Null-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady* Amir Hertz* Kfir Aberman Yael Pritch
Daniel Cohen-Or

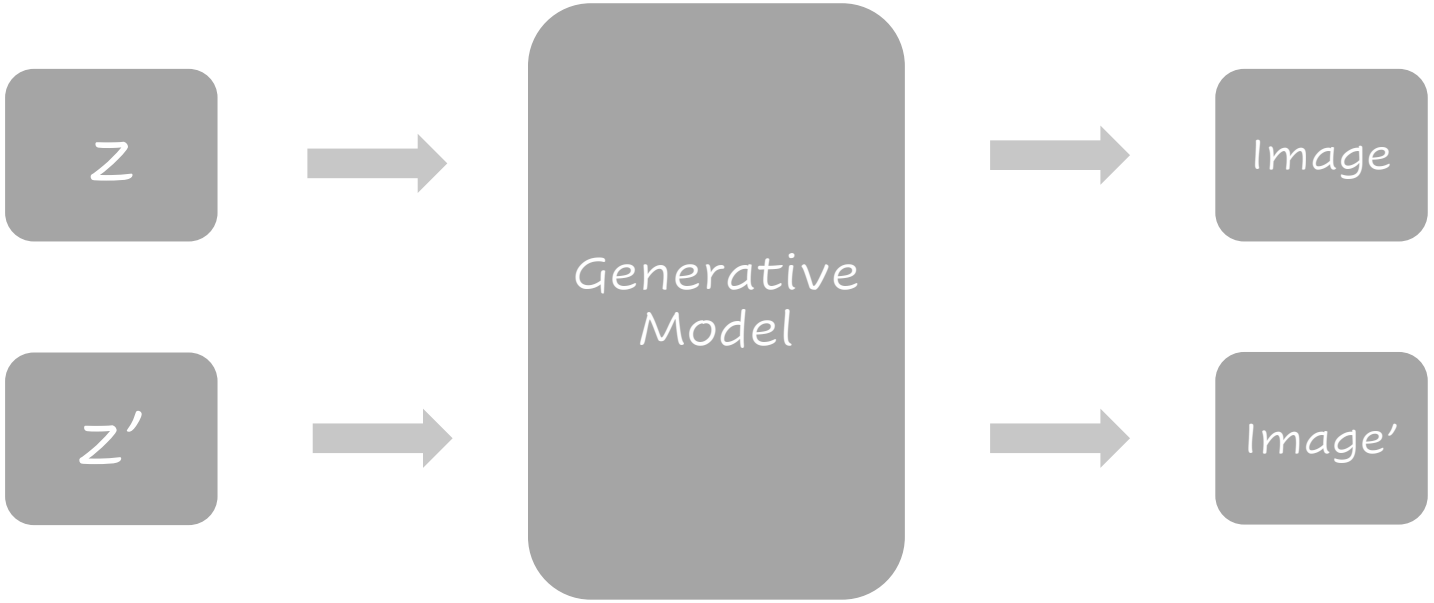
STRUCT Group Seminar
Presenter: Zhengbo Xu
2023.11.05

