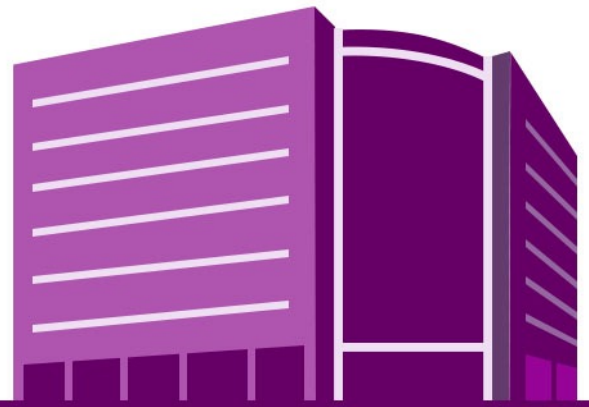
多模态可控生成的“道”与“术”

——对可控文本、图像、多模态内容生成的前沿探讨与深度思考

Summer Workshop

马志远
清华大学，协同交互智能研究中心(C3I)
2023年7月28日



Outline

- **（战略）**可控多模态生成（CMG）方向的研究热点及趋势：

  多模态统一、多条件控制、多目标编辑、个性化定制、高效扩散、可解释性、复杂场景生成（知识交互、推理能力）……

- **（基本战术）**扩散模型和大模型高效微调的原则、进展和实践（背景技术）：

  DDPM、DDIM、Inversion、Sampling、Latent Diffusion、ControlNet、Adapter、LoRa、Unified……

- **（进阶战术）**可控生成的四大“法宝”（主流方案）：

  Adapter微调（ControlNet）、Prompting微调（DreamBooth）、参数编辑（Prompt2Prompt）、新约束优化（Imagic）

- **（修炼特技）**关于可控多模态生成的创新性思考和具体规划

  非刚性编辑、创意场景图文生成、文本的视觉化生成、新的高效扩散方式、逻辑和推理驱动的视觉生成

- **（用武之地）**可控多模态生成（CMG）方向的前沿应用

  对象消失术、创意内容的数字签名、海报的排版生成、文本艺术字生成、high-level内容生成（视频、3D、VR、虚拟场景）

➢ 1-1. 文本的可控生成： 条件概率模型框架（CTRL）

CTRL: A Conditional Transformer Language Model

- Given sequences $x = (x_1, \ldots, x_n)$ where each x_i comes from a fixed set of symbols, the goal of language modeling is to learn $p(x)$:

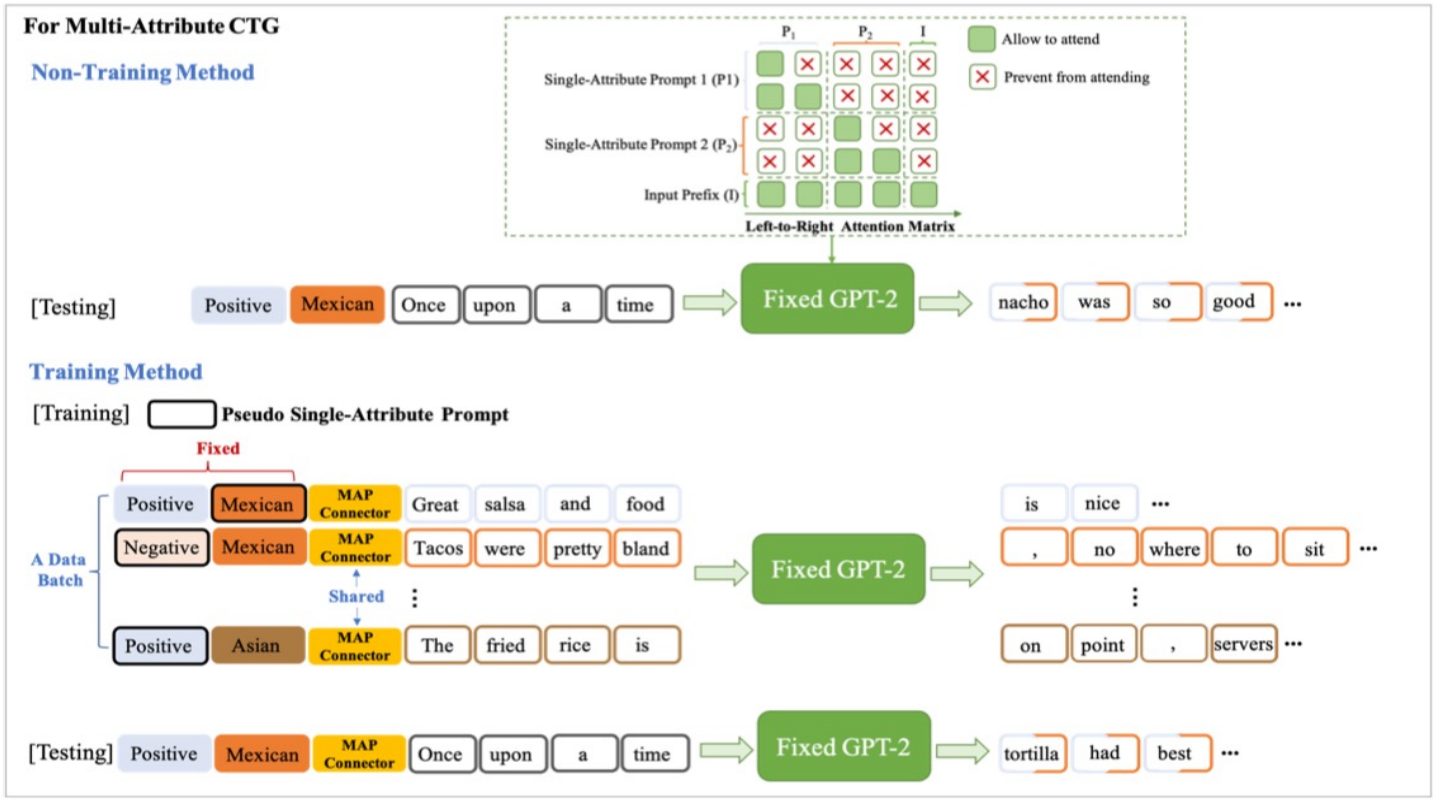
$$p(x) = \prod_{i=1}^{n} p(x_i | x_{<i})$$

- CTRL is a conditional language model that is always conditioned on a control code c and learns the distribution $p(x|c)$:

$$p(x|c) = \prod_{i=1}^{n} p(x_i | x_{<i}, c) \qquad \mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(x_i^k | x_{<i}^k, c^k)$$

【1】CTRL: A Conditional Transformer Language Model for Controllable Generation, Salesforce, Arxiv, 2019 (Citations 824)
【2】Tailor: A Prompt-Based Approach to Attribute-Based Controlled Text Generation, Alibaba, Arxiv, 2022 (Citations 15)

➢ 1-2. 文本的可控生成（CTG）：多属性条件下的可控文本生成

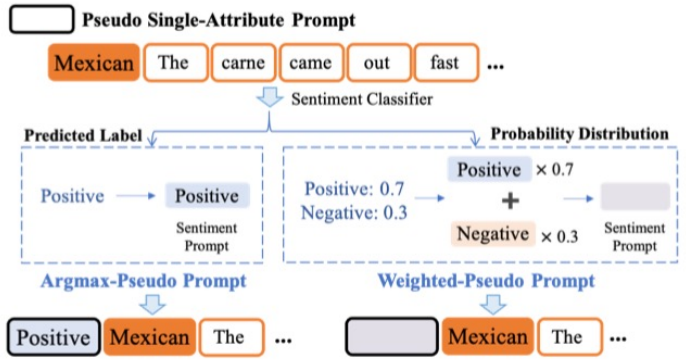


◘ 基于多属性控制条件的提示微调

$$\mathcal{L}_{single} = \sum_{t=1}^{n} \log P_{\theta_g;\theta_{S_k}}(x_t|S_k, x_{<t})$$

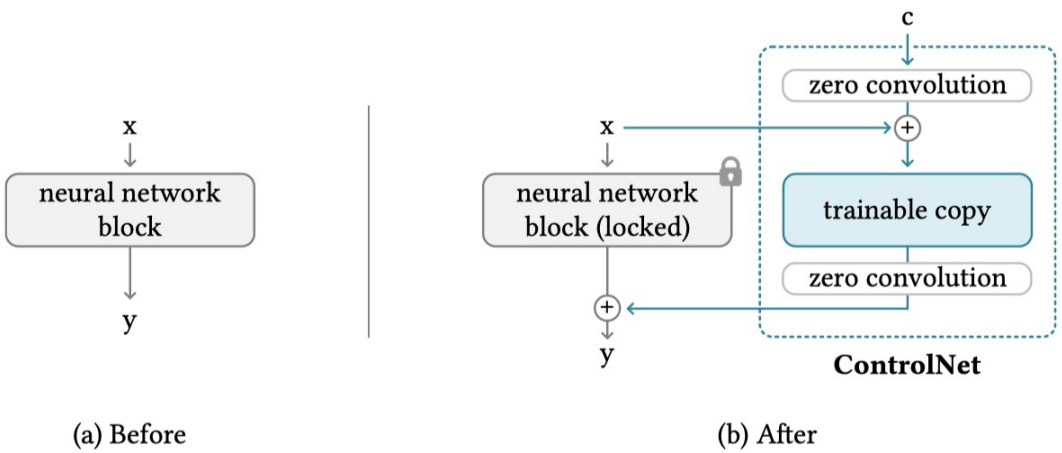
$$A = \text{Softmax}(\frac{QK^\top}{\sqrt{d}} + M_p) \in \mathbb{R}^{(l_p+n)\times(l_p+n)},$$

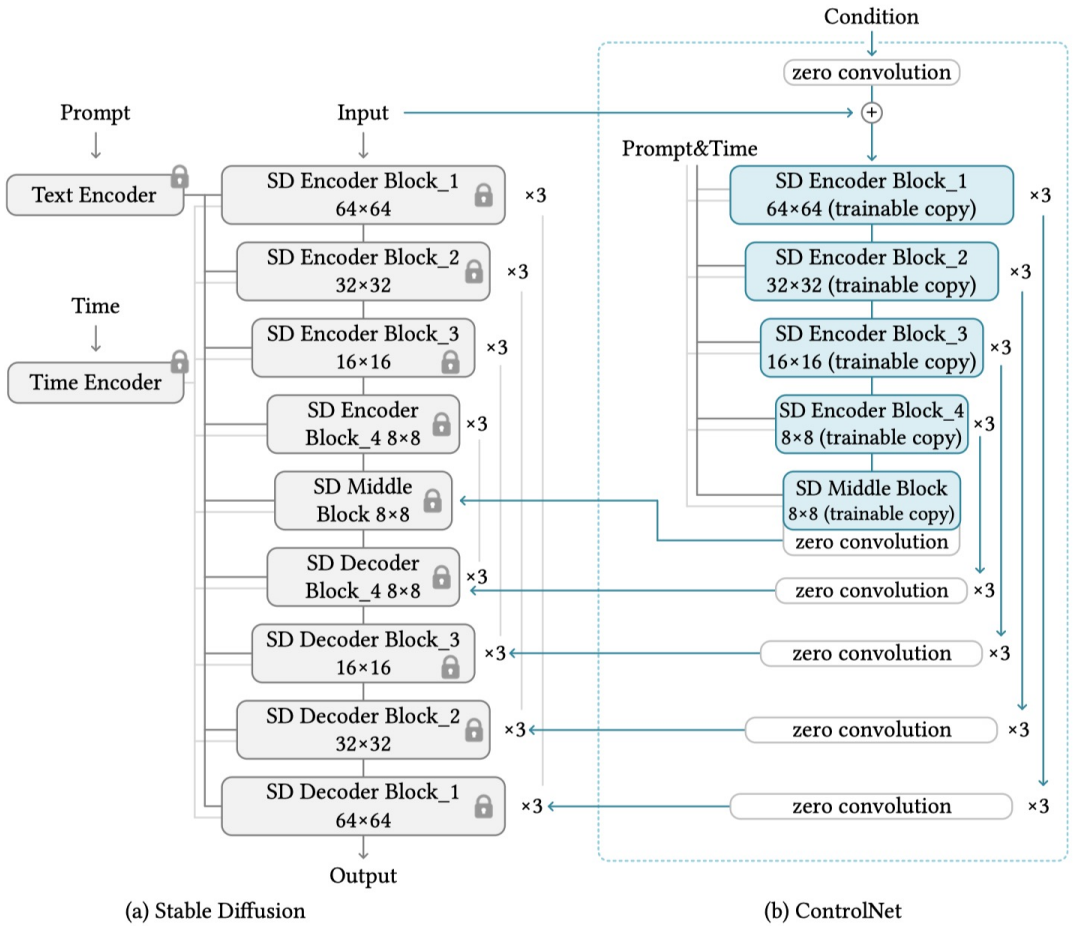
$$M_p^{ij} = \begin{cases} -\infty & i \in [l_u, l_v] \text{ and } j \in [0, l_u], \\ 0 & \text{otherwise}, \end{cases}$$



【1】CTRL: A Conditional Transformer Language Model for Controllable Generation, Salesforce, Arxiv, 2019 (Citations 824)
【2】Tailor: A Prompt-Based Approach to Attribute-Based Controlled Text Generation, Alibaba, Arxiv, 2022 (Citations 15)

➢ 2-1. 图像的可控生成： 鲁棒、高效和有效的可控文生图框架（ControlNet）





$$\boxed{\ } \quad \boldsymbol{y} = \mathcal{F}(\boldsymbol{x};\Theta)$$

$$\boxed{\ } \quad \boldsymbol{y}_c = \mathcal{F}(\boldsymbol{x};\Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c};\Theta_{z1});\Theta_c);\Theta_{z2})$$

$$\boldsymbol{y}_c = \boldsymbol{y} \quad \Rightarrow \quad \begin{cases} \mathcal{Z}(\boldsymbol{c};\Theta_{z1}) = \boldsymbol{0} \\ \mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c};\Theta_{z1});\Theta_c) = \mathcal{F}(\boldsymbol{x};\Theta_c) = \mathcal{F}(\boldsymbol{x};\Theta) \\ \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c};\Theta_{z1});\Theta_c);\Theta_{z2}) = \mathcal{Z}(\mathcal{F}(\boldsymbol{x};\Theta_c);\Theta_{z2}) = \boldsymbol{0} \end{cases}$$

