GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing

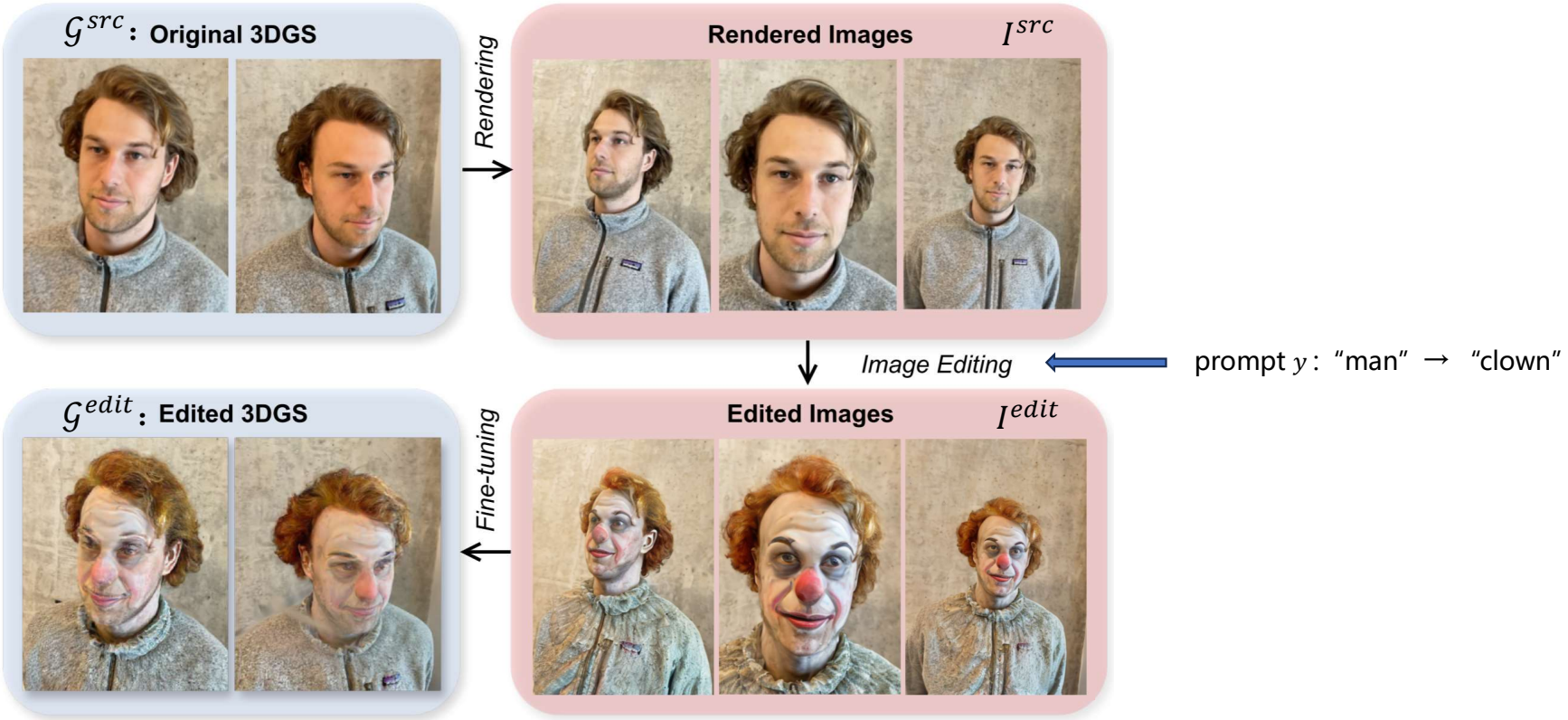
Jing Wu*[1], Jia-Wang Bian*[1], Xinghui Li[1], Guangrun Wang[1], Ian Reid[2], Philip Torr[1], and Victor Adrian Prisacariu[1]

[1] University of Oxford
[2] Mohamed bin Zayed University of Artificial Intelligence