OUTLINE

- Background
- Method
- Experiments
- Conclusion

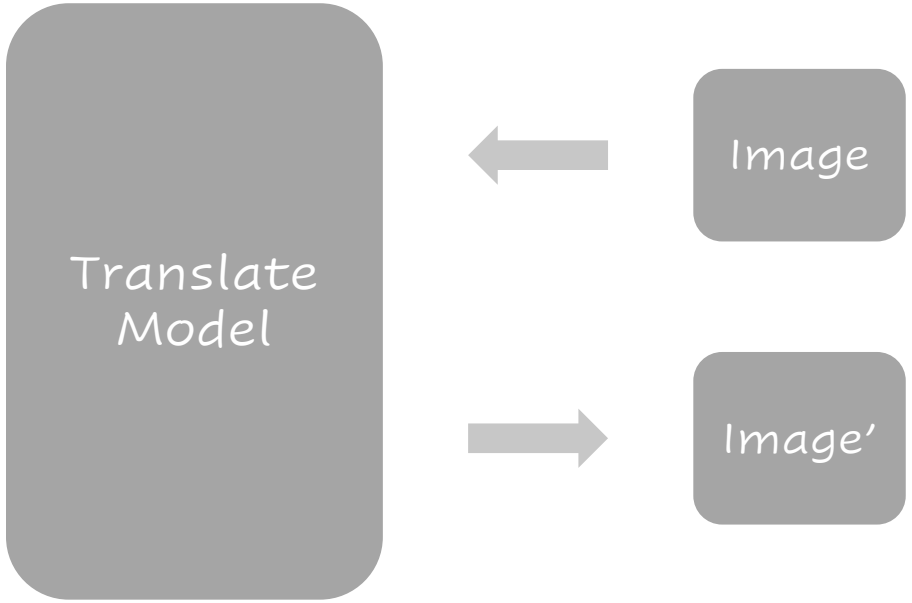
BACKGROUND: Image Editing

Image Translation



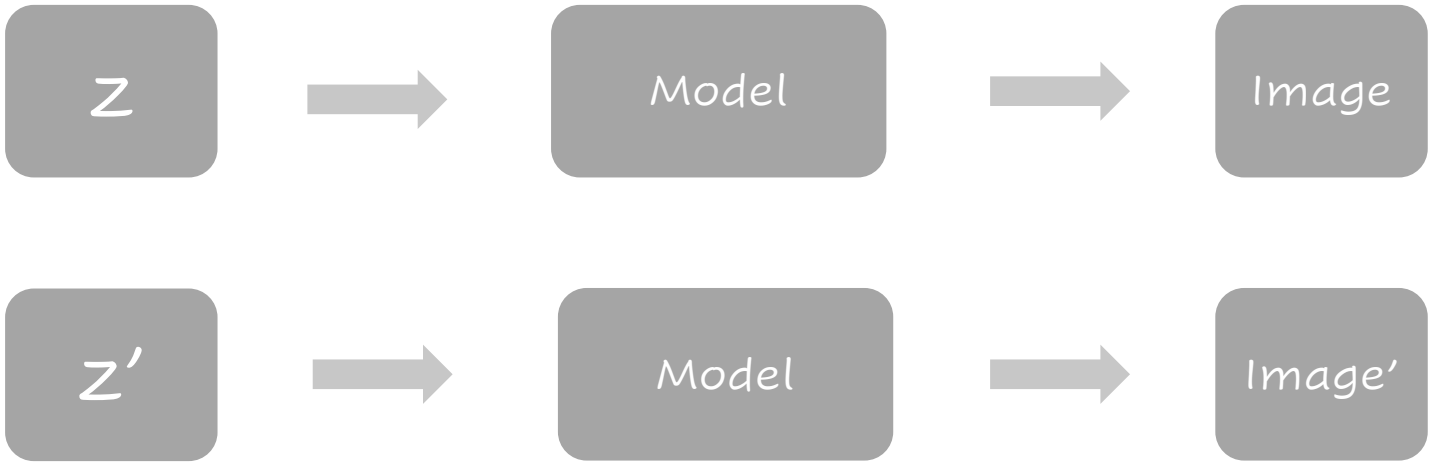
BACKGROUND: Image Editing

Train a Model
CycleGAN/Style Transfer…



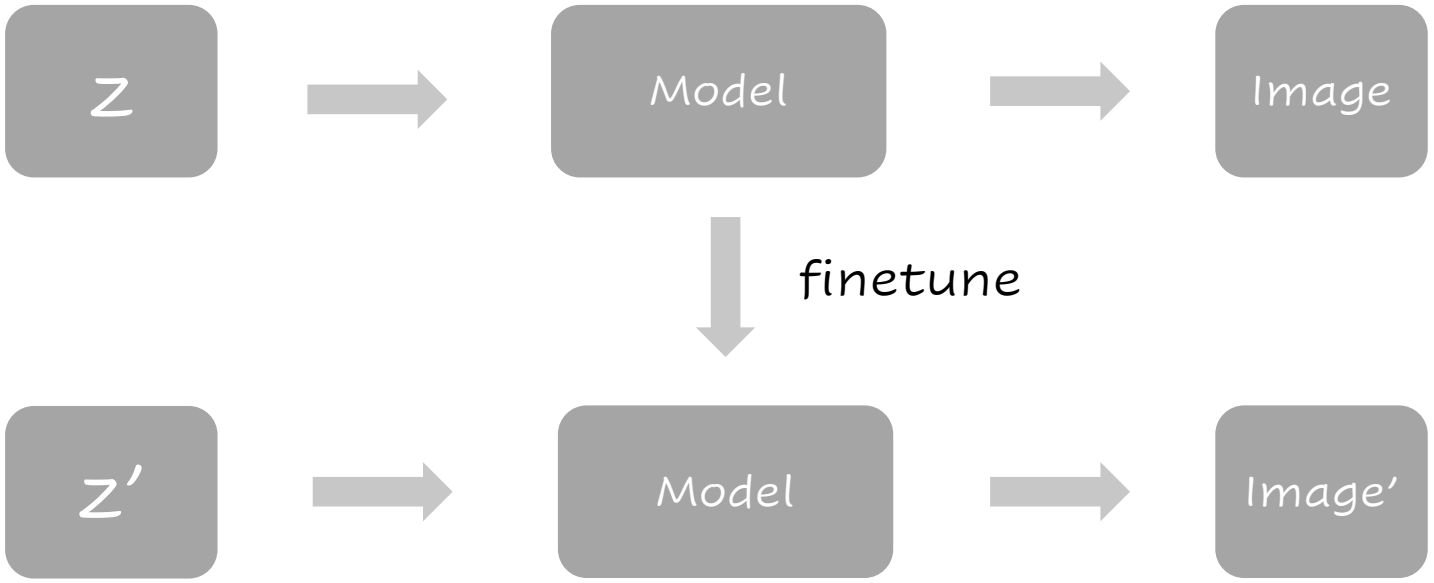