【1】High-Resolution Image Synthesis with Latent Diffusion Models, Heidelberg University, CVPR, 2022 (Citations 2004)
【2】Adding Conditional Control to Text-to-Image Diffusion Models, Stanford University, Arxiv, 2023 (Citations 179)

➢ 2-2. 图像的可控生成： 多条件的统一控制生成（UniControl）

- (Base) Unconditional Target

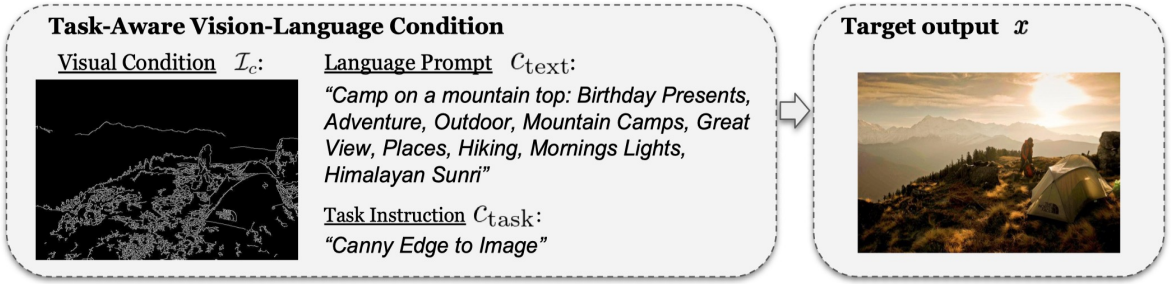
$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}^x} \|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2,$$

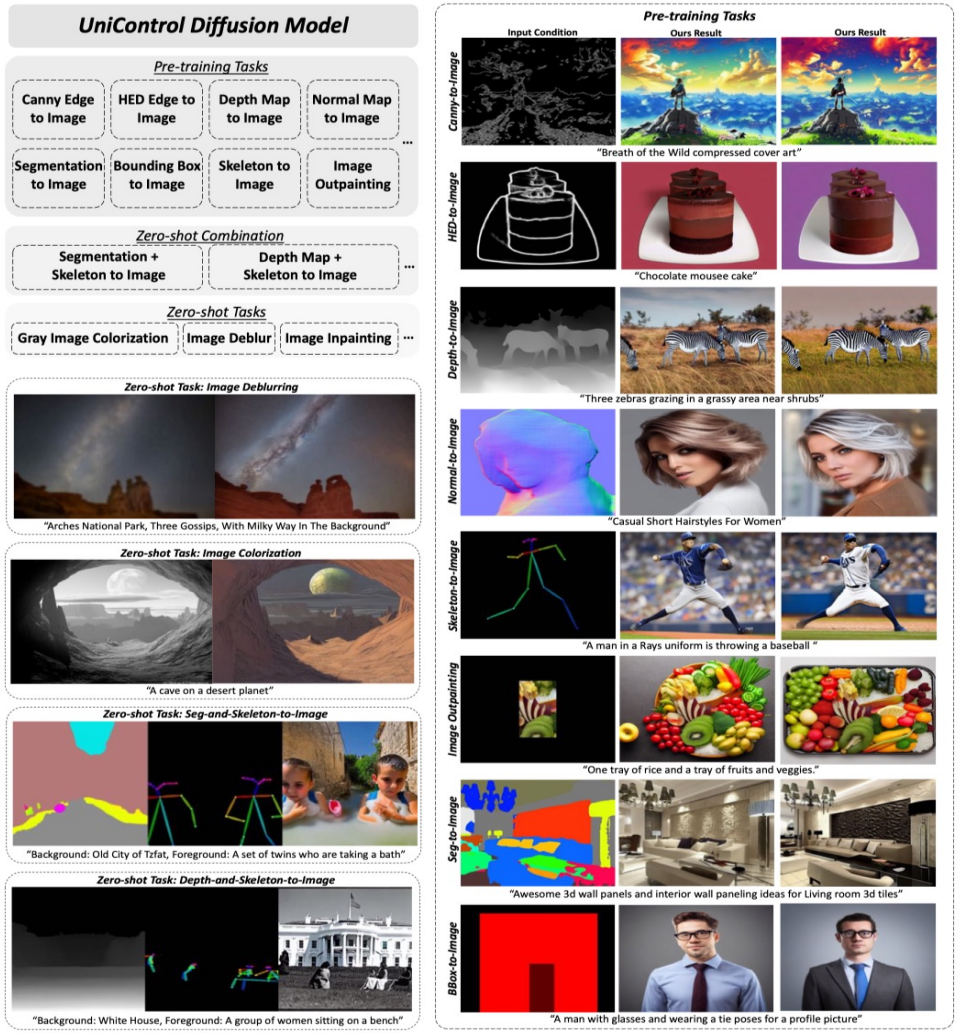
- UniControl Target

Diffusion Step　　Visual Condition

$$\ell^k(\theta) := \mathbb{E}_{z,\varepsilon,t,c_{\text{task}},c_{\text{text}},\mathcal{I}_c} \left[\|\varepsilon - \varepsilon_\theta(z_t, t, c_{\text{task}}, c_{\text{text}}, \mathcal{I}_c)\|_2^2\right]$$

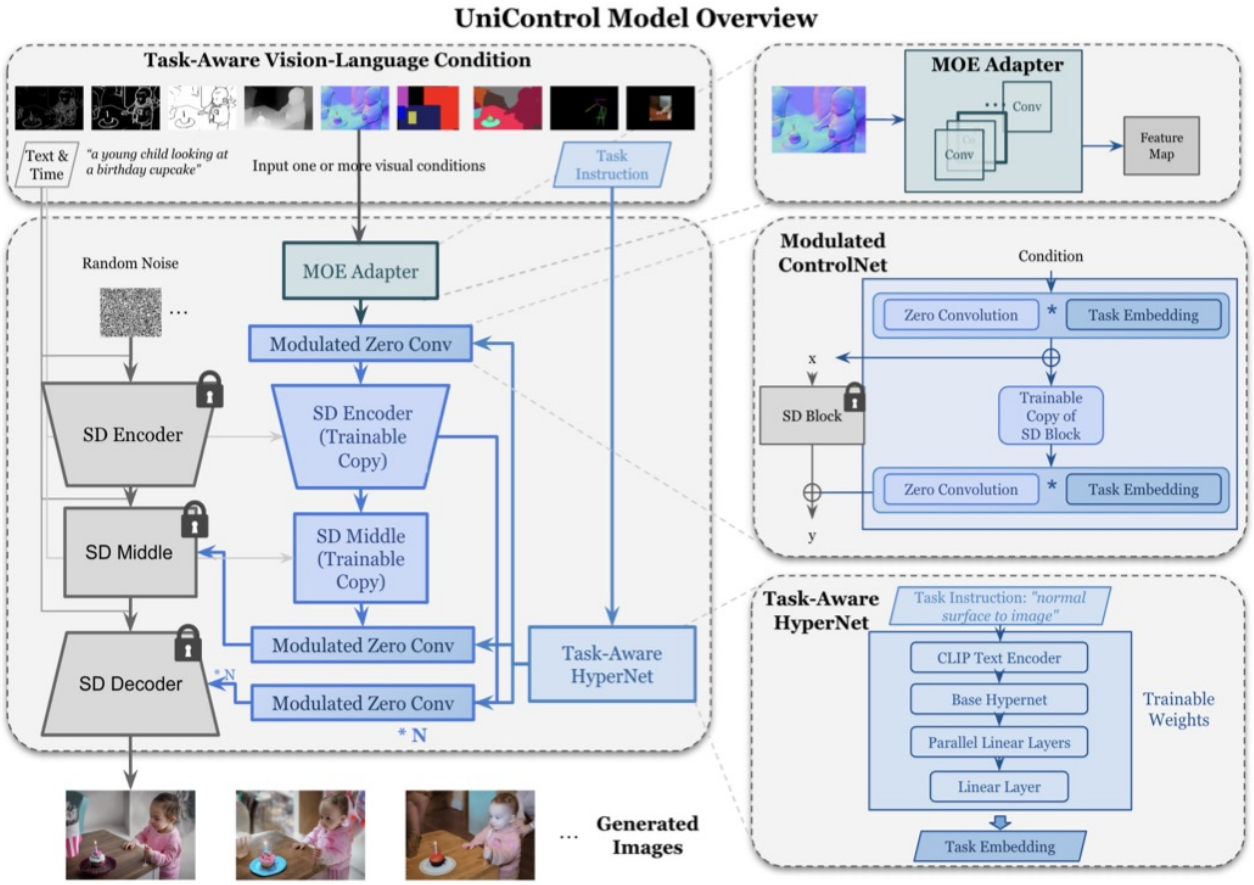
Task Instruction　　Text Prompt





【1】UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild, Salesforce, Arxiv, 2023 (Citations 1)

➢ 2-2. 图像的可控生成： 多条件的统一控制生成（UniControl）



【1】UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild, Salesforce, Arxiv, 2023 (Citations 1)

☐ MOE-Style Adapter

$$\mathcal{F}_{\text{Adapter}}(\mathcal{I}_c^k) := \sum_{i=1}^{K} \mathbb{1}(i == k) \cdot \mathcal{F}_{\text{Cov1}}^{(i)} \circ \mathcal{F}_{\text{Cov2}}^{(i)}(\mathcal{I}_c^k)$$

Indicator function

Conditioned image

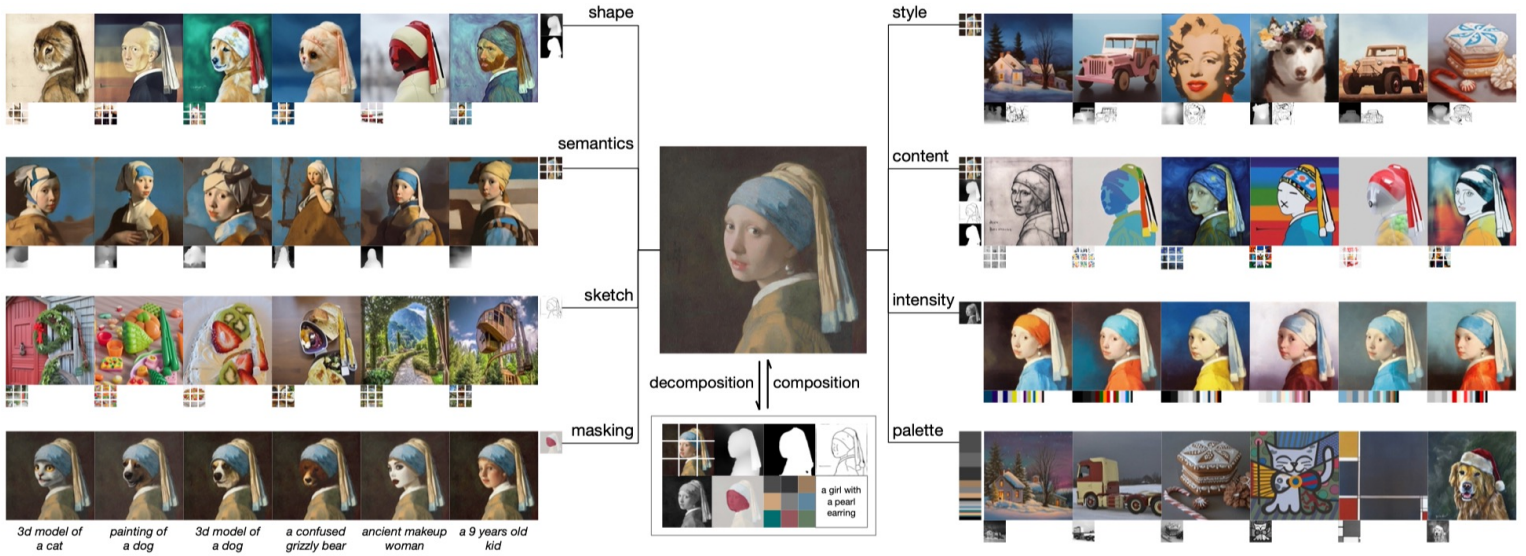
- A group of **convolution modules** to serve as the adapter for UniControl to capture features of various **low-level visual conditions**.

☐ Task-Aware HyperNet

- The HyperNet modulates the **zero-convolution** modules of ControlNet with the **task instruction condition** c_{task}:

MOE Hybrid Tasks Generalization　　MOE Zero-Shot New Task Generalization

➢ 2-3. 图像的可控生成：基于组合条件的可控图像生成（Composer）



“通过有限的表达来定义无限的变化”

“The infinite use of finite means.”

