of interpolated image?
  - Fréchet Inception Distance (FID) [2] as the quality proxy



03

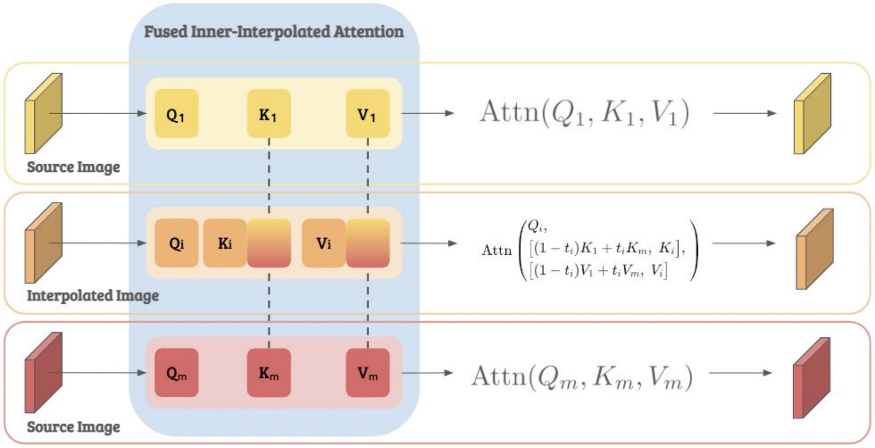
How to improve this?

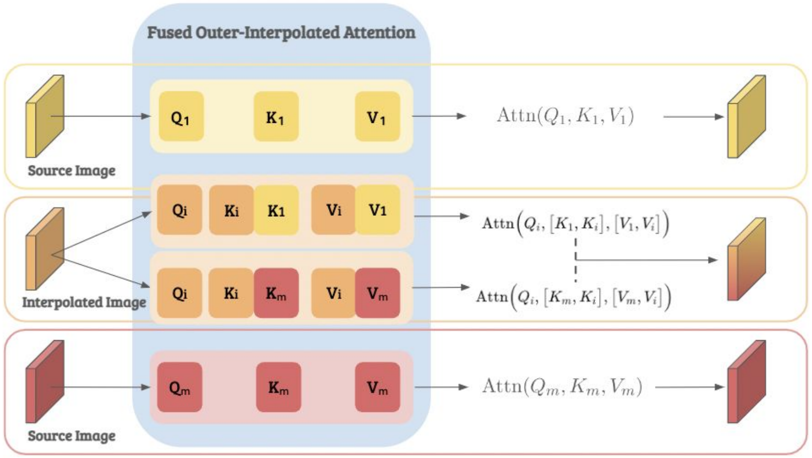
04 PAID: Prompt-guided Attention Interpolation of Diffusion

- Fully training-free
- Fused interpolated attention
  - “Interpolated”: boost consistency
  - “Fused”: boost quality
- Beta prior selection
  - Boost smoothness
- Prompt guidance
  - Enable selecting interpolation path via the third condition

04 PAID Workflow: Fused interpolated attention

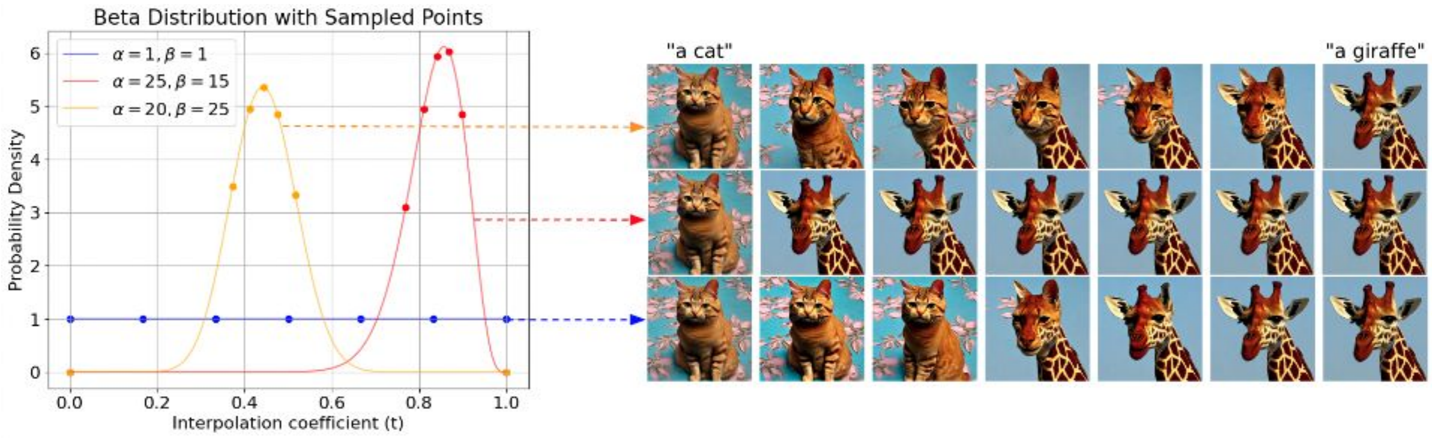
- Expand interpolation to **both cross-attention & self-attention**
- Concatenate with original self-attention to enhance fidelity
- Inner / outer interpolation