Accepted by ECCV 2024

一、介绍

One Iteration



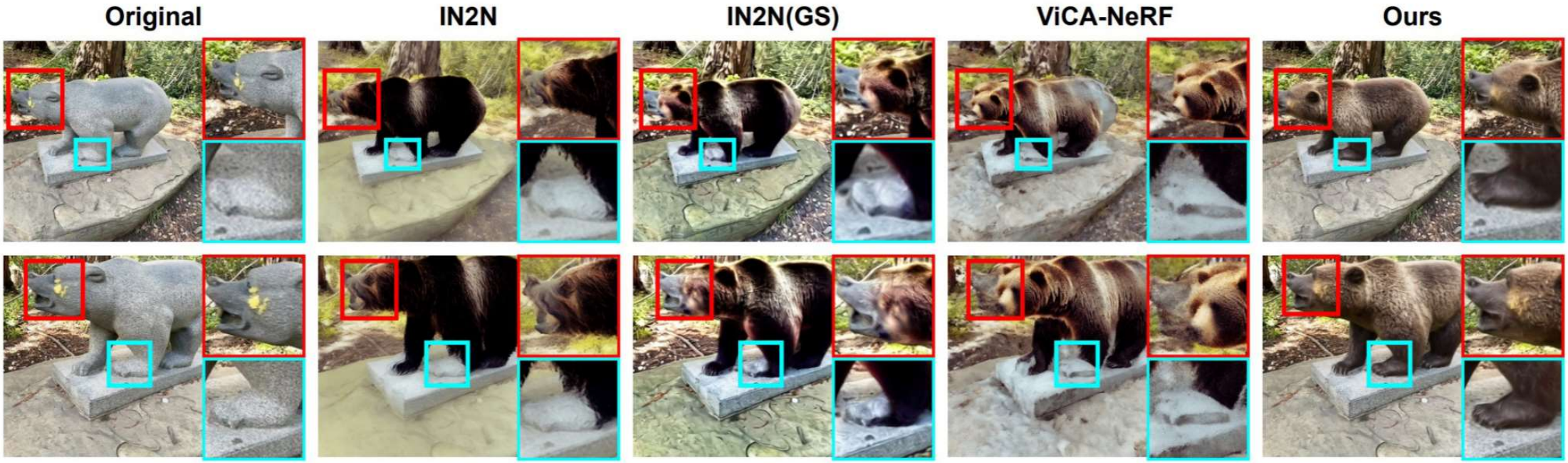
目的：给定 $\mathcal{G}^{src}$ 和 修改提示词 y，将 $\mathcal{G}^{src}$ 转换为与 y 对齐的编辑版本 $\mathcal{G}^{edit}$

- 从多个视图 $\mathcal{V}$=$\{v\}$ 对3DGS进行渲染，生成源图像集合I^{src}
- 利用2D编辑模型将I^{src} 根据文本提示词 y 转换为 I^{edit}
- 最后将 I^{edit} 用作训练指导，将 $\mathcal{G}^{src}$ 微调为 $\mathcal{G}^{edit}$