– Noam Chomsky (Chomsky, 1965)

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0,\mathbf{c},\boldsymbol{\epsilon},t}(\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(a_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}, \mathbf{c})\|_2^2),$$

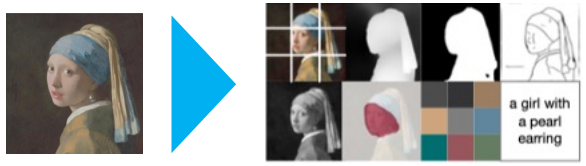
$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = \omega\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}) + (1-\omega)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$$

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, \mathbf{c}) = \omega\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_2) + (1-\omega)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}_1)$$

多条件组合控制

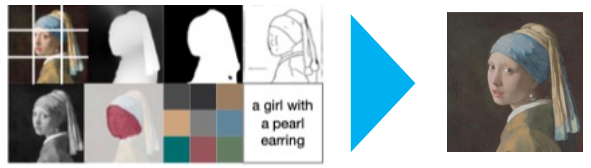
◘ 关键技术：

① Decomposition



Caption
Semantics and style
Color
Sketch
Instances
Depthmap
Intensity
Masking

② Recomposition



【1】 Composer: Creative and Controllable Image Synthesis with Composable Conditions, Alibaba Group, Arxiv, 2023 (Citations 21)

➢ 3-1. 可控文生图：主题驱动的可控文生图微调 （DreamBooth）

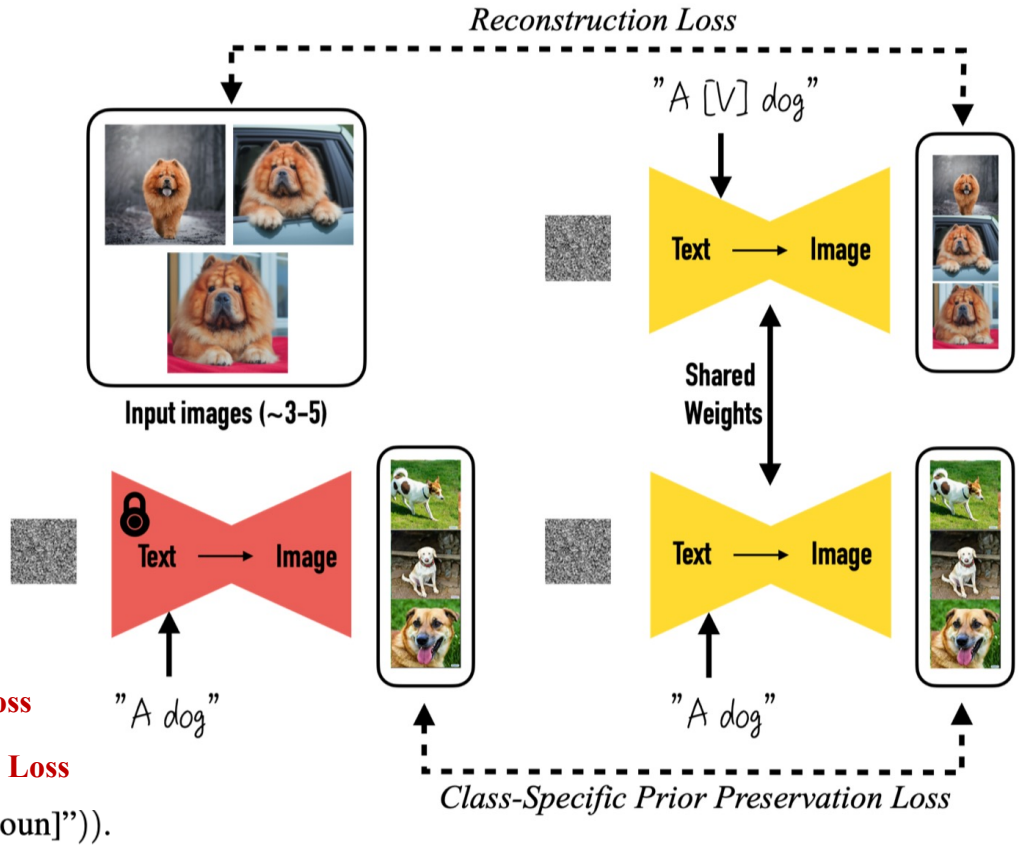
关键：实现主题保持（绑定）的个性化文生图训练

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2\right]$$

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+ \rightarrow \text{Subject-Driven Guidance Loss}$$

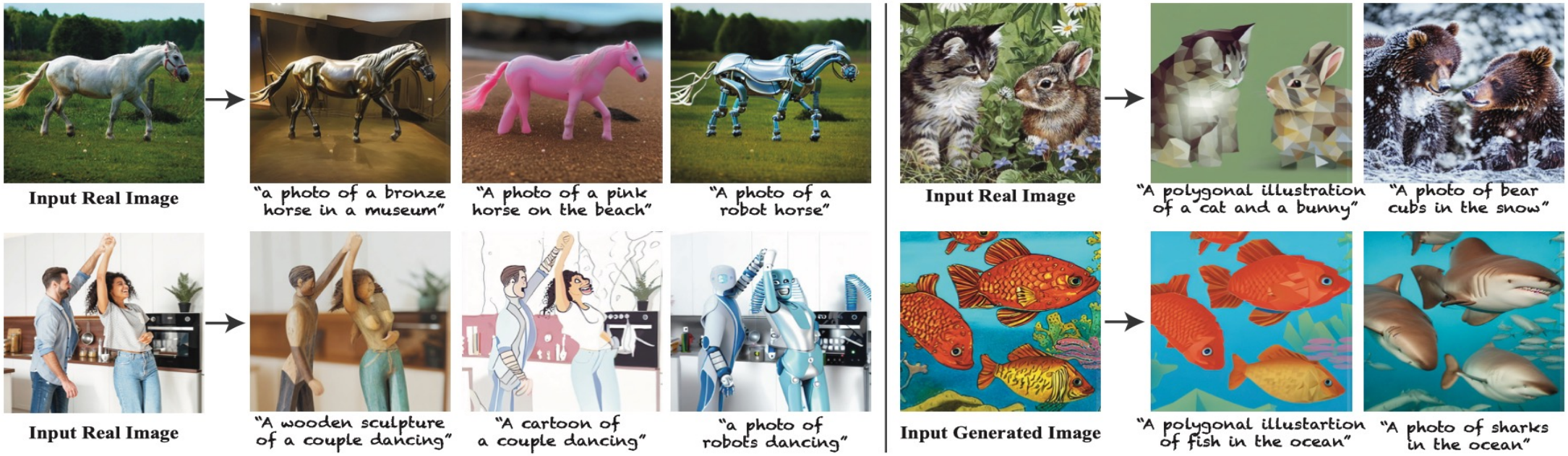
$$\lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2] \rightarrow \text{Prior-Preservation Loss}$$

$$\mathbf{c}_{\text{pr}} := \Gamma(f(\text{"a [class noun]"})).$$



【1】 DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Google Research, CVPR, 2023 (Citations 287)

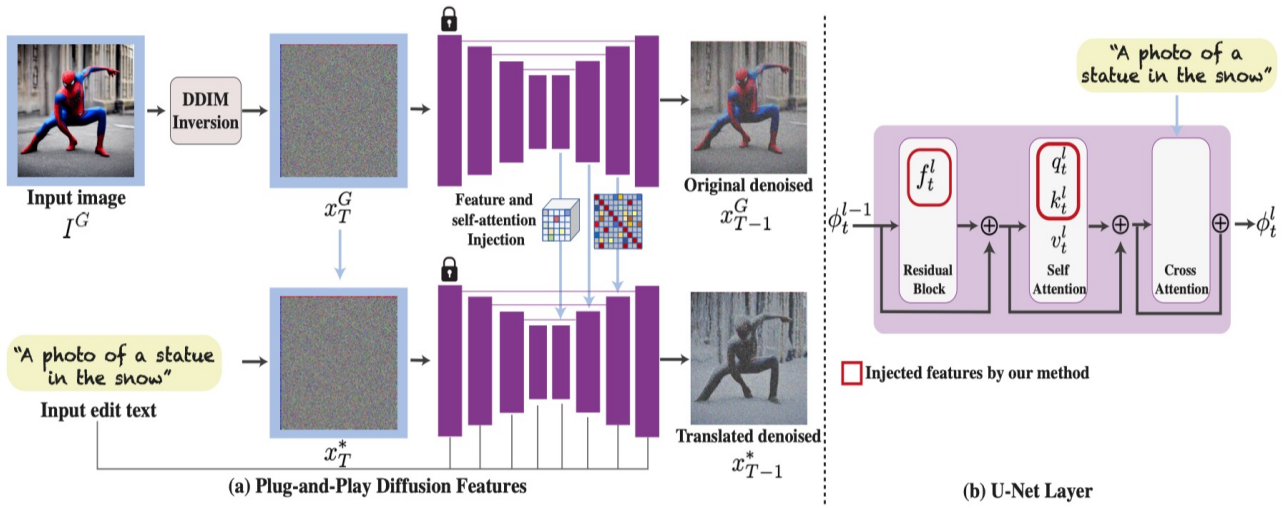
➢ 3-2. 可控文生图：即插即玩的主题绑定和风格迁移的图像生成（Plug-and-Play Diffusion）

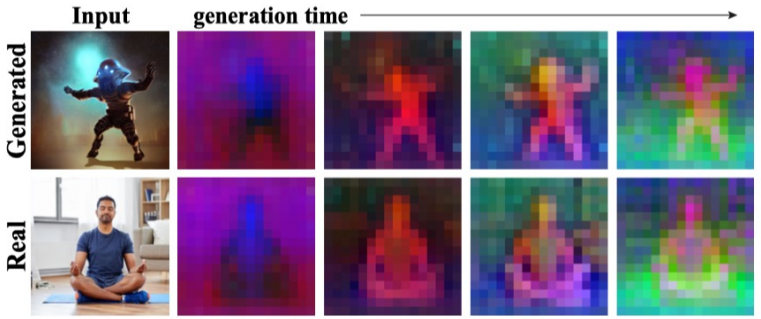


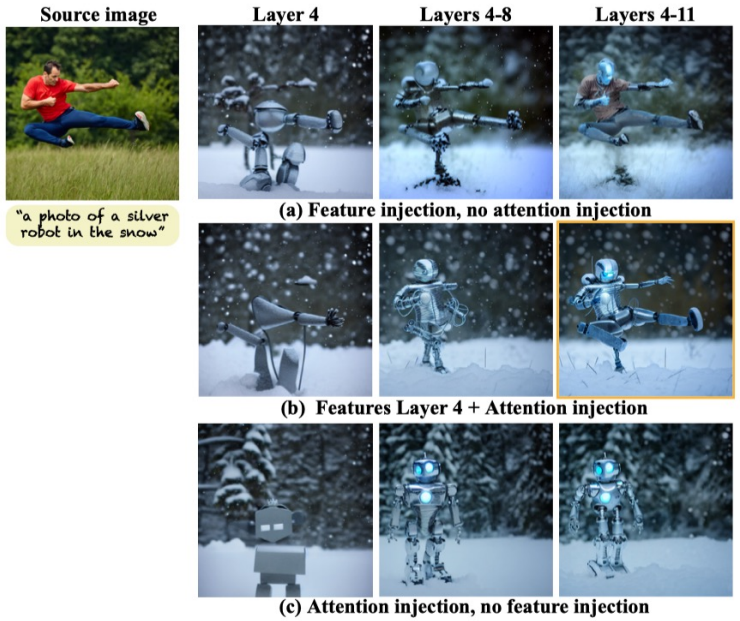
- 关键是如何实现主题绑定？
  - 结构不变
  - 布局不变
- 如何在绑定主题的基础上实现风格迁移？
  - 文本引导
  - 风格切换

【1】 Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Weizmann, CVPR, 2023 (Citations 45)

➢ 3-2. 可控文生图：即插即玩的主题绑定和风格迁移的图像生成（Plug-and-Play Diffusion）







- 关键技术：特征注入

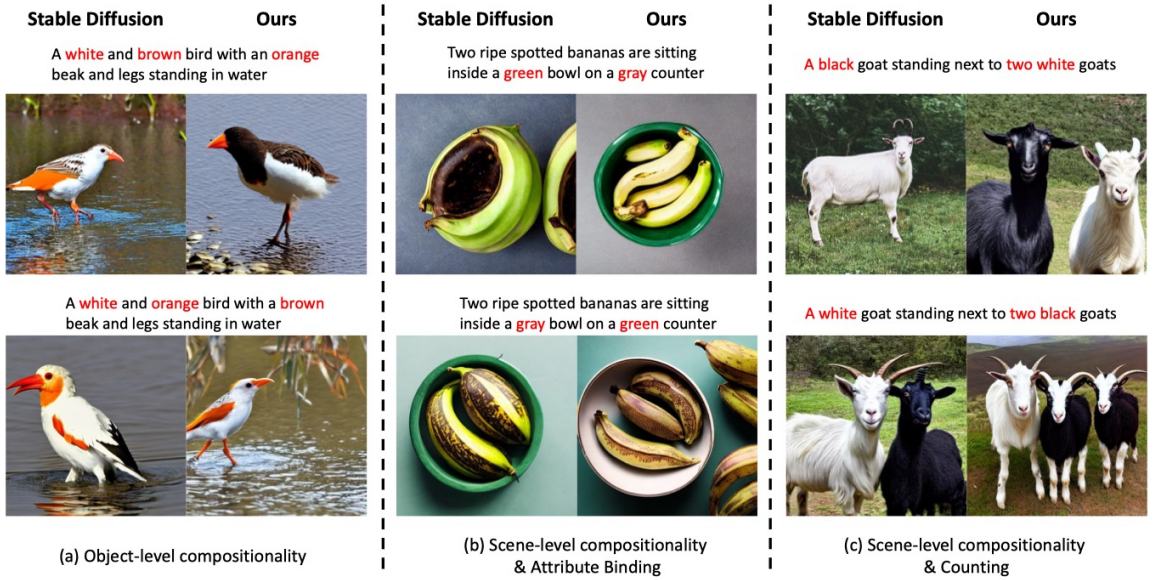
Negative-prompting

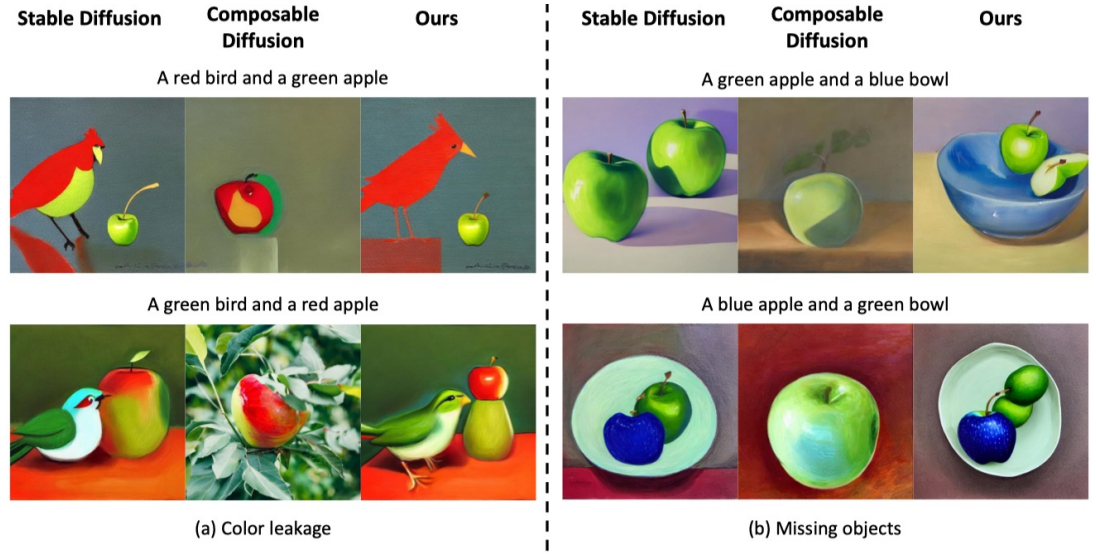
$$z_{t-1}^* = \hat{\epsilon}_\theta(x_t^*, P, t\,;\,\{f_t^l\}) \qquad \tilde{\epsilon} = \alpha\epsilon_\theta(x_t, \varnothing, t) + (1-\alpha)\epsilon_\theta(x_t, P_n, t)$$

$\hat{\epsilon}_\theta(\cdot\,;\,\{f_t^l\})$: denotes the modified denoising step with the injected features $\{f_t^l\}$.