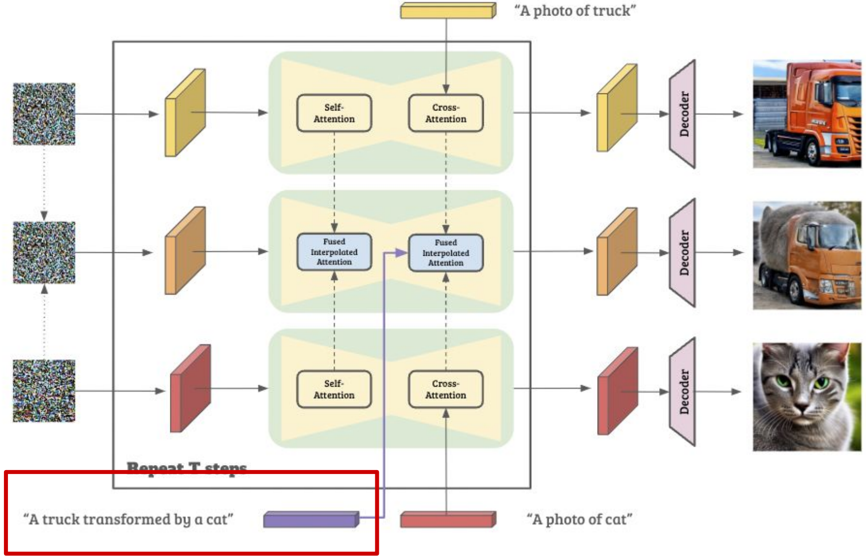
04 PAID Workflow: Beta prior selection

- Beta prior is good for three reasons:
  - Bell shape: encourage more selection in a small range
  - Uniform distributed point as the lower bound
  - Can make up for the bias towards one side
- We use dynamic selection to auto-fit the alpha and beta of the prior



04 PAID Workflow: Prompt guidance

- Since self-attention already poses spatial constraints
    - Loose semantic constraints in cross-attention
- Inject the guidance prompt in cross-attention



(f) *"Photo of a dog"* to *"Photo of a car"*, guided with *"A dog driving car"* (top), *"A car with dog furry texture"* (middle), and *"A toy named dog-car"* (bottom).