BACKGROUND: Image Editing

Image Translation



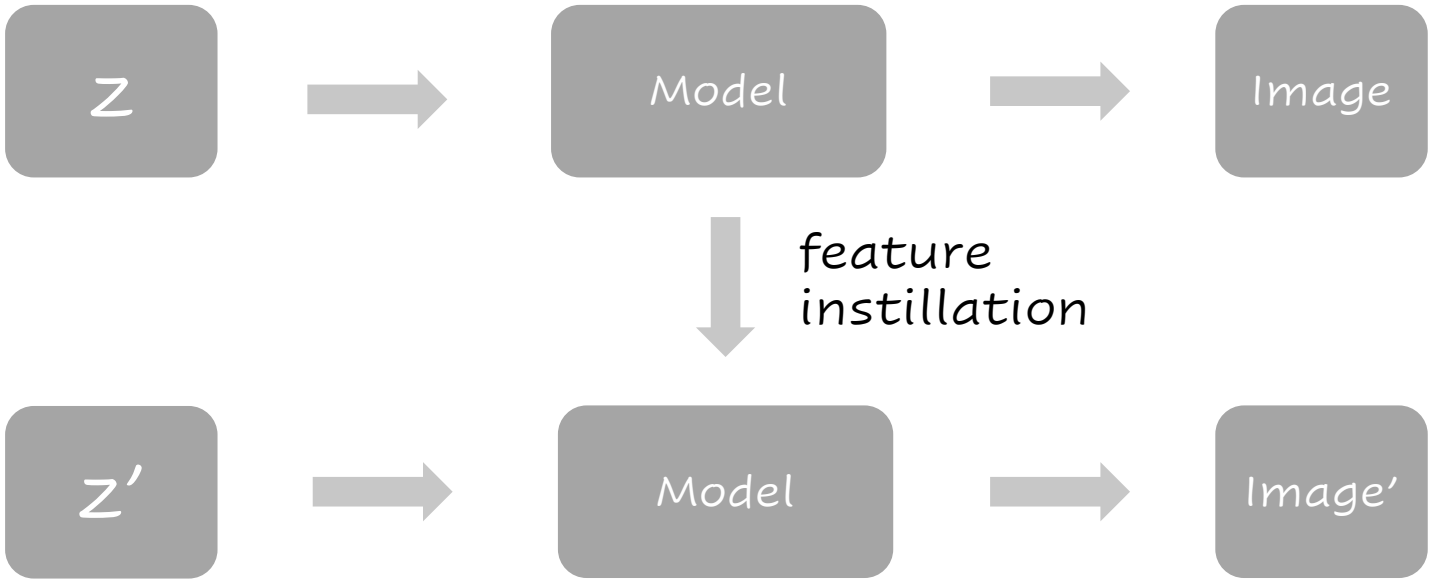
BACKGROUND: Image Editing

Finetune



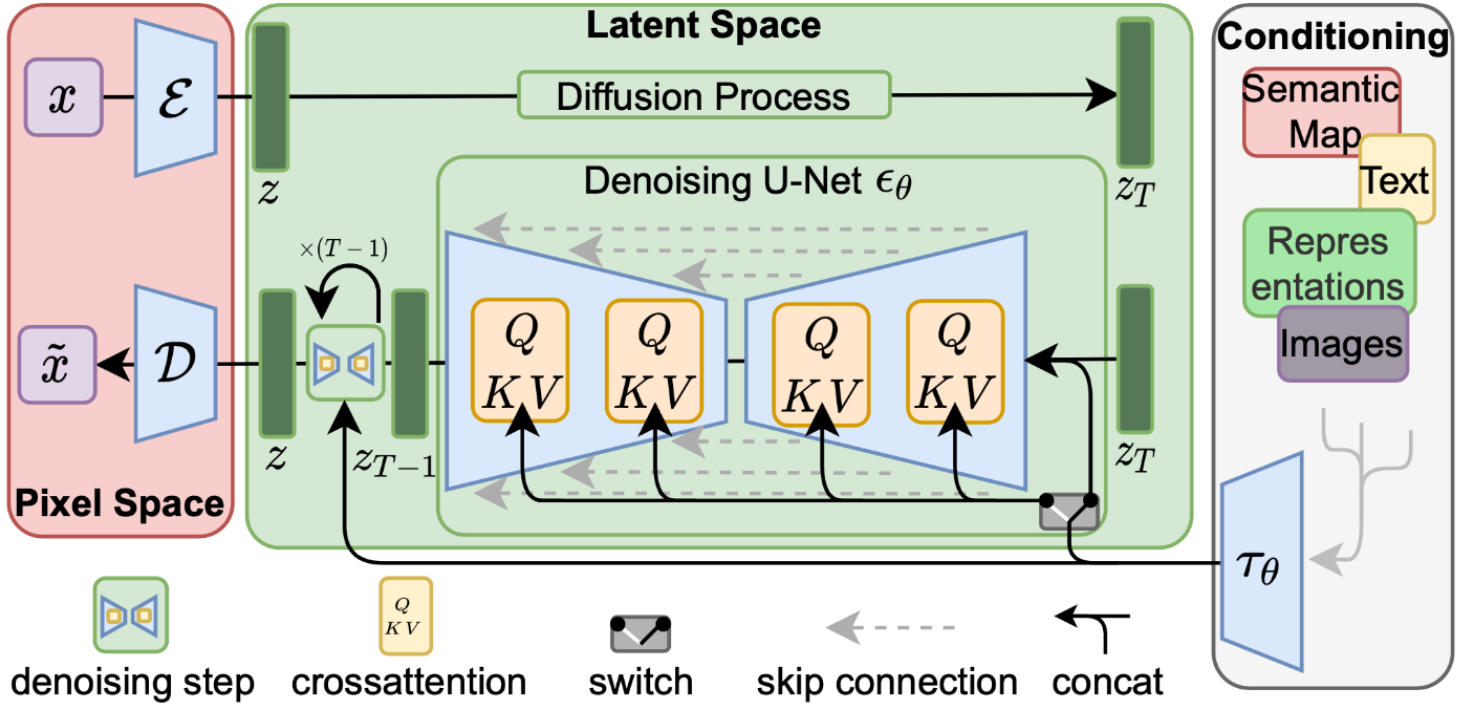
BACKGROUND: Image Editing

Feature Instillation



BACKGROUND: Stable Diffusion

Latent diffusion model