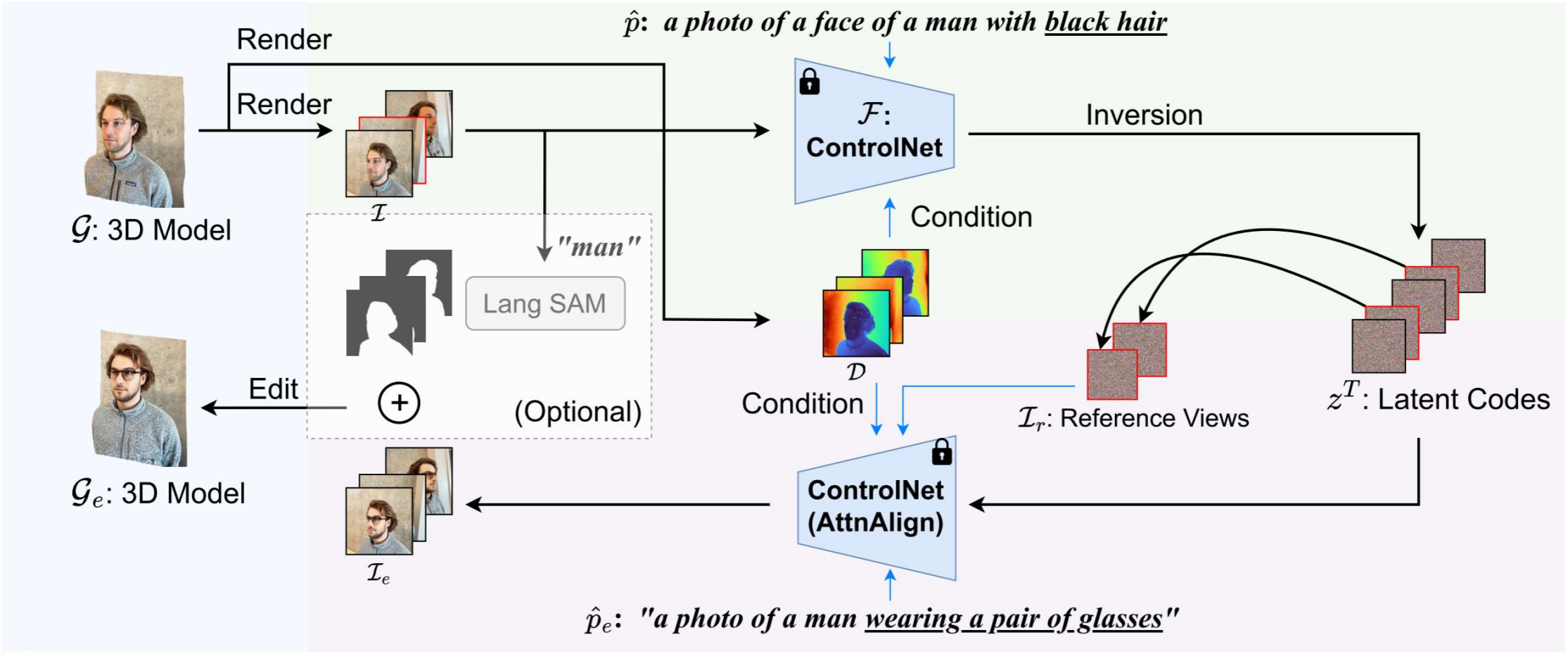
$$\mathcal{G}^{\text{edit}} = \underset{\mathcal{G}}{\operatorname{argmin}} \sum_{v \in \mathcal{V}} \mathcal{L}_{\text{Edit}}(\mathcal{R}(\mathcal{G}, v),\ \mathcal{I}^{\text{edit}}),$$

where $\mathcal{R}$ represents the rendering function that projects 3DGS to image given a specific view v.

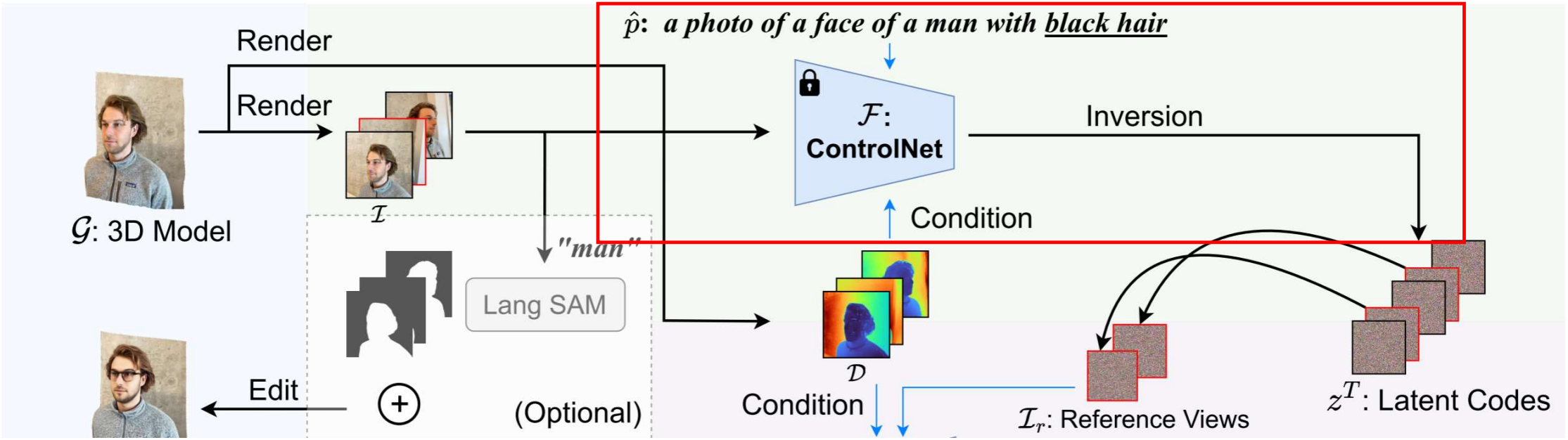
一、介绍



- 其他方法存在伪影，跨视图编辑不一致性
- 通过基于深度图约束和多视图潜在特征对齐，GaussCtrl具有更好的视觉效果。



- 利用多视图的深度图和source prompt，结合controlnet预测噪声，利用该噪声进行DDIM Inversion获得 latent code
- 在去噪过程中，利用参考视图，将编辑视图的self-attention和参考视图的self-attention进行特征对齐