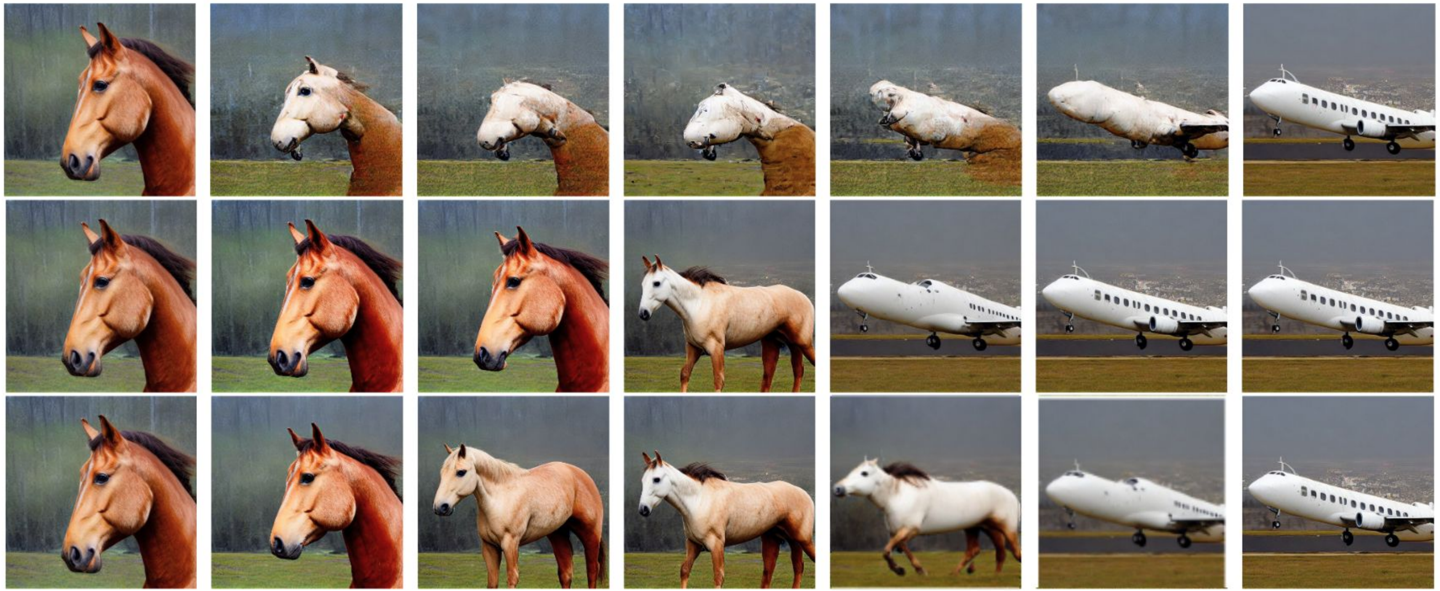
05

Experiment: Ablation Study

attention interpolation

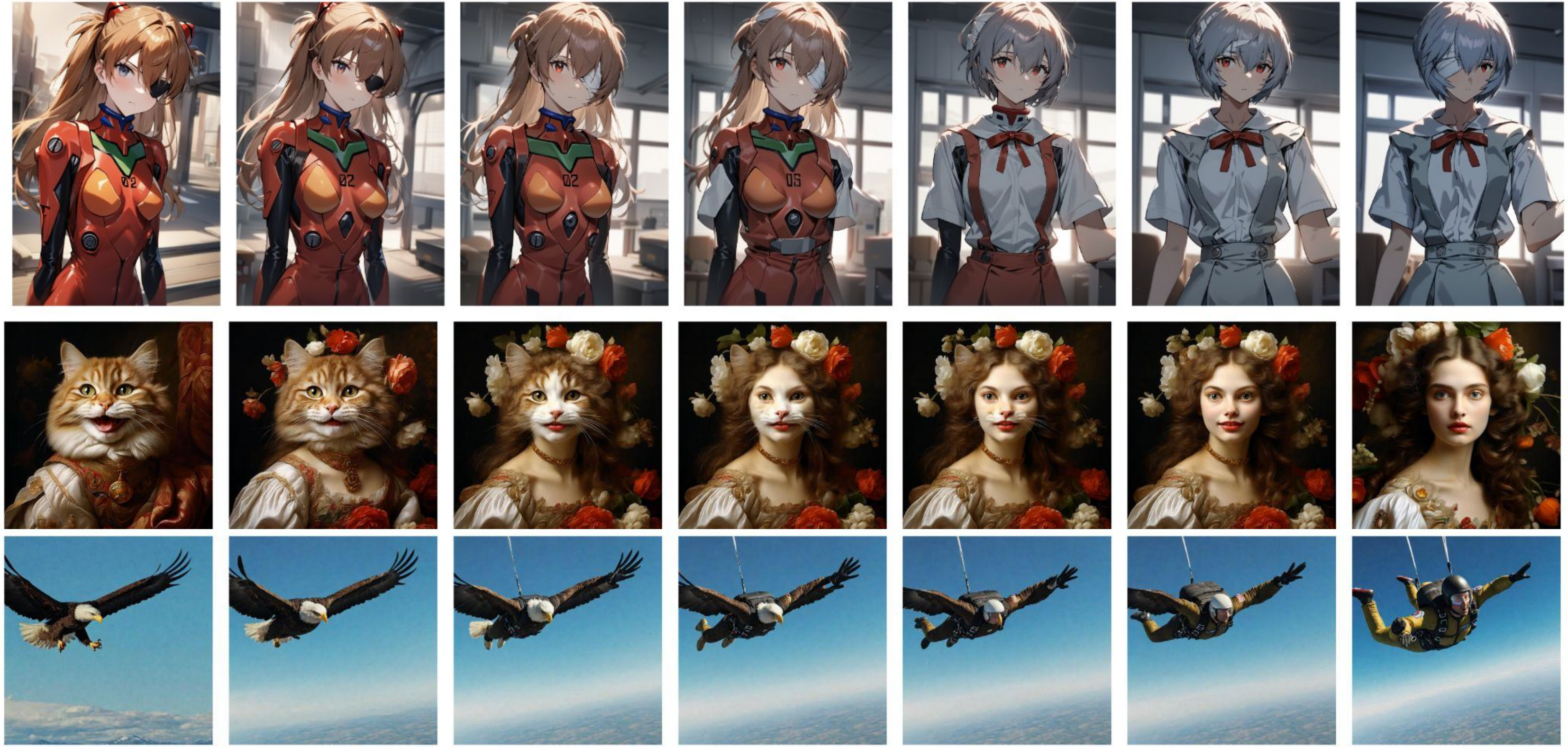
Fused interpolated attention

Fused interpolated attention
+ Beta prior selection



05

Experiment: Qualitative results



05 Experiment: Quantitative Results

- Boost three evaluation metrics by a large margin
  - Especially consistency & fidelity