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

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

$$\text{where } \mathbf{Q} = \mathbf{W}_Q^{(i)}\cdot\varphi_i(\mathbf{z}_i),$$

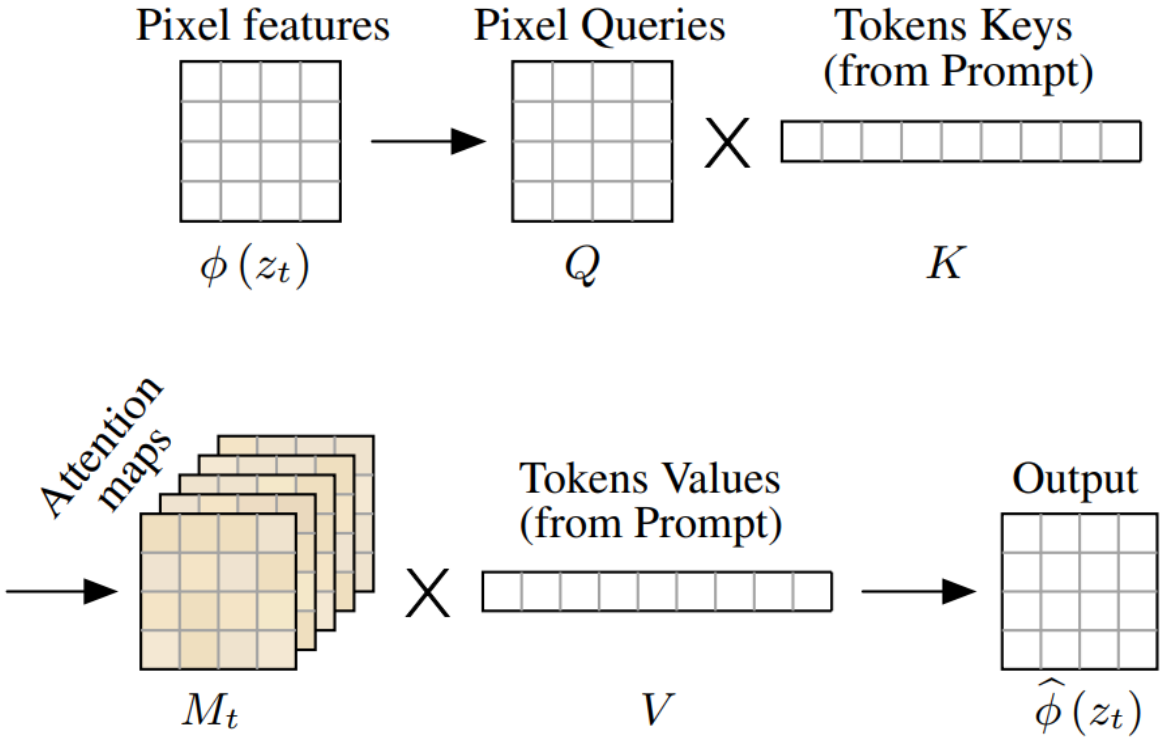
$$\mathbf{K} = \mathbf{W}_K^{(i)}\cdot\tau_\theta(y),$$