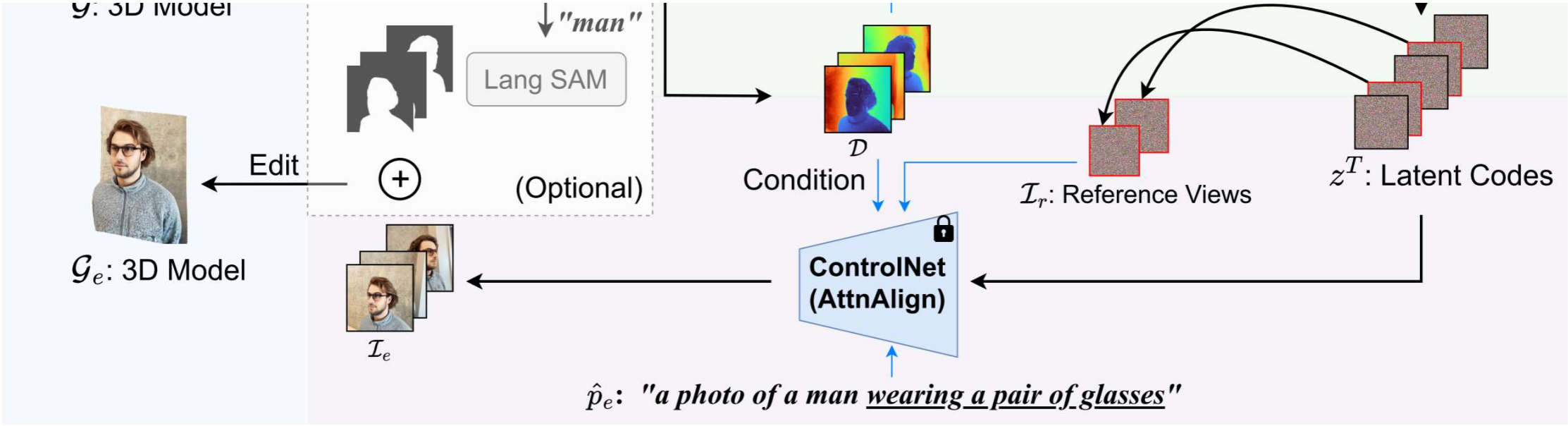
DDIM Inversion阶段

Given a to-be-edited image $\mathcal{I}$ and its corresponding description prompt $\hat{p}$, we begin by computing its latent code z^0, using the VAE encoder of the ControlNet. We then iteratively invert it to its corresponding Gaussian noise z^T via DDIM inversion. Mathematically, the inversion can be described as follows:

$$\epsilon^t = \mathcal{F}_U(z^t, t, \hat{p}, \mathcal{F}_C(z^t, t, \hat{p}, \mathcal{D}))$$

$$z^{t+1} = \sqrt{\alpha_{t+1}} \frac{z^t - \sqrt{1-\alpha_t} \cdot \epsilon^t}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t+1}} \epsilon^t$$

$\mathcal{F}_U$表示Unet， $\mathcal{F}_C$表示ControlNet



特征对齐阶段——将自身的self-attention和与参考视图的交叉特征对齐进行加权混合

Specifically, we first define the attention between two latent codes z_i and z_j as:

$$\text{Attn}_{i,j} = \text{Softmax}(\frac{W_q(z_i)W_k(z_j)^\top}{\sqrt{c}})W_v(z_j), \tag{7}$$

$$\text{AttnAlign}_e = \lambda \cdot \text{Attn}_{e,e} + (1-\lambda) \cdot \frac{1}{N_r}\sum_{i=1}^{N_r}\text{Attn}_{e,i}$$