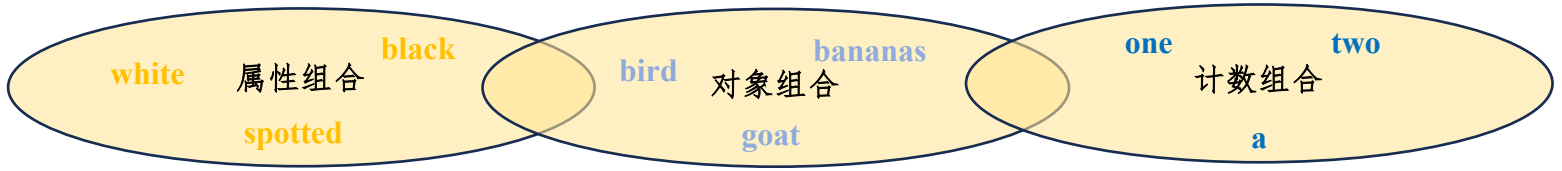
【1】 Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Weizmann, CVPR, 2023 (Citations 45)

➢ 3-3. 可控文生图：基于属性绑定和多对象组合的结构化图像生成（StructureDiffusion）





- 在复杂的文生图场景中，扩散模型应该具备怎样的能力？

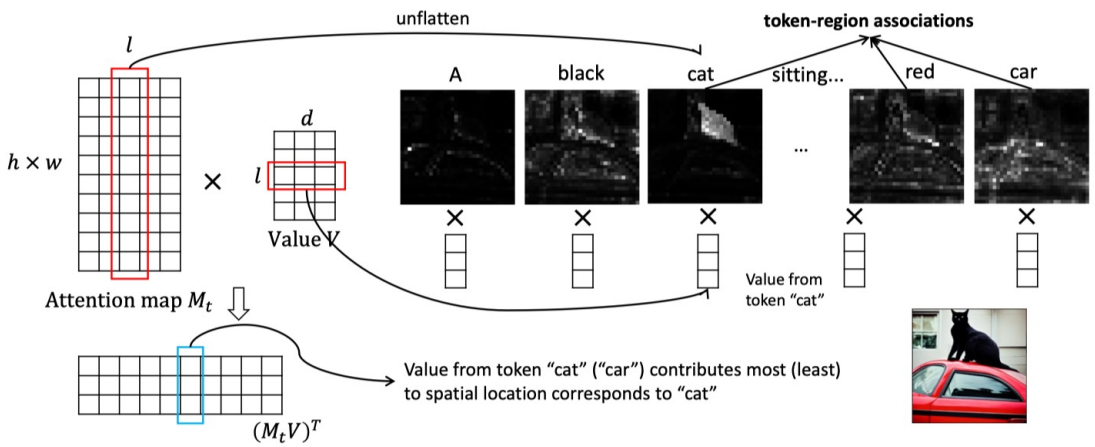


推理（逻辑、因果、方位……）

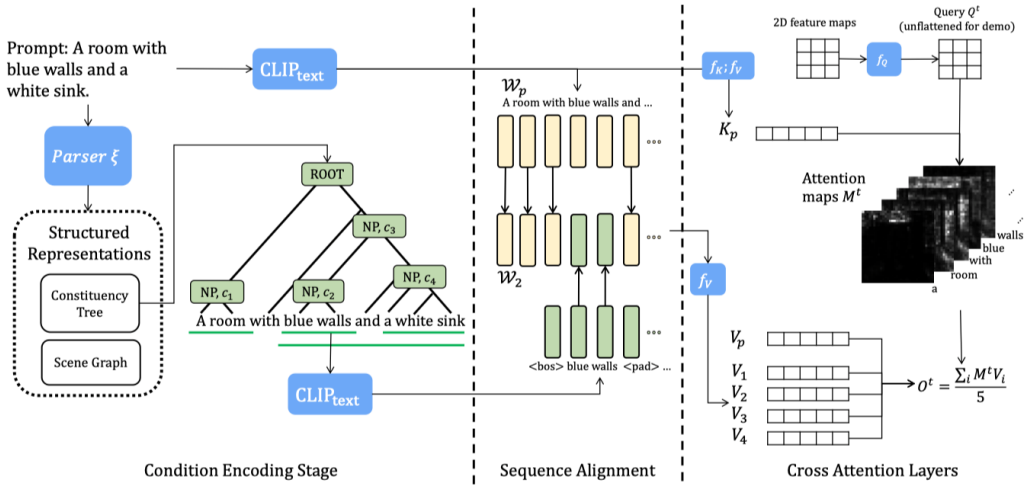
知识（常识、领域知识……）

【1】 Training-free structured diffusion guidance for compositional text-to-image synthesis, UC, ICLR, 2023 (Citations 47)

➢ 3-3. 可控文生图：基于属性绑定和多对象组合的结构化图像生成（StructureDiffusion）



cross-attention operations



StructureDiffusion

□ 核心点：

Algorithm 1 StructureDiffusion Guidance.

Require:
Input: Prompt $\mathcal{P}$, Parser ξ, decoder ψ, trained diffusion model ϕ.
Output: Generated image x.

Retrieve concept set $\mathcal{C} = [c_1, \ldots, c_k]$ by traversing $\xi(\mathcal{P})$;
$\mathcal{W}_\mathrm{p} \leftarrow \mathrm{CLIP}_\mathrm{text}(\mathcal{P})$, $\mathcal{W}_i \leftarrow \mathrm{CLIP}_\mathrm{text}(c_i)$;
for $t = T, T-1, \ldots, 1$ **do**
  for each cross attention layer in ϕ **do**
    Obtain previous layer's output $\mathcal{X}^t$.
    $Q^t \leftarrow f_Q(\mathcal{X}^t)$, $K_\mathrm{p} \leftarrow f_K(\mathcal{W}_\mathrm{p})$, $V_i \leftarrow f_V(\overline{\mathcal{W}}_i)$;
    Obtain attention maps M^t from Q^t, K_p;
    Obtain O^t from $M^t, \{V_i\}$, and feed to following layers;
  end for
end for
Feed z^0 to decoder $\psi(\cdot)$ to generate x.

$$\mathbb{W} = [\mathcal{W}_\mathrm{p}, \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_k],\ \mathcal{W}_i = \mathrm{CLIP}_\mathrm{text}(c_i),\ i = 1, \ldots k.$$

CLIP Text Encoding

$$\mathbb{V} = [f_V(\mathcal{W}_\mathrm{p}), f_V(\overline{\mathcal{W}}_1), \ldots, f_V(\overline{\mathcal{W}}_k)] = [V_\mathrm{p}, V_1, \ldots, V_k].$$

Calculate Value Vector of Objects !

$$\mathbb{K} = \{f_K(\mathcal{W}_i)\},\ \mathbb{M}^t = \{f_M(Q^t, K_i)\},\ i = \mathrm{p}, 1, 2, \ldots, k.$$

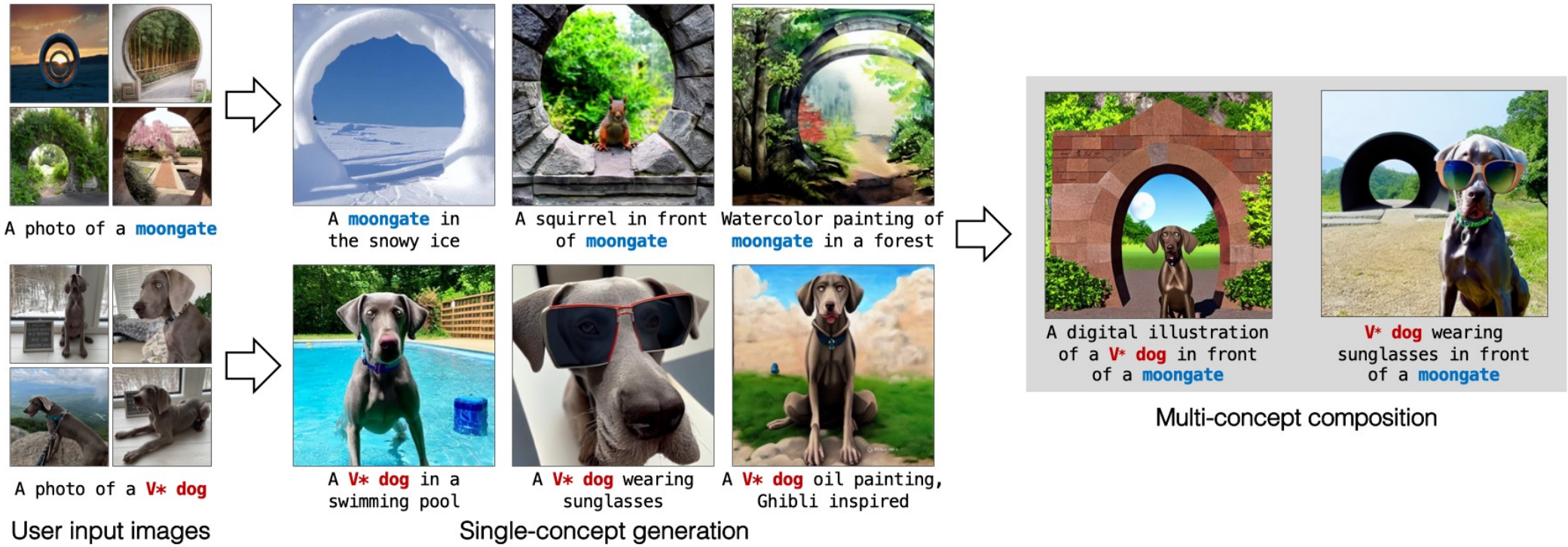
General Key Vectors and Cross-attention Map

$$O^t = \frac{1}{(k+1)} \sum_i (M_i^t V_k), i = \mathrm{p}, 1, 2, \ldots, k.$$

Compositional Output

【1】 Training-free structured diffusion guidance for compositional text-to-image synthesis, UC, ICLR, 2023 (Citations 47)

➢ 3-4. 可控图像生成：基于多概念组合的定制化文生图（Custom Diffusion）



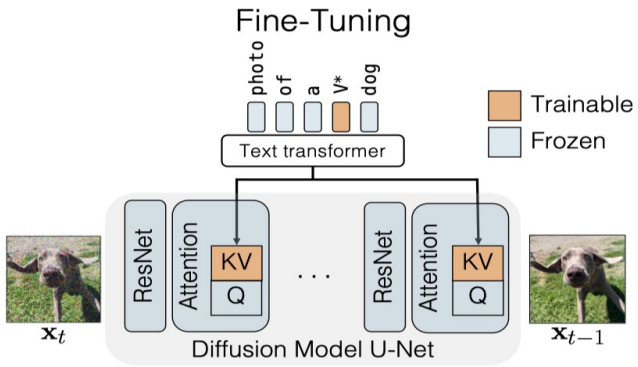
如何在训练的同时尽量不影响模型的先验？

◘ 关键洞察：

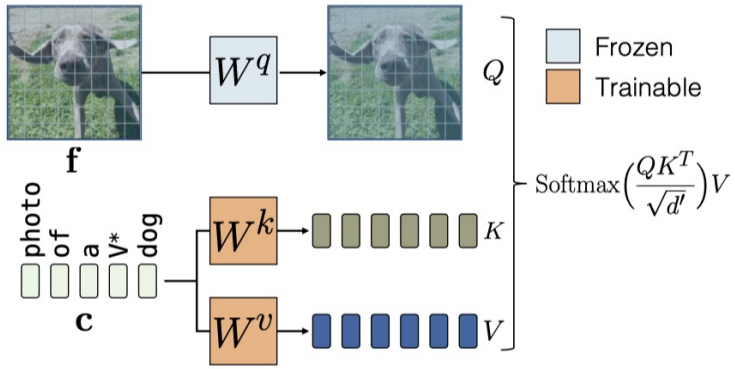
① 如何从用户给定的例子中学习并保存个性化（新）概念？ ② 如何组合多个新概念并生成文本定制的图像？

【1】 Multi-Concept Customization of Text-to-Image Diffusion, CMU, CVPR, 2023 (Citations 57)

➢ 3-4. 可控图像生成：基于多概念组合的定制化文生图（Custom Diffusion）



Custom Diffusion.



Single-head Cross-Attention.

Case Studies.