$$\mathbf{V} = \mathbf{W}_V^{(i)}\cdot\tau_\theta(y)$$

Attention

BACKGROUND: Spatial Attention

Spatial Attention



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

$$\text{where } \mathbf{Q} = \mathbf{W}_Q^{(i)}\cdot\varphi_i(\mathbf{z}_i),$$

$$\mathbf{K} = \mathbf{W}_K^{(i)}\cdot\tau_\theta(y),$$

$$\mathbf{V} = \mathbf{W}_V^{(i)}\cdot\tau_\theta(y)$$

Attention

BACKGROUND: DDIM Inversion

Reverse Direction

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

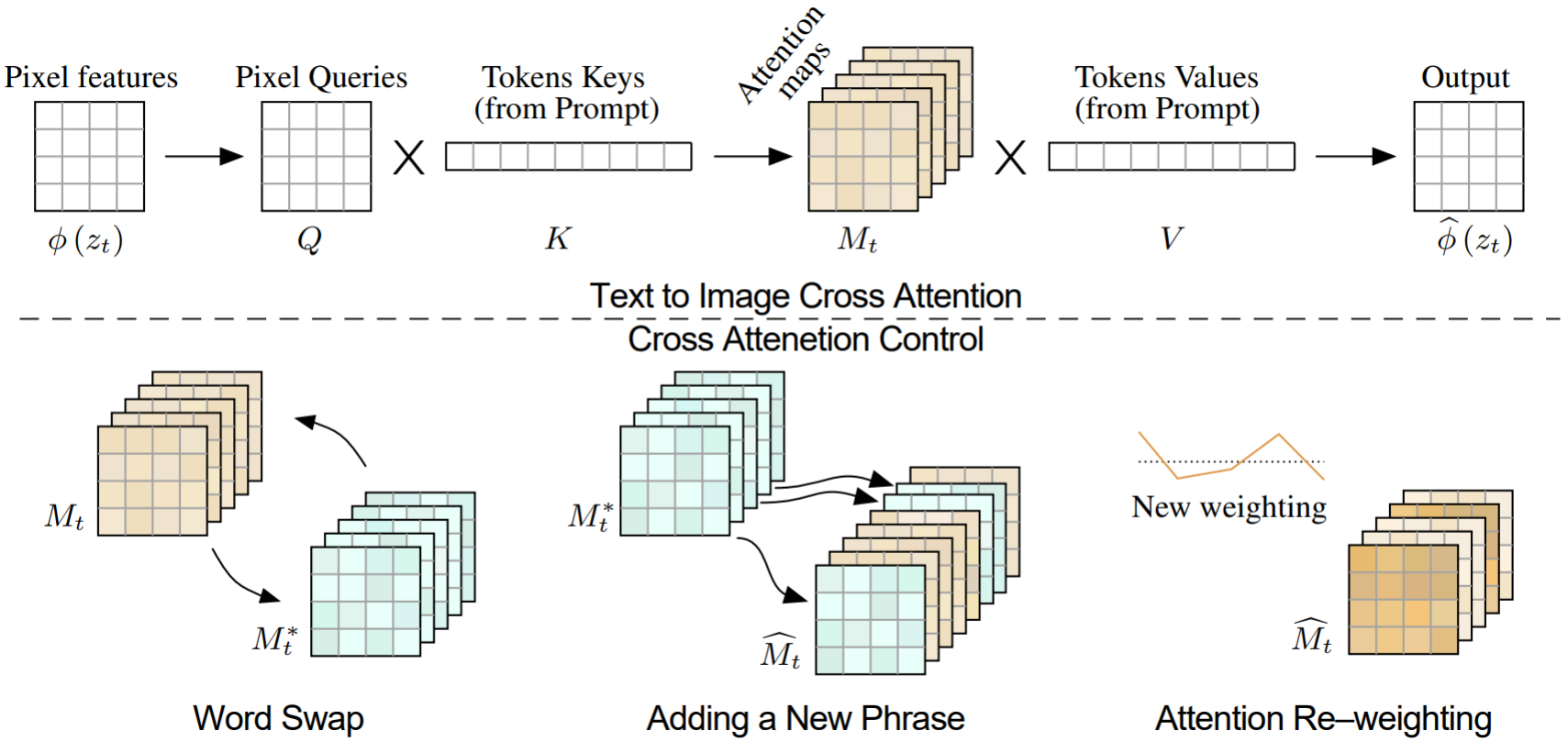
$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