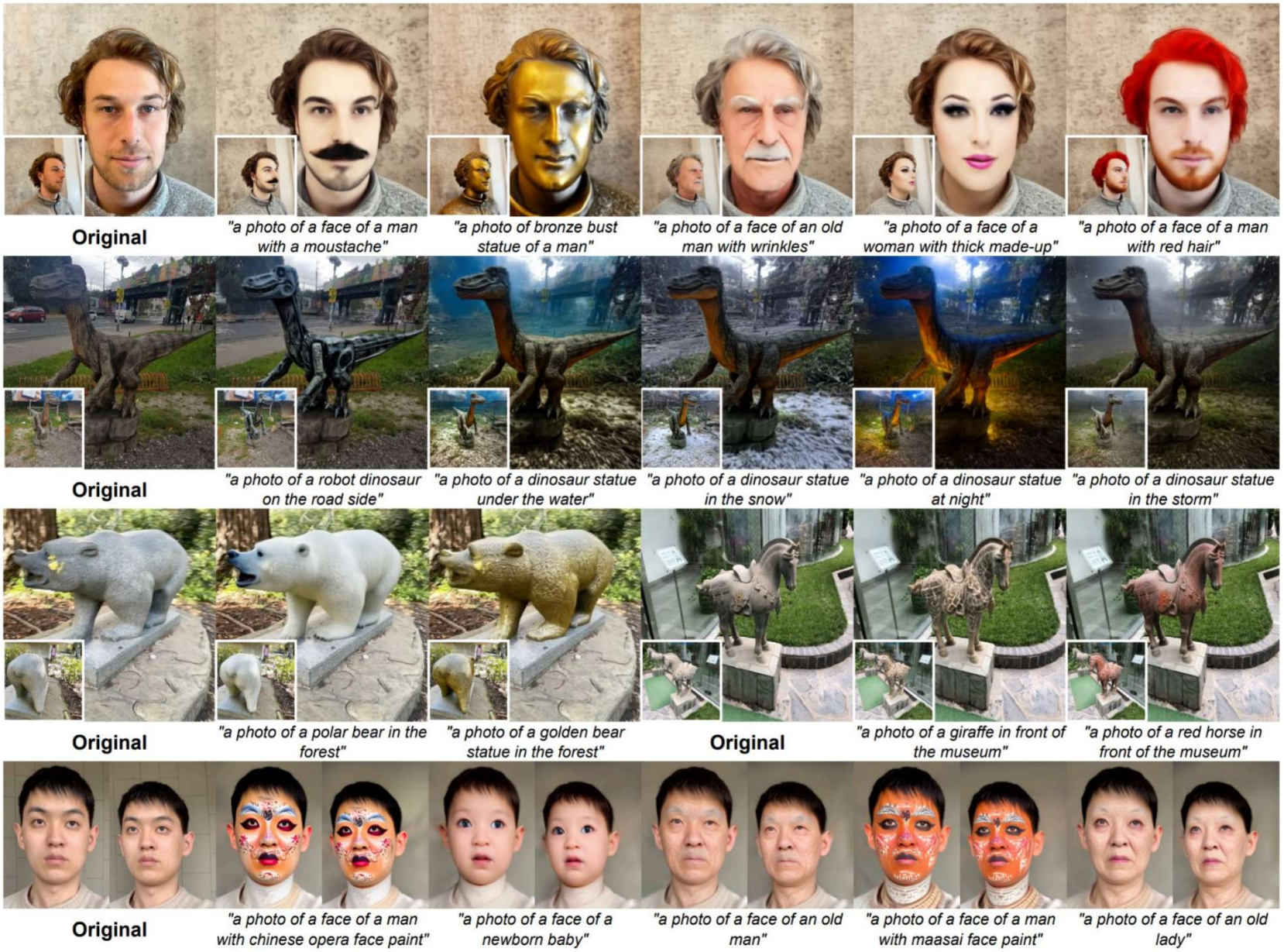
$$\epsilon^t = \mathcal{F}_U(z^t, t, \hat{p}, \mathcal{F}_C(z^t, t, \hat{p}, \mathcal{D}))$$

$$z^{t+1} = \sqrt{\alpha_{t+1}}\frac{z^t - \sqrt{1-\alpha_t}\cdot\epsilon^t}{\sqrt{\alpha_t}} + \sqrt{1-\alpha_{t+1}}\epsilon^t$$