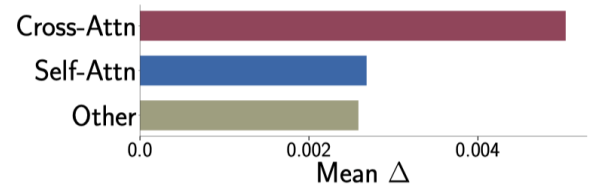
- 核心点：

$$p_\theta(\mathbf{x}_0) = \int \left[p_\theta(\mathbf{x}_T) \prod p_\theta^t(\mathbf{x}_{t-1}|\mathbf{x}_t) \right] d\mathbf{x}_{1:T},$$

$$\mathbb{E}_{\epsilon,\mathbf{x},\mathbf{c},t}[w_t||\epsilon - \epsilon_\theta(\mathbf{x}_t,\mathbf{c},t)||]$$

问题：计算不高效、容易过拟合

解决方案：在模型参数的小规模子集上进行微调



$$\hat{W} = \arg\min_W ||WC_{\text{reg}}^\top - W_0 C_{\text{reg}}^\top||_F$$

$$\text{s.t. } WC^\top = V, \text{ where } C = [\mathbf{c}_1 \cdots \mathbf{c}_N]^\top$$

$$\text{and } V = [W_1\mathbf{c}_1^\top \cdots W_N\mathbf{c}_N^\top]^\top.$$

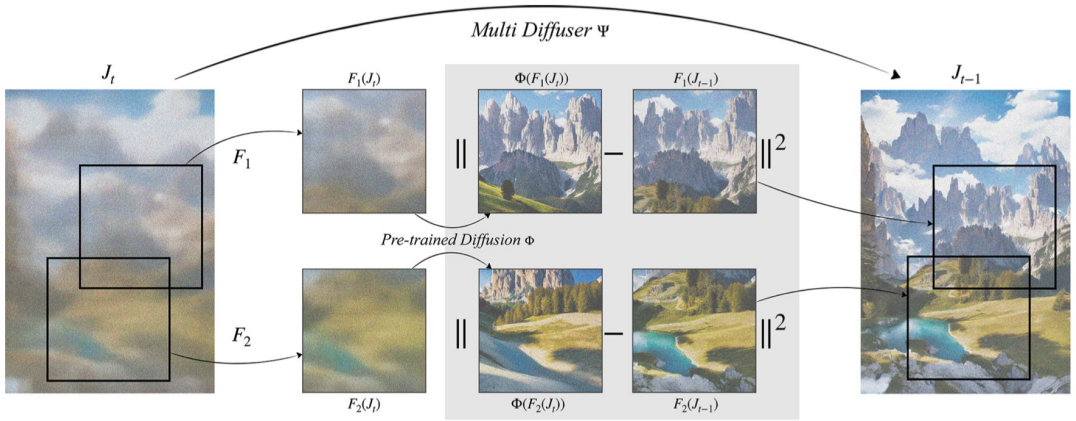
【1】 Multi-Concept Customization of Text-to-Image Diffusion, CMU, CVPR, 2023 (Citations 57)

➢ 3-5. 可控图像生成：统一多个扩散过程的多目标可控生成（MultiDiffusion）

- **优点：**（1）无需优化模型；（2）无需修改文本prompt

$$\Phi : \mathcal{I} \times \mathcal{Y} \to \mathcal{I} \qquad I_T, I_{T-1}, \ldots, I_0 \quad \text{s.t.} \quad I_{t-1} = \Phi(I_t|y)$$

$$\Psi : \mathcal{J} \times \mathcal{Z} \to \mathcal{J} \qquad J_T, J_{T-1}, \ldots, J_0 \quad \text{s.t.} \quad J_{t-1} = \Psi(J_t|z)$$



$$I_t^i = F_i(J_t), \quad y_i = \lambda_i(z)$$

$$\Psi(J_t|z) = \arg\min_{J \in \mathcal{J}} \mathcal{L}_{\text{FTD}}(J|J_t, z)$$

$$\mathcal{L}_{\text{FTD}}(J|J_t, z) = \sum_{i=1}^{n} \left\| W_i \otimes \left[F_i(J) - \Phi(I_t^i|y_i) \right] \right\|^2$$

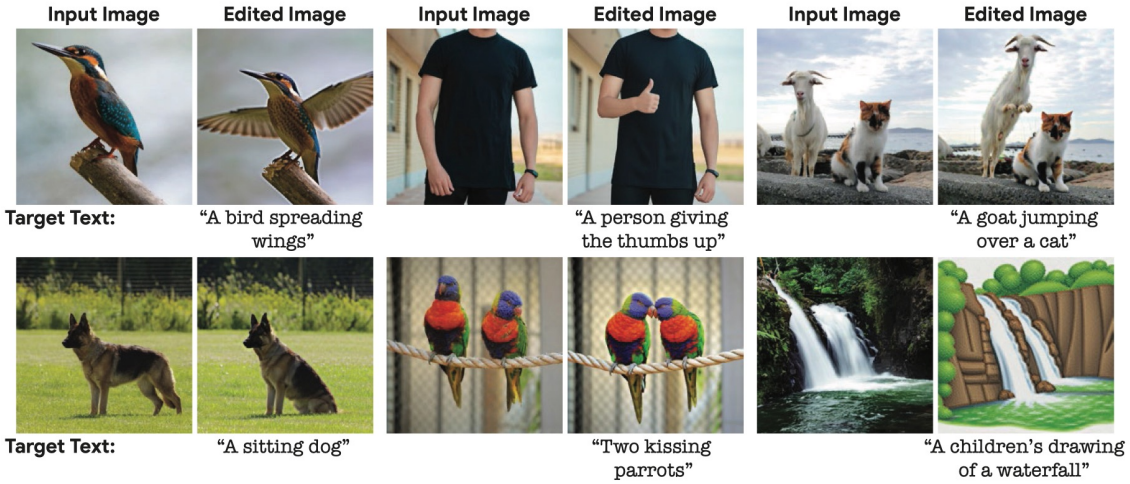
Algorithm 1 MultiDiffusion sampling.

Input : Φ ▷ pre-trained Diffusion Model
$\{F_i\}_{i=1}^n$ ▷ image space mappings
$\{y_i\}_{i=1}^n$ ▷ text-prompts conditioning
$\{W_i\}_{i=1}^n$ ▷ per-pixel weights
$J_T \sim P_{\mathcal{J}}$ ▷ noise initialization
for $t = T, \ldots, 1$ **do**
  $I_{t-1}^i \leftarrow \Phi(F_i(J_t), y_i) \quad \forall i \in [n]$ ▷ diffusion updates
  $J_{t-1} \leftarrow \texttt{MultiDiffuser}(\{I_{t-1}^i\}_{i=1}^n)$ ▷ Eq. 5
Output : J_0

【1】 MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, WIoS & Meta AI, ICML, 2023 (Citations 16)

➢ 4-1. 可控图像编辑：基于文本指令的真实图像编辑（Imagic）

关键：实现文本语义引导的真实场景图像编辑而保持原始图像特征不变

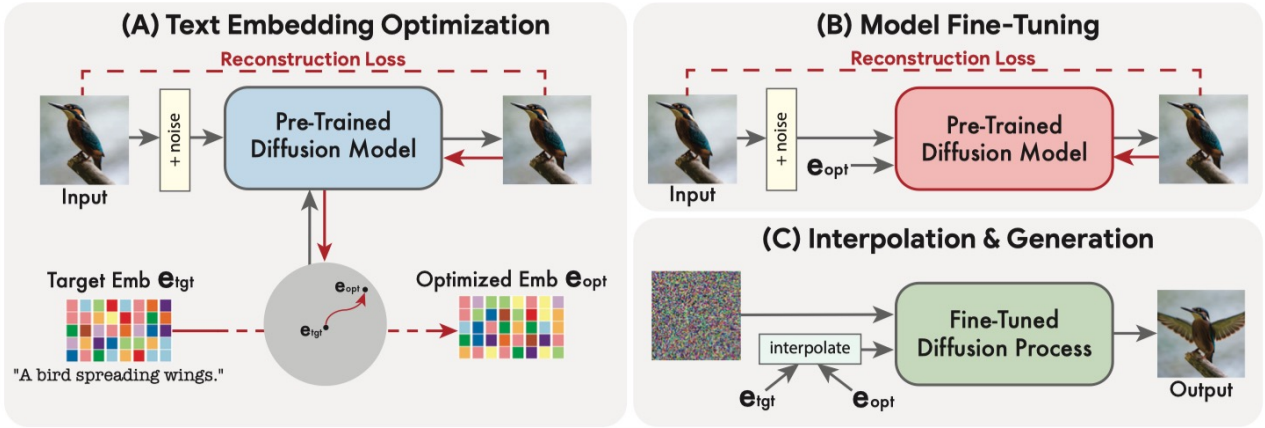


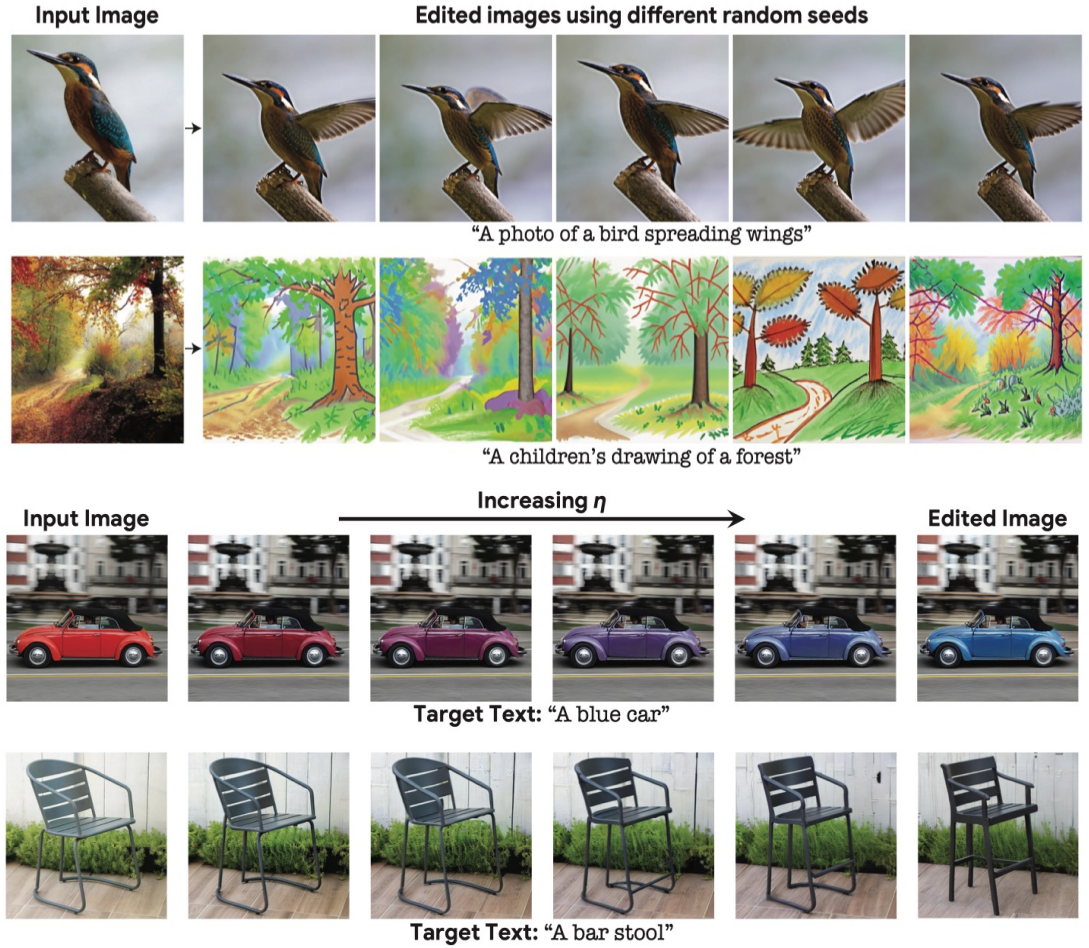
- 如何实现文本语义引导的非刚性图像编辑？
- 如何保持原始图像特征在细节上不改变？

【1】 Imagic: Text-Based Real Image Editing with Diffusion Models, Google, CVPR, 2023 (Citations 162)

➢ 4-1. 可控图像编辑：基于文本指令的真实图像编辑（Imagic）

关键：实现文本语义引导的真实场景图像编辑而保持原始图像特征不变





- **Text embedding optimization**

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - f_\theta(\mathbf{x}_t, t, \mathbf{e})\|_2^2\right]$$

- **Model fine-tuning**

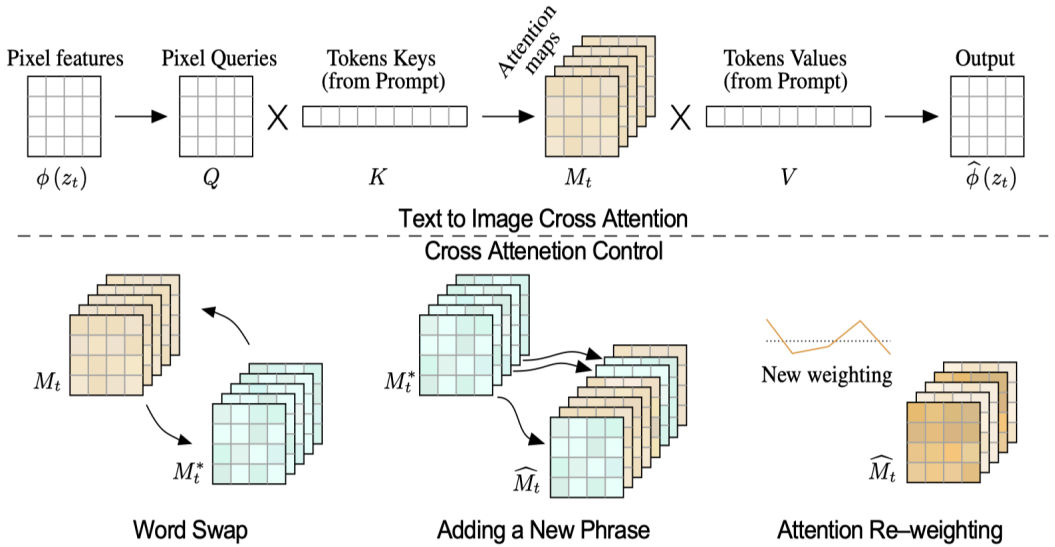
- **Text embedding interpolation**

$$\bar{\mathbf{e}} = \eta \cdot \mathbf{e}_{tgt} + (1-\eta) \cdot \mathbf{e}_{opt},$$