OUTLINE

- Background
- Method
- Experiments
- Conclusion

METHOD

Prompt-to-Prompt

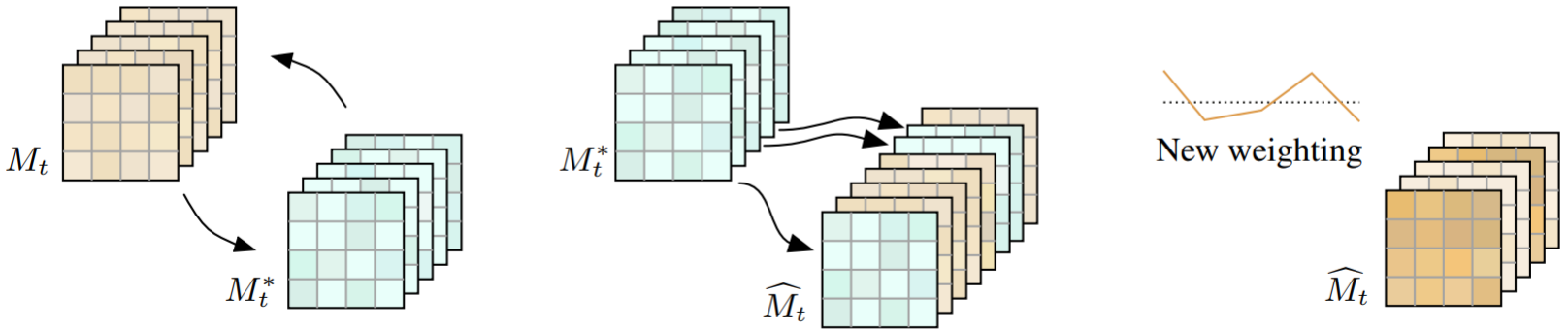


METHOD

Prompt-to-Prompt

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases} \qquad (Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

Cross Attenetion Control



Word Swap | Adding a New Phrase | Attention Re–weighting

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

BACKGROUND: DDIM Inversion

Reverse Direction

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

reverse

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

METHOD

Problem: Classifier-free Guidance

In Stable Diffusion, we have

$$\tilde{\varepsilon}_\theta(z_t, t, \mathcal{C}, \varnothing) = w \cdot \varepsilon_\theta(z_t, t, \mathcal{C}) + (1 - w) \cdot \varepsilon_\theta(z_t, t, \varnothing).$$

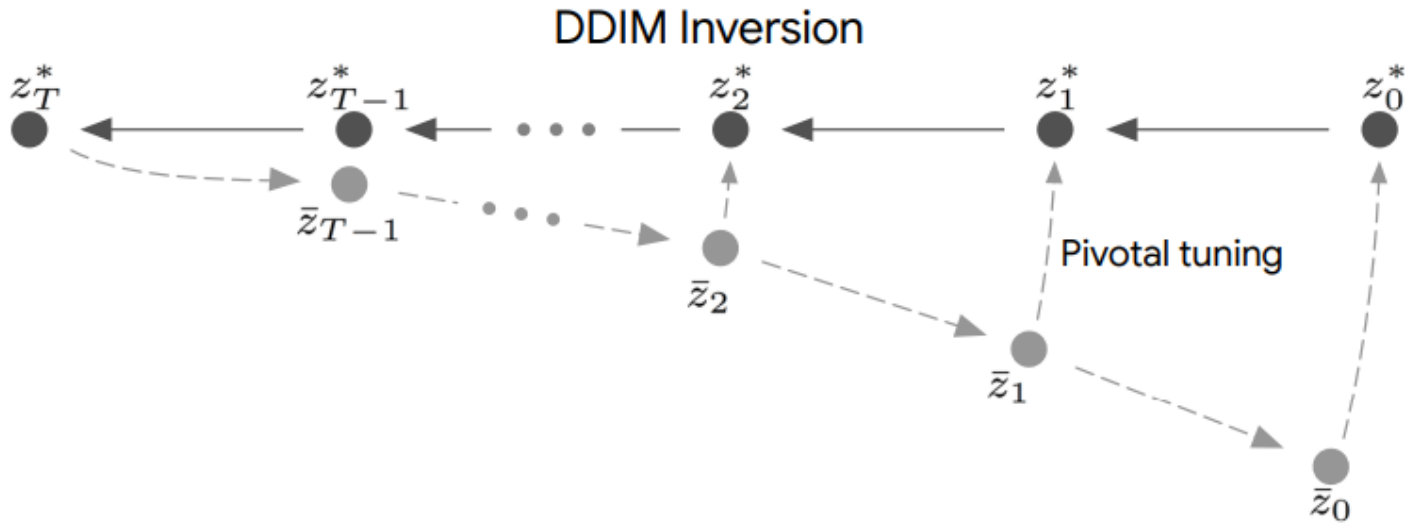
So we cannot use DDIM Inversion directly.

Tips: Just adding the weight does not work.