Dataset	Method	Smoothness (↑)	Consistency (↓)	Fidelity (↓)
CIFAR-10	TEI	0.7531	0.3645	118.05
	DI	0.7564	0.4295	87.13
	AID-O	0.7831	**0.2905***	**51.43***
	AID-I	**0.7861***	0.3271	101.13
LAION-Aesthetics	TEI	0.7424	0.3867	142.38
	DI	0.7511	0.4365	101.31
	AID-O	0.7643	**0.2944***	**82.01***
	AID-I	**0.8152***	0.3787	129.41

05 Experiment: Human study

- Dominantly preferred by human evaluation

Interpolation method	Near Object	Far Object	Scene	Object+Scene
TEI	8.75%	1.16%	0%	1.26%
AID-I	**53.75%**	**50%**	45.2%	45.57%
AID-O	36.25%	46.5%	**50%**	**51.90%**
Hard to determine	1.25%	2.32%	4.76%	1.26%

05 Applications

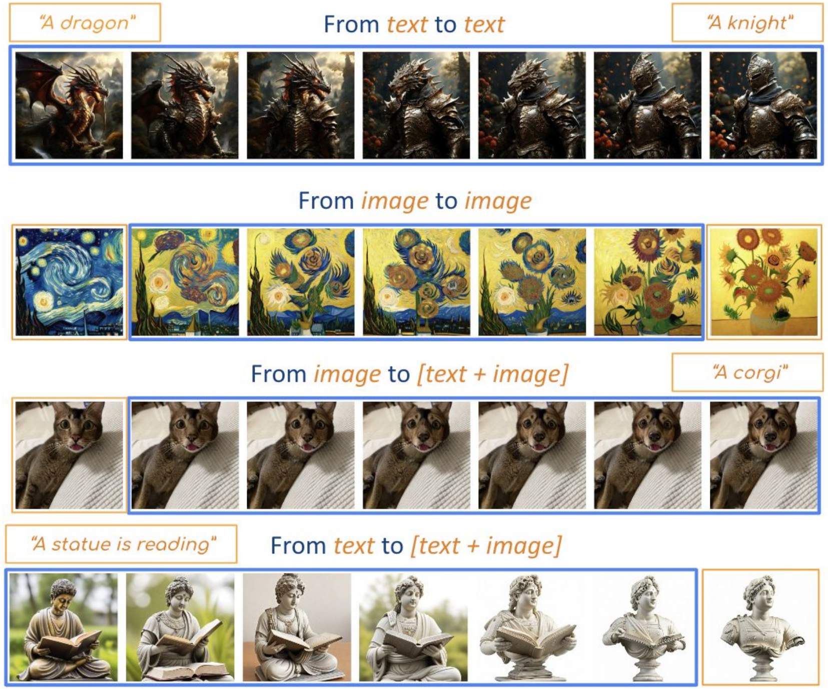


Image morphing

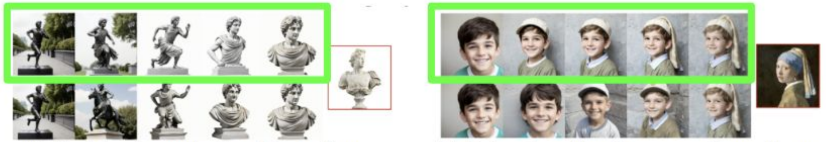
Compositional Generation

Image Editing Control

(a) "A ~~cat~~ dog sitting on the grass."   (b) "A boy is ~~angry~~ happy."

Figure 5: **Results of image editing control.** Our method boosts the controlling ability over editing. The first row of (a) and (b) is generated by P2P + AID while the second row is P2P + TEI.

Control the scale of image prompt



(b) "A statue is running." + global reference   (c) "A boy is smiling." + composition reference

06 Summary

Github Repo →✨

- We formulate a problem called conditional deep interpolation and propose corresponding evaluation metric
- We analyze the behaviour of attention interpolation and tackle the formulated problem by a training-free method PAID
    - Interpolated attention ⟶ consistency
    - Fused interpolated attention ⟶ fidelity
    - Beta prior ⟶ smoothness
    - Prompt guidance ⟶ one-to-many answers
- This approach provides stronger control ability on various application
    - Text-to-Text: compositional generation
    - Image-to-[Text+Image]: image editing
    - Image-to-Image: image morphing
    - Text-to-[Text+Image]: image-control generation