【1】 Imagic: Text-Based Real Image Editing with Diffusion Models, Google, CVPR, 2023 (Citations 162)

➢ 4-2. 可控图像编辑：基于Cross-Attention的精准图像编辑 (Prompt-to-Prompt)

关键：实现文本语义引导的真实场景图像编辑而保持原始图像特征不变



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

□ Cross-Attention Map $\qquad (Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} \\ (M_t)_{i,j} \end{cases} \quad \begin{matrix} \text{if } j = j^* \\ \text{otherwise.} \end{matrix}$

$$M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \quad \hat{\phi}(z_t) = MV$$

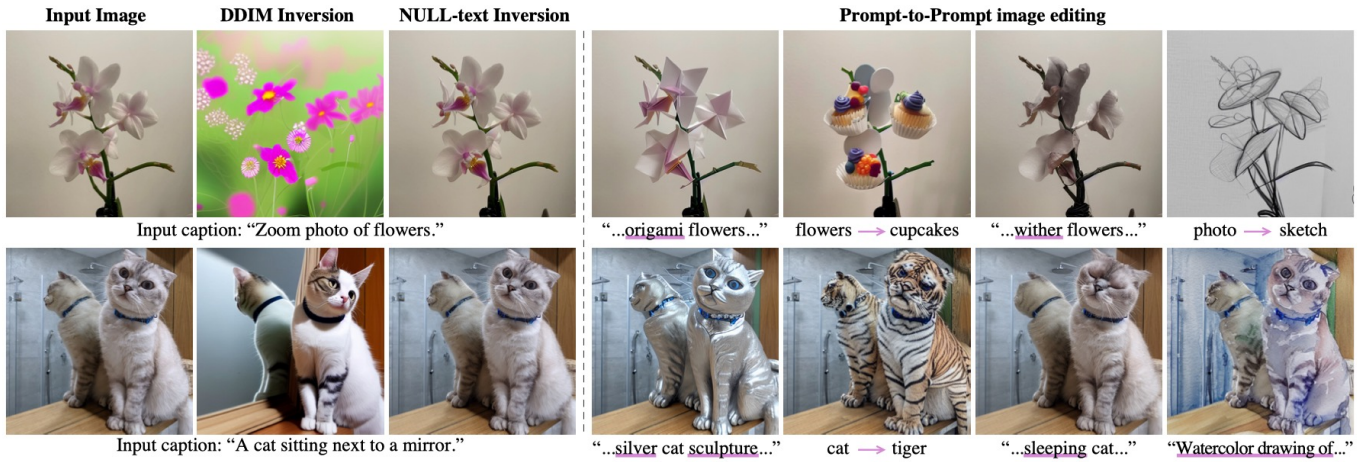
◆ **Object Replace** → Word Swap

◆ **Global Change** → Adding a New Phrase

◆ **Degree Reweight** → Attention Re-weighting

【1】 **Prompt-to-Prompt Image Editing with Cross Attention Control, Google, CVPR, 2023 (Citations 204)**

➢ 4-3. 可控图像编辑：基于空文本反演的真实图像编辑 （Null-text Inversion）



- 回顾DDIM反演过程，思考有没有一些问题？

Stable Diffusion (LDM)

$$\min_{\theta} E_{z_0,\varepsilon\sim N(0,I),t\sim\text{Uniform}(1,T)} \|\varepsilon - \varepsilon_\theta(z_t,t,\mathcal{C})\|_2^2 \quad \textbf{DDPM (Conditional)}$$

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \sqrt{\alpha_{t-1}} \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t) \quad \textbf{DDIM Sampling}$$

$$z_0 = E(x_0)$$

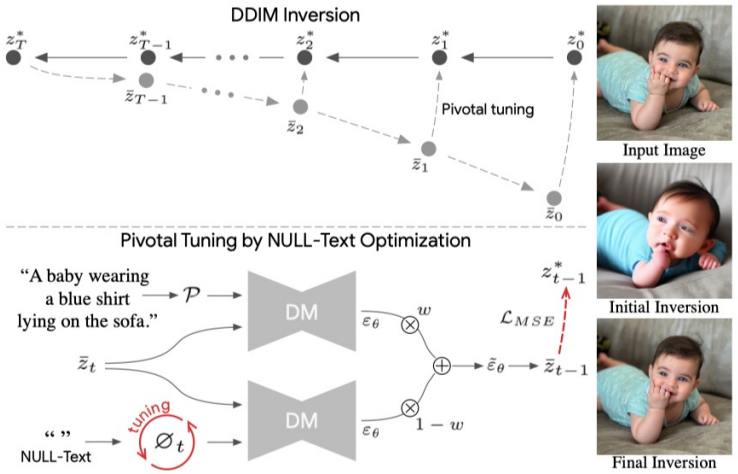
$$x_0 = D(z_0).$$

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

- 关键洞察：对Diffusion的预测可能会逐渐偏离，引入空文本（提示参数）进行微调优化！

【1】 Null-text Inversion for Editing Real Images using Guided Diffusion Models, Google, CVPR, 2023 (Citations 81)

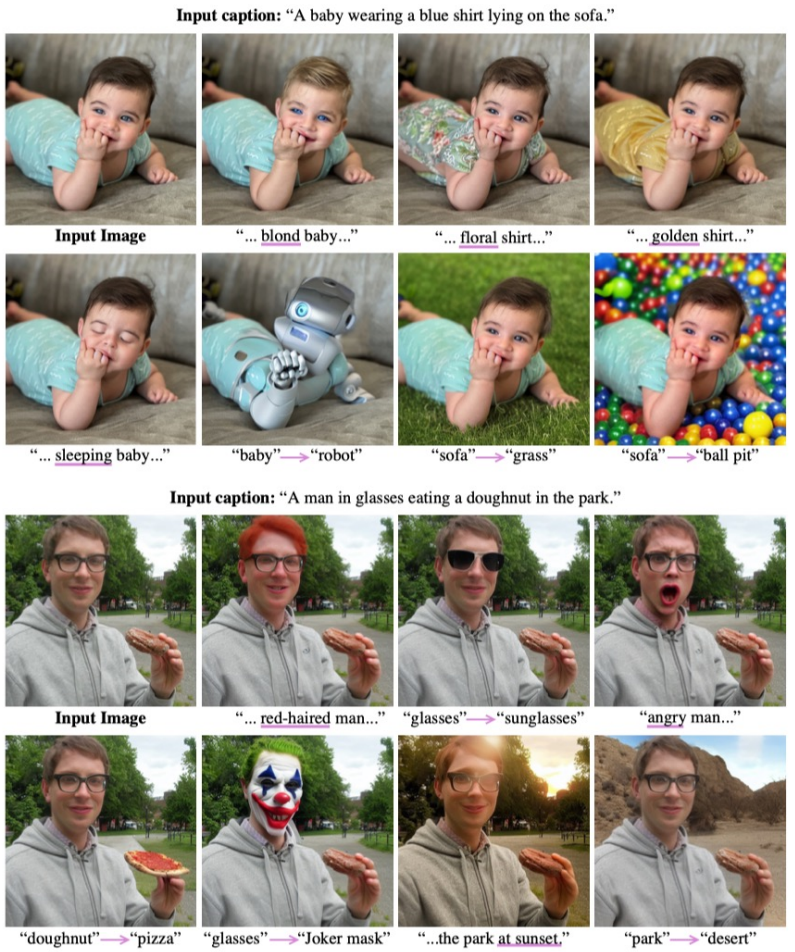
➢ 4-3. 可控图像编辑：基于空文本反演的真实图像编辑（Null-text Inversion）



Algorithm 1: Null-text inversion

Input: A source prompt embedding $\mathcal{C} = \psi(\mathcal{P})$ and input image $\mathcal{I}$.
Output: Noise vector z_T and optimized embeddings $\{\varnothing_t\}_{t=1}^{T}$.

Set guidance scale $w = 1$;
Compute the intermediate results $z_T^*, \ldots, z_0^*$ using DDIM inversion over $\mathcal{I}$;
Set guidance scale $w = 7.5$;
Initialize $\bar{z_T} \leftarrow z_T^*$, $\varnothing_T \leftarrow \psi(\text{""})$;
for $t = T, T-1, \ldots, 1$ **do**
  for $j = 0, \ldots, N-1$ **do**
    $\varnothing_t \leftarrow \varnothing_t - \eta \nabla_{\varnothing} \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \varnothing_t, \mathcal{C}) \right\|_2^2$;
  end
  Set $\bar{z}_{t-1} \leftarrow z_{t-1}(\bar{z}_t, \varnothing_t, \mathcal{C})$, $\varnothing_{t-1} \leftarrow \varnothing_t$;
end
Return $\bar{z_T}, \{\varnothing_t\}_{t=1}^{T}$



- 回顾DDIM反演过程，思考有没有一些问题？

$$\min_{\theta} E_{z_0,\varepsilon\sim N(0,I),t\sim\text{Uniform}(1,T)} \left\| \varepsilon - \varepsilon_\theta(z_t, t, \mathcal{C}) \right\|_2^2$$

DDPM (Conditional)

Stable Diffusion (LDM)

$$z_0 = E(x_0)$$

$$x_0 = D(z_0).$$

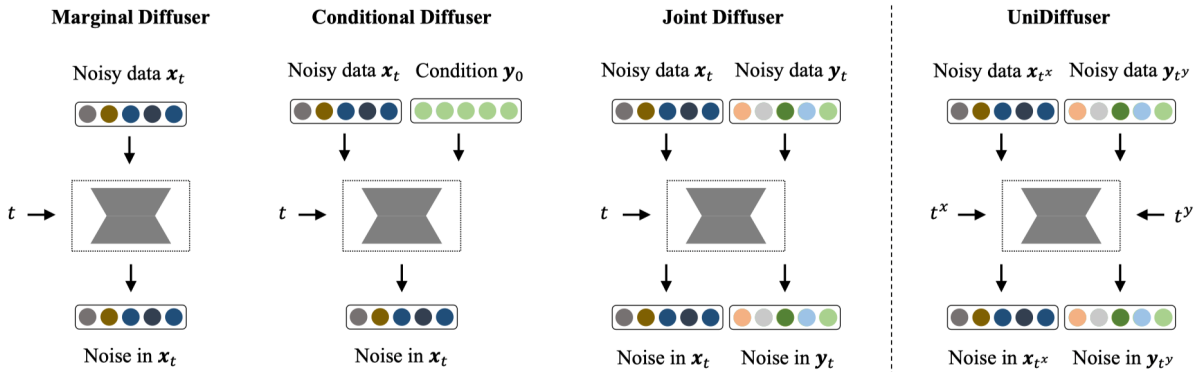
$$\min_{\varnothing_t} \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \varnothing_t, \mathcal{C}) \right\|_2^2.$$

Null-Text Optimization

- 优点：（1）无需优化模型；（2）无需修改文本prompt

【1】 Null-text Inversion for Editing Real Images using Guided Diffusion Models, Google, CVPR, 2023 (Citations 81)

➢ 5-1. 统一图文生成： 图文统一扩散框架 (UniDiffuser)



$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}),$$ Forward Process （无需训练）

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}),$$ Reverse Predication （DDPM训练）

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}^x} \|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|_2^2,$$ Unconditional Loss

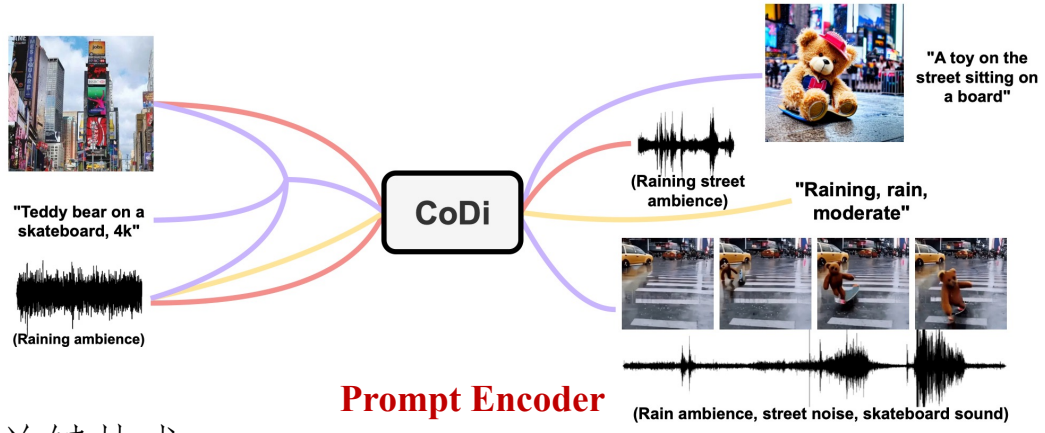
$$\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{\epsilon}^x} \|\boldsymbol{\epsilon}^x - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_0, t)\|_2^2.$$ Conditional Loss

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{y}_0,\boldsymbol{\epsilon}^x,\boldsymbol{\epsilon}^y,t^x,t^y} \|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t^x}, \boldsymbol{y}_{t^y}, t^x, t^y) - [\boldsymbol{\epsilon}^x, \boldsymbol{\epsilon}^y]\|_2^2$$ **Unified Loss !**

【1】UniDiffuser: One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale, Tsinghua & Huawei, ICLR, 2022 (Citations 12)
【2】Classifier-Free Diffusion Guidance (CFDG), Google Research, NeurIPS, 2021 (Citations 562)