METHOD

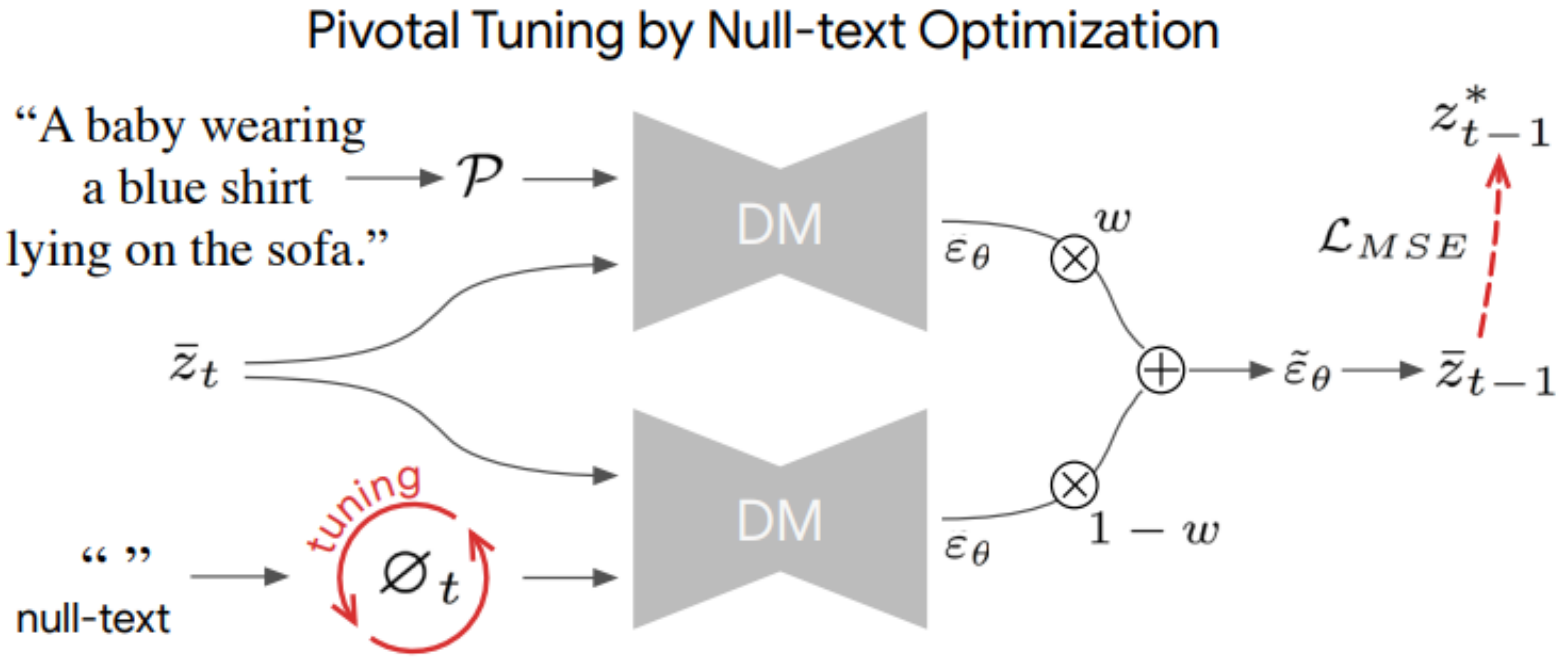
Considering a simple situation:

$$\tilde{\varepsilon}_\theta(z_t, t, \mathcal{C}, \varnothing) = w \cdot \varepsilon_\theta(z_t, t, \mathcal{C}) + (1 - w) \cdot \varepsilon_\theta(z_t, t, \varnothing).$$



METHOD

Change the null-text input:



OUTLINE

- Background
- Method
- Experiments
- Conclusion

Ablation Study

VQAE: Only encoder-decoder

Random Pivot: Use latents when adding noises

Global null-text embedding: One null-text latent for all steps

Random Caption: Random prompts

Textual Inversion: Optimize C

Input caption: "A black dinning room table sitting in a yellow dinning room."

Input Image | DDIM inversion | VQAE reconstruction | Textual inversion | Random pivot | Global null-text | Random caption | Null-text (ours)