➢ 5-2. 统一多模态生成：基于组合扩散的通用多模态生成（CoDi）

- 通用多模态可控生成： 任意组合的多模态输入 ＋ 任意组合的多模态输出



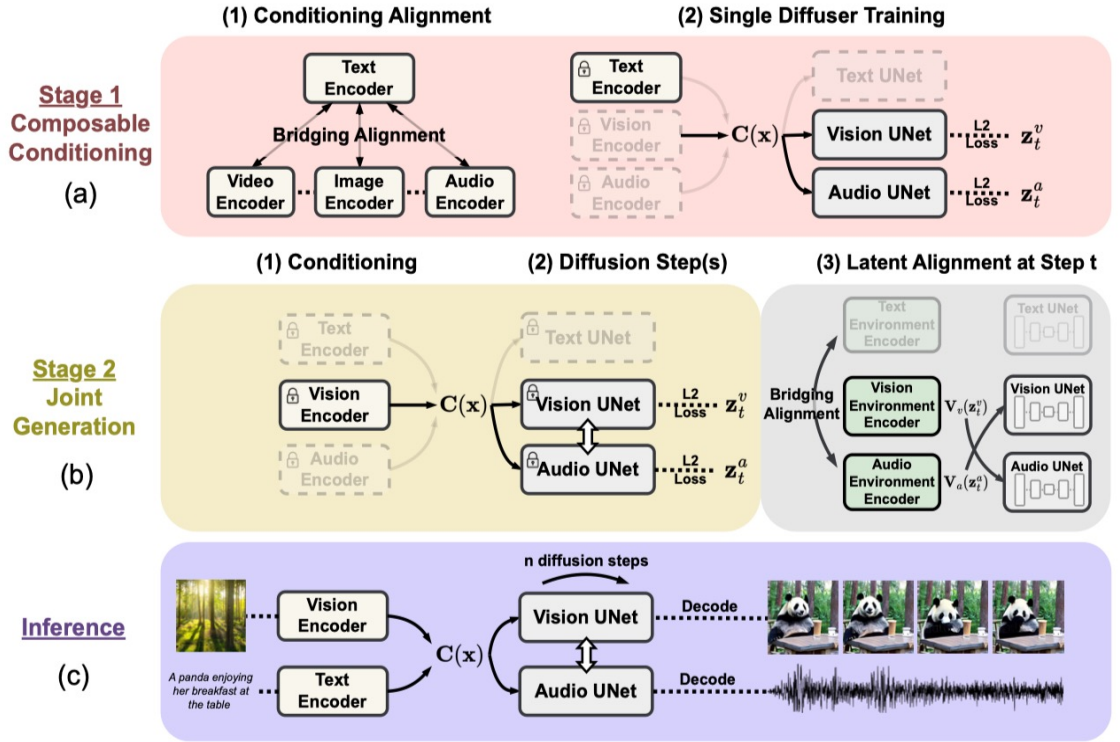
- 关键技术：

$$C(x_t, x_i, x_v, x_a) = \sum_m \alpha_m C(m)$$ 桥接对齐 ＋ 特征插值！

$$\mathcal{L}^A_{Cross} = \mathbb{E}_{\boldsymbol{z},\epsilon,t}\|\epsilon - \epsilon_{\theta_c}(\boldsymbol{z}^A_t, V_B(\boldsymbol{z}^B_t), t, C(\boldsymbol{y}))\|^2_2,$$

$$\mathcal{L}^A_{Cross} + \mathcal{L}^B_{Cross}.$$

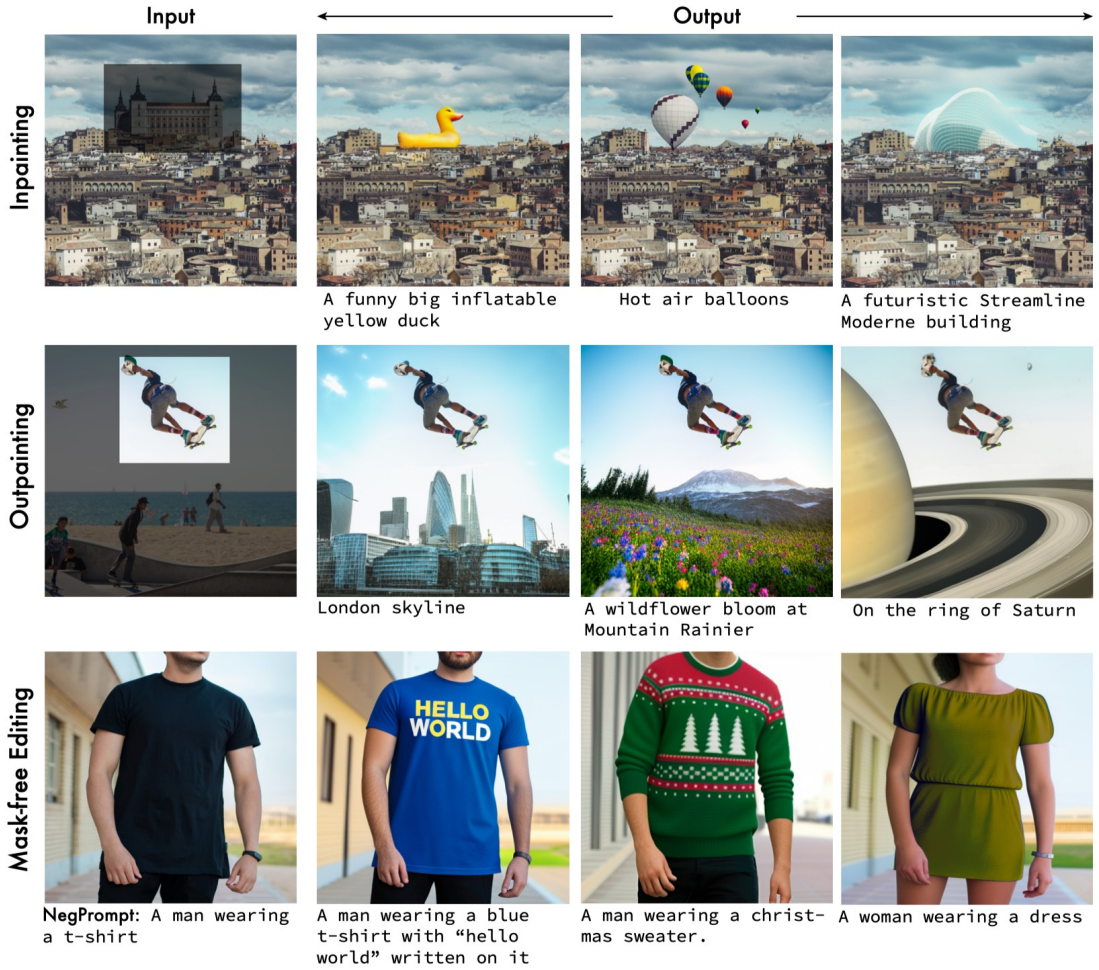
} 联合多模态生成

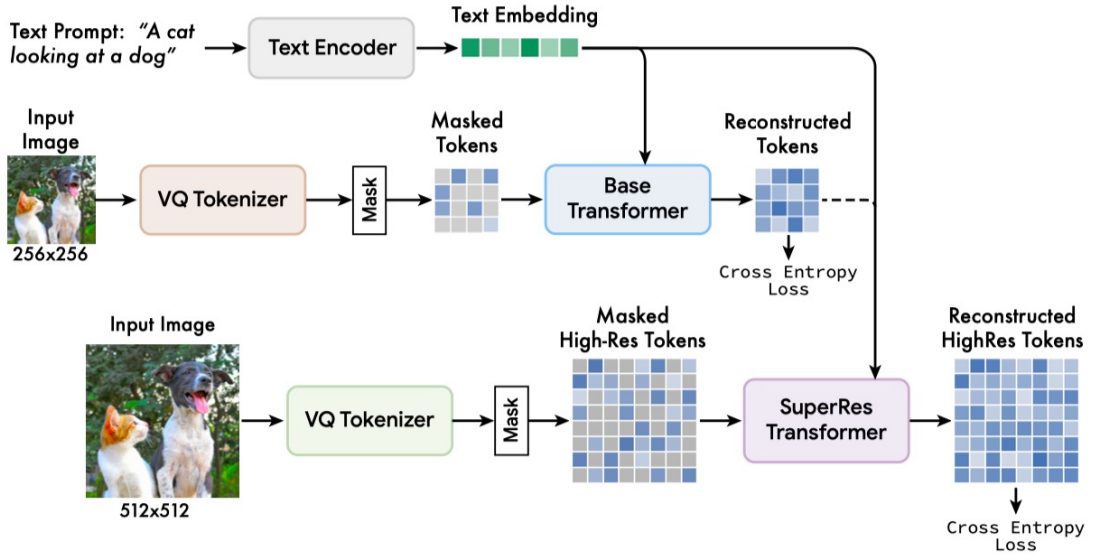


CoDi model architecture.

【1】 Any-to-Any Generation via Composable Diffusion, UNC & Microsoft, Arxiv, 2023 (Citations 1)

➢ 5-3. 统一图文生成：基于掩码生成式 Transformer 的可控文生图（Muse）





◘ 关键点：

① Pre-trained T5 Text-Encoder　丰富的语义信息

② VQGAN Visual-Tokenizer　多样且可感知的视觉元素

③ Iterative Parallel Transformer Decoder　高效的解码生成

④ Conditional Guidance with Negative Prompts　有效的语义编辑

$$\ell_g = (1+t)\ell_c - t\ell_u$$

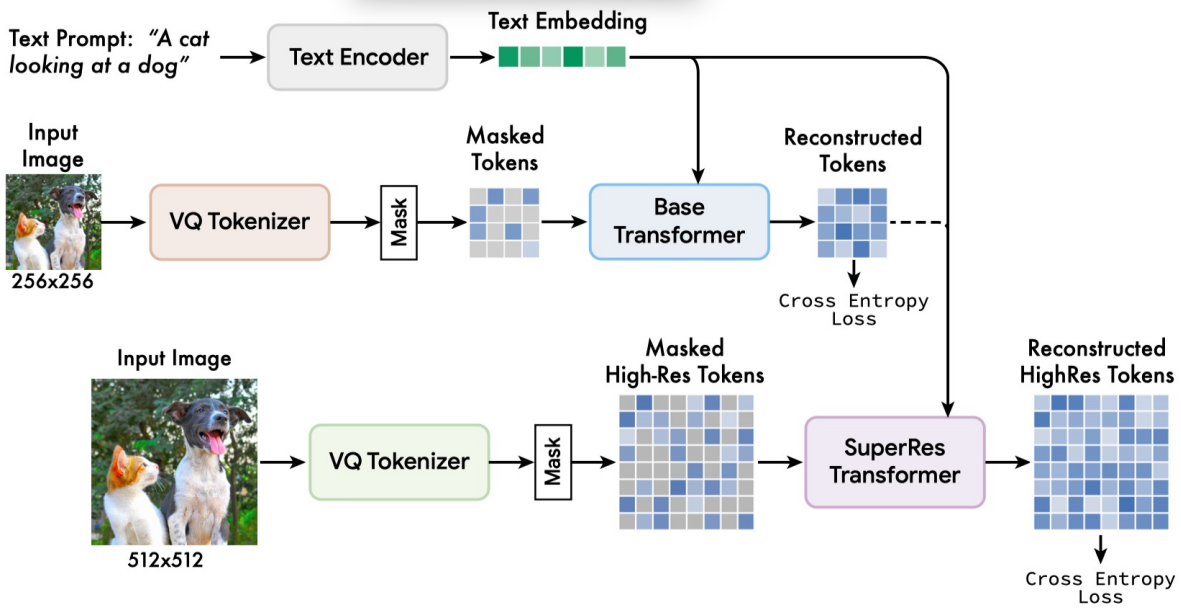
【1】 Muse: Text-To-Image Generation via Masked Generative Transformers, Google Research, Arxiv, 2023 (Citations 86)

图文编创领域的统一生成范式

非统一范式：Visual-ChatGPT～！

➢ 方案一：图像兼容文本生成（ICTG）：

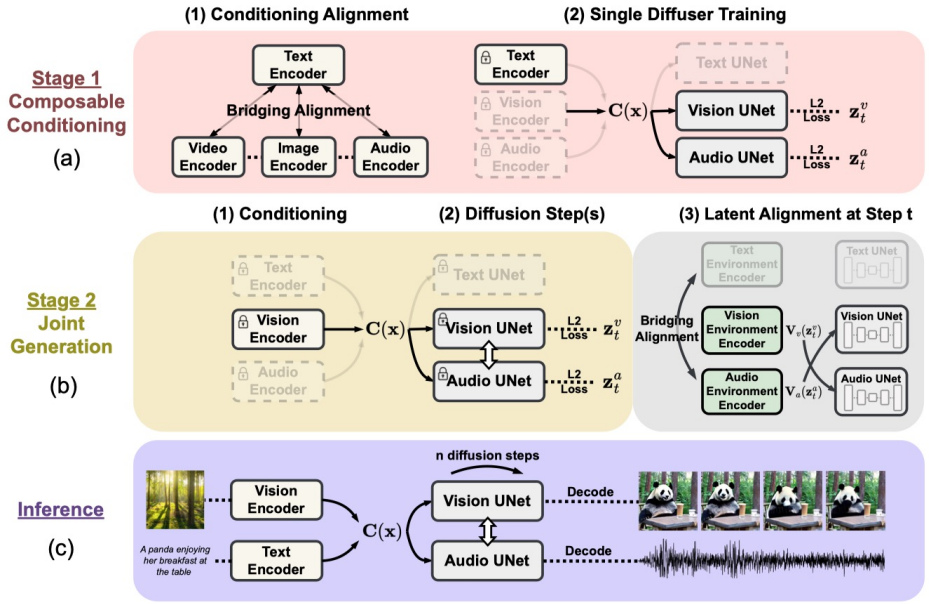
Transformer 架构



Muse, Google Research, 2023

方案二：文本兼容图像生成（TCIG）：

Diffusion 架构



CoDi, UNC& Microsoft, 2023

- 技术挑战：Visual Tokenizer（Codebook）、多样性（视觉生成）
- 技术挑战：Textual Diffuser（Decoding）、准确性（文本生成）

（修炼特技-1）非刚性编辑（Non-Rigid / Soft Editing）

工作一：面向时序和空域连续变化的自适应编辑算法（Adaptive Editing）

2023.7.12-2023.9.15

□ 动机：

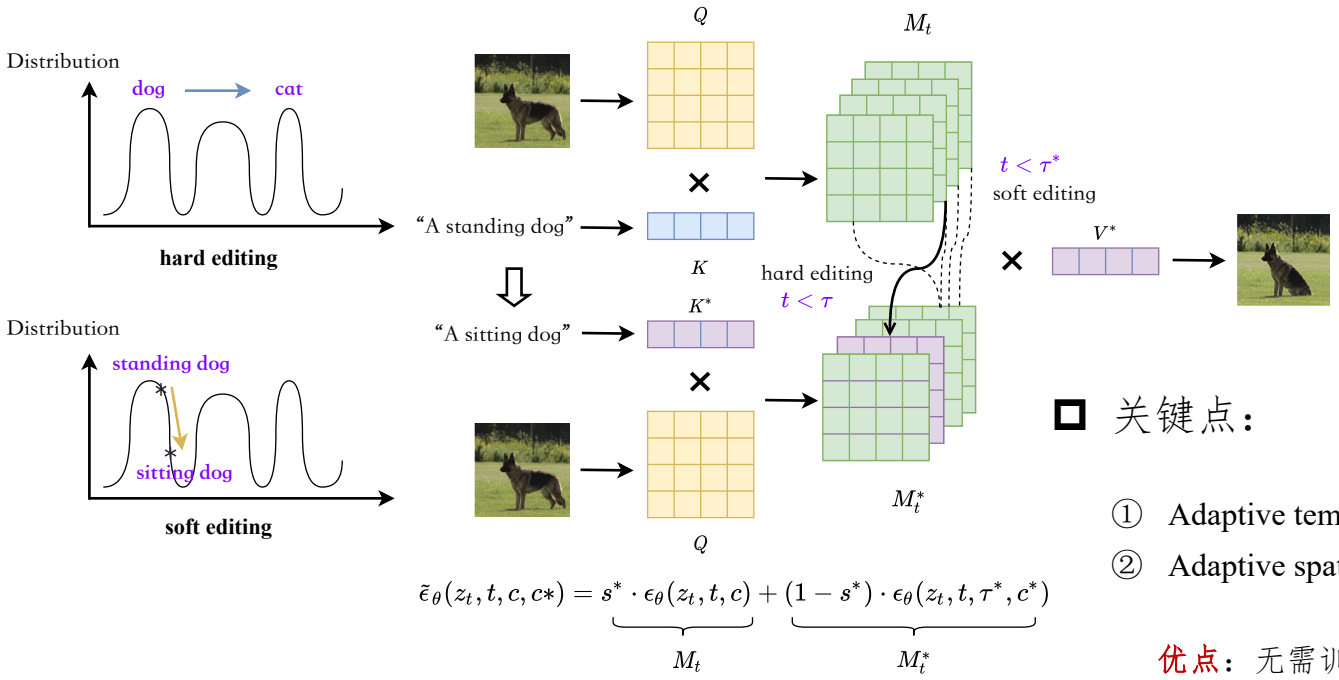
非刚性编辑具有时空连续性，依赖于自适应的条件扩散算法

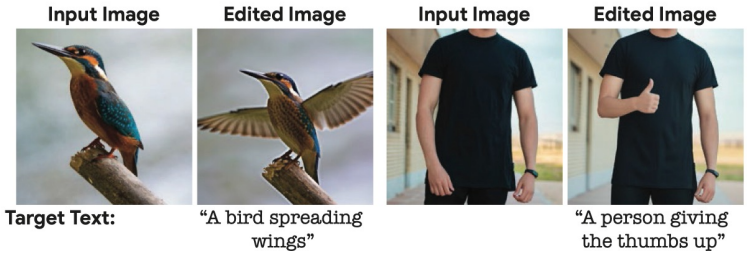
SD： Backbone → **ControlNet：** Conditional Control → **Prompt-to-Prompt：** Training-free, Hard Editing → **Adaptive Editing：** Non-Rigid / Soft Editing

● 针对性提升：

① 软编辑能力

② 语义连续性

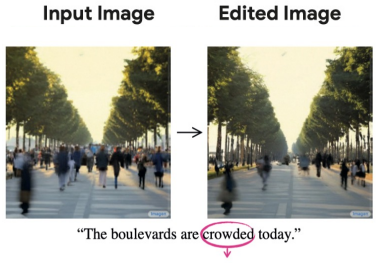




□ 关键点：

① Adaptive temporal guidance $s^* = f_{cross}(M_t, M_t^*)$

② Adaptive spatial guidance $\tau^* = f_{self}(M_t, M_t)$



优点：无需训练、提升零样本场景的软编辑能力；

（修炼特技-1）非刚性编辑（Non-Rigid / Soft Editing）

工作二：视觉知识指导的软编辑算法（Visual-Knowledge Guided Soft Editing）

2023.7.12-2023.9.15

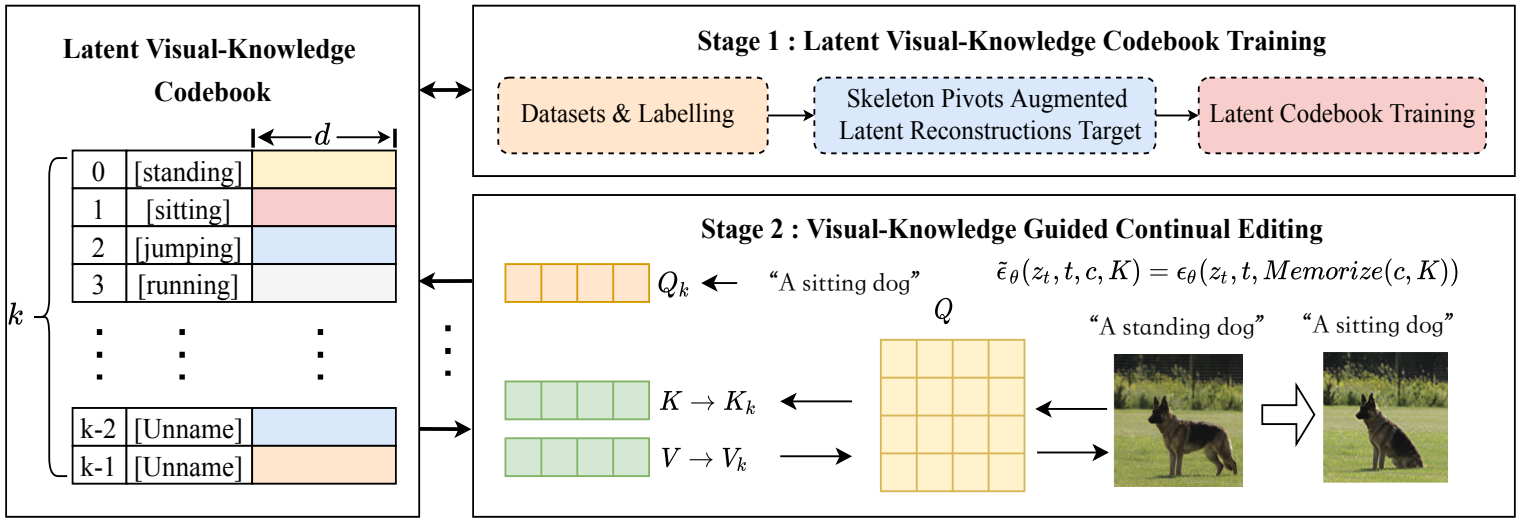
□ 动机：
非刚性编辑需要动作和姿势等视觉知识的感知、组合和生成

SD：Backbone → ControlNet：Conditional Control → VQGAN：Latent Knowledge / Codebook Sampling → VK-Editing：Action / Pose Guidance

● 针对性提升：
① 软编辑能力
② 视觉知识理解

□ 关键点： ① Latent Codebook Learning ② Visual-Knowledge Guided Continual Editing

优点：无需训练、提升零样本场景的软编辑能力；



From: https://github.com/EricGuo5513/HumanML3D

（修炼特技-2）创意场景图文生成（Creative Scene Image-Text Generation）

“ When AIGC starts to implement content customization, everything starts to really become valuable.” -- Small Pony.Ma

工作三：文本内容的视觉化生成（ Text-to-LOGO / Text-to-Wordart / Text-to-Chart ）

2023.9.16-2024.3.15

□ 动机：

文本内容的视觉化是创意场景图文生成的基础，文本像素化生成也是独辟蹊径的做法

SD：Backbone → ControlNet：Conditional Control → TextDiffuser：Latent Knowledge / Codebook Sampling → Text-to-Vision：Action / Pose Guidance

● 针对性提升：

① 文字像素级生成

② 内容生成的准确性

□ 关键点： ① Denoising Target ② Content Supervision ③ Contrastive Supervision

优点：可以对生成的文字、视觉内容进行监督、易拓展；

Text-Content Guided Conditions

“Please **customize a LOGO** for the [150th] anniversary of [Tsinghua University]”

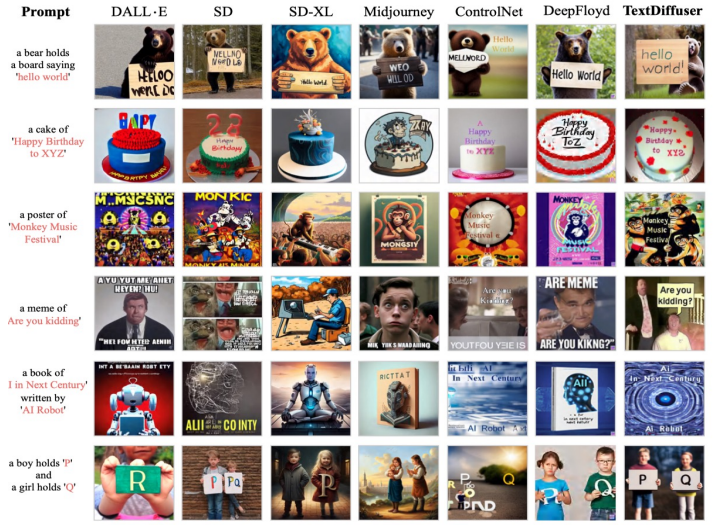
τ

Pretrained Diffusion Models

Visual-Content Guided Conditions

ν

$$\tilde{\epsilon}_\theta(z_t,t,\tau,\nu)=\omega_\mathrm{v}\cdot||\epsilon-\epsilon_\theta(z_t,t,f(\tau,\nu))||_2^2 + \omega_\mathrm{c}\cdot||g(c)-g(\tilde{\alpha}z_t+\tilde{\beta}\epsilon_\theta(z_t))||_2^2 + (1-\omega_\mathrm{v}-\omega_\mathrm{c})\cdot||\epsilon-\epsilon_\theta(z_t),f(\bar{\tau},\bar{\nu}))||_2^2$$



（修炼特技-2）创意场景图文生成（Creative Scene Image-Text Generation）

“ When AIGC starts to implement content customization, everything starts to really become valuable.” -- Small Pony.Ma

工作四：基于统一图文元素的海报生成（ Prompt-to-Poster, P2P）

2023.9.16-2024.3.15

□ 动机：
非刚性编辑具有时空连续性，依赖于自适应的条件扩散算法

SD： Backbone → **ControlNet：** Conditional Control → **Layout Diffusion：** Latent Knowledge / Codebook Sampling → **Prompt-to-Poster：** Action / Pose Guidance

● 针对性提升：
① 统一图文生成
② 个性化定制模版

□ 关键点： ① Denoising Target ② Content Supervision ③ Contrastive Supervision

优点：个性化定制内容和版式、端到端生成能力；

$$\tilde{\ell}_\theta(z_t,t,\tau,\nu)=\frac{1}{2}\ell_{\theta,T2V}+\frac{\eta}{2}\cdot||\epsilon_\iota-\epsilon_\theta(z_t,t,\iota,f(\tau,\nu))||_2^2+\frac{(1-\eta)}{2}\cdot||\epsilon-\epsilon_\theta(z_t,t,f(\tau,\nu))||$$

Layout Supervision Item

课题参考：

1. 图像和文本的统一扩散模型UniDiffusion，实现图文的统一生成（双流、单流架构）
2. 多源控制条件的文生图模型Multi-Controller（多指令、多模态、多条件控制源）
3. 可交互的图文编辑创作模型（类似SAM的交互形式或基于对话的人机交互图文编辑与创作）
4. **创意场景的图文统一生成（根据已有文本或视觉Prompts生成海报、宣传页、广告、PPT、图纸、产品说明书等含有图文元素的image）**
5. **带有文字标注和箭头示意的阐释型图像生成（输入蛋白质描述/化学公式/数理题干/科学论文/流程描述/神经网络描述，输出蛋白质结构示意/化学反应图示/数理图示/论文框架图/流程或UML图/神经网络图等）**
6. 主体对象保持的个性化图像编辑与生成（图像主体保持不变、创意风格个性化改变）
7. 基于协同扩散的可控文生图模型（扩散小模型协作完成，全局生成与局部编辑、风格迁移与主体对象保持、轮廓生成和细节刻画等）
8. **非刚性对象的图像编辑（属性、风格、动作、姿势、表情等的特定控制）**
9. 高效的扩散模型研究用于快速文生图部署（跳跃扩散、稀疏扩散、动量扩散、记忆扩散、局部扩散、注意力扩散等）
10. **场景(创意)文本可控生成（探索文本内容如何用diffusion的形式转换为特定风格、颜色、纹理、组合的艺术字图片或LOGO）**