Ablation Study

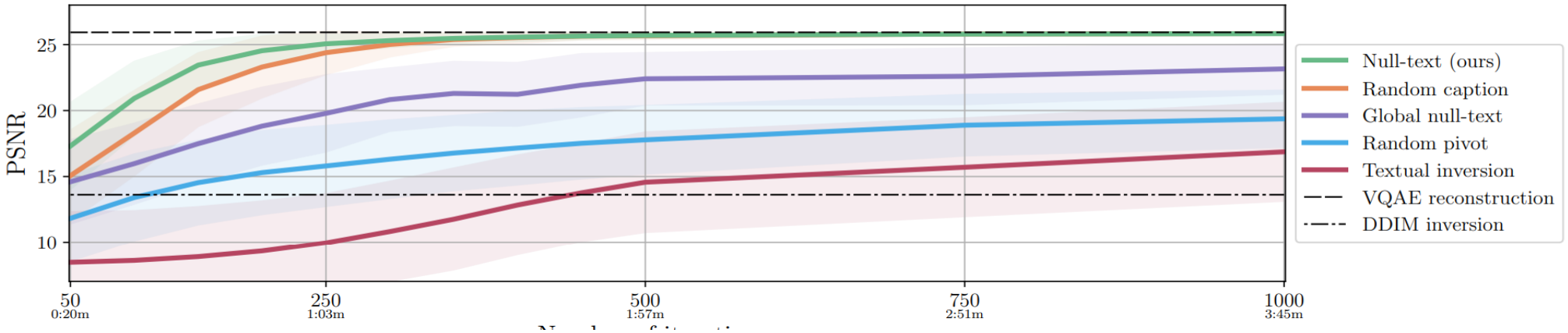
VQAE: Only encoder-decoder

Random Pivot: Use latents when adding noises

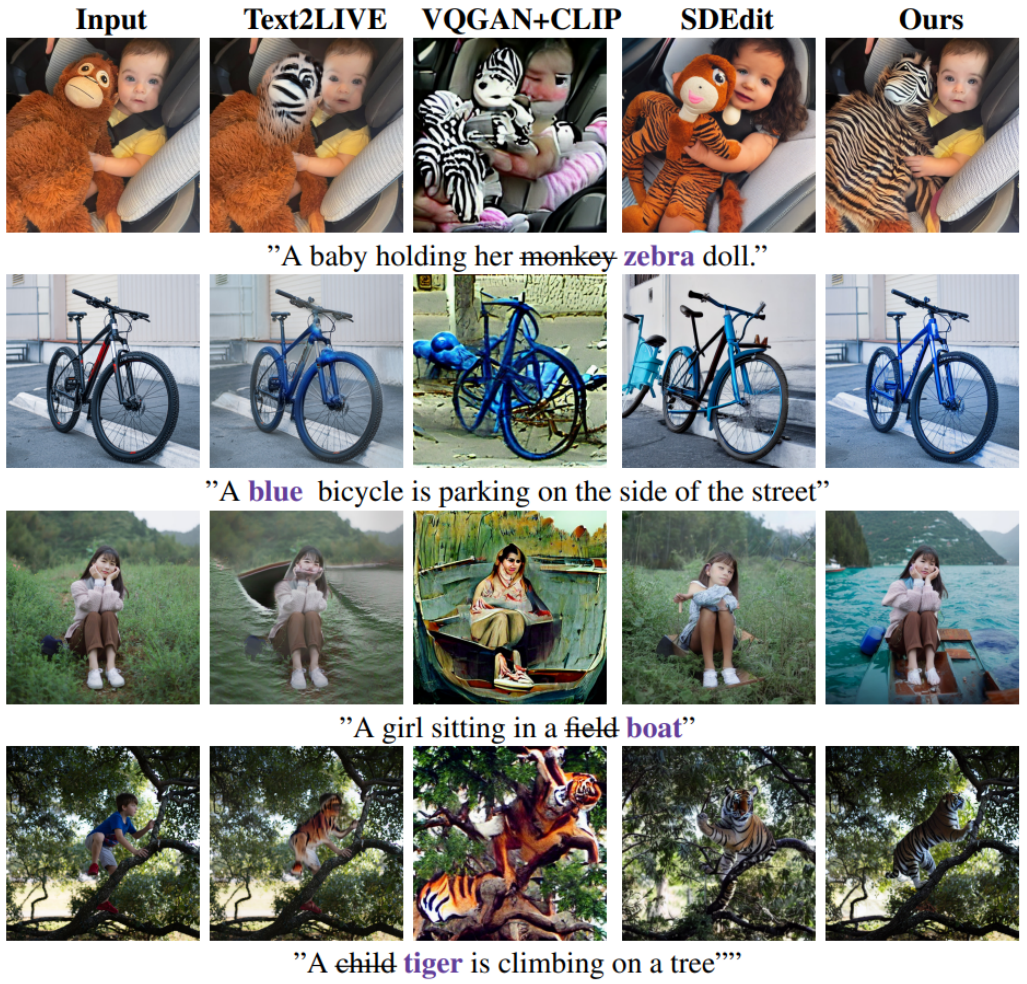
Global null-text embedding: One null-text latent for all steps

Random Caption: Random prompts

Textual Inversion: Optimize C



EXPERIMENTS



EXPERIMENTS

Input Image **Modifed caption:** "A girl sitting in a dry field."

Input Image **Modifed caption:** "A living room with a zebra dense pattern couch and pillows"

Input caption: "Two crochet birds sitting on a branch."

Input caption: "A basket with apples kittens on a chair."

Input Image+Mask **Blended–Diffusion** **Glide** **SD Inpainting** **Ours**

OUTLINE

- Background
- Method
- Experiments
- Conclusion

CONCLUSION

- A way for image editing on diffusion model
- A way for finding latent on Stable Diffusion Model
- A good experiment result

Thanks for listening!