$\mathcal{F}_U$表示Unet， $\mathcal{F}_C$表示ControlNet

Where $\lambda \in [0,1]$, e 表示当前正在编辑的图像

latent codes of N_r reference images, where $i = 1,2,\ldots,N_r$

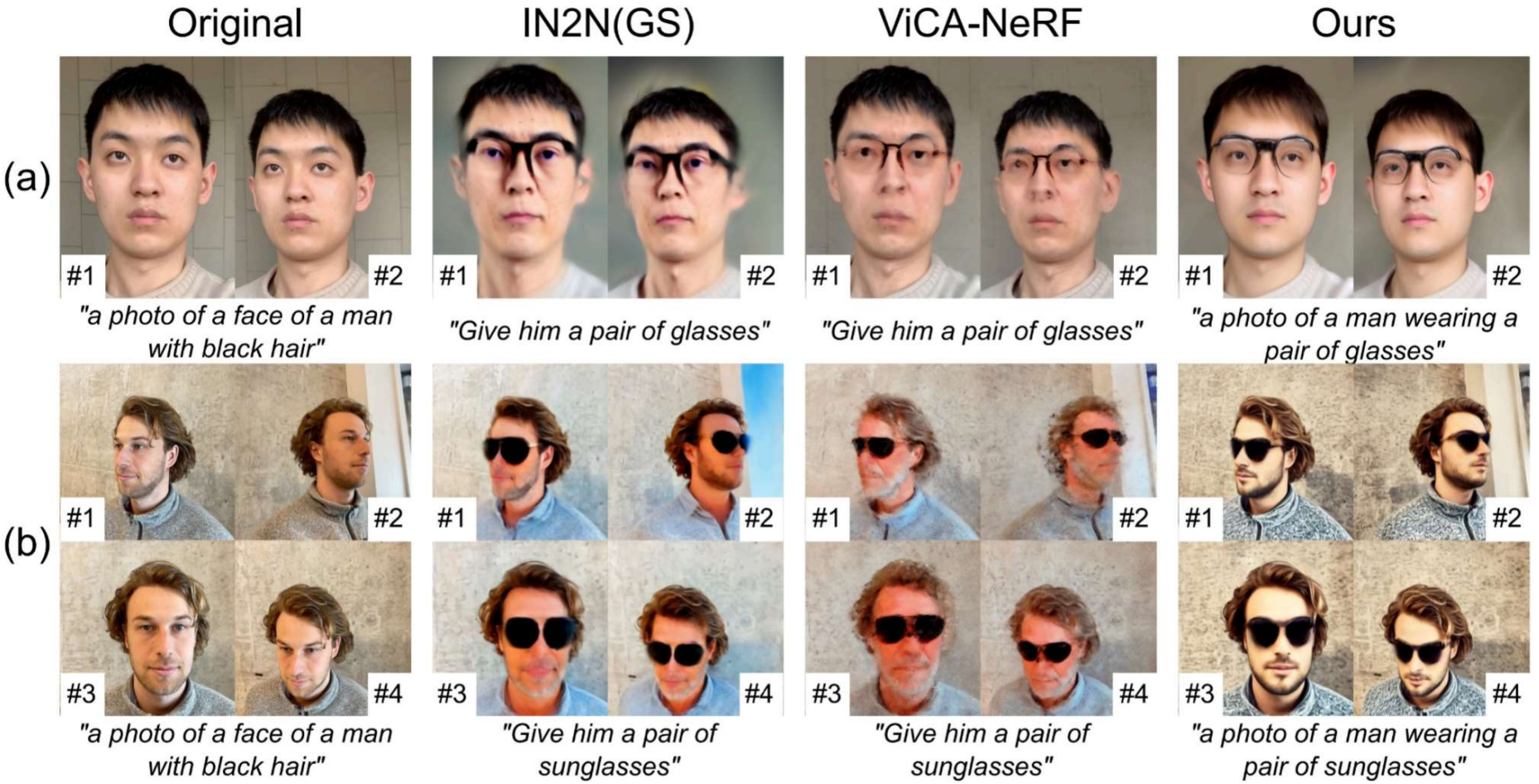


该方法展示了文本引导编辑
在不同场景下的不同结果，
从编辑对象到调整环境，
例如改变目标人的外表和年龄，
以及修改环境

三、Qualitative comparison

该方法相比以前的方法产生更一致和更高质量的图像。

三、Qualitative comparison



正面场景的定性结果。该方法能产生更真实的结果，具有更好的质量、一致性和更少的伪影。

三、Qualitative comparison

编辑一致性比较。

- IN2N和ViCA在#6上出现了背后视角生成脸部的情况
- 在（视图#1,2,4,8,10）上存在编辑不一致情况，导致了伪影和熊脸的模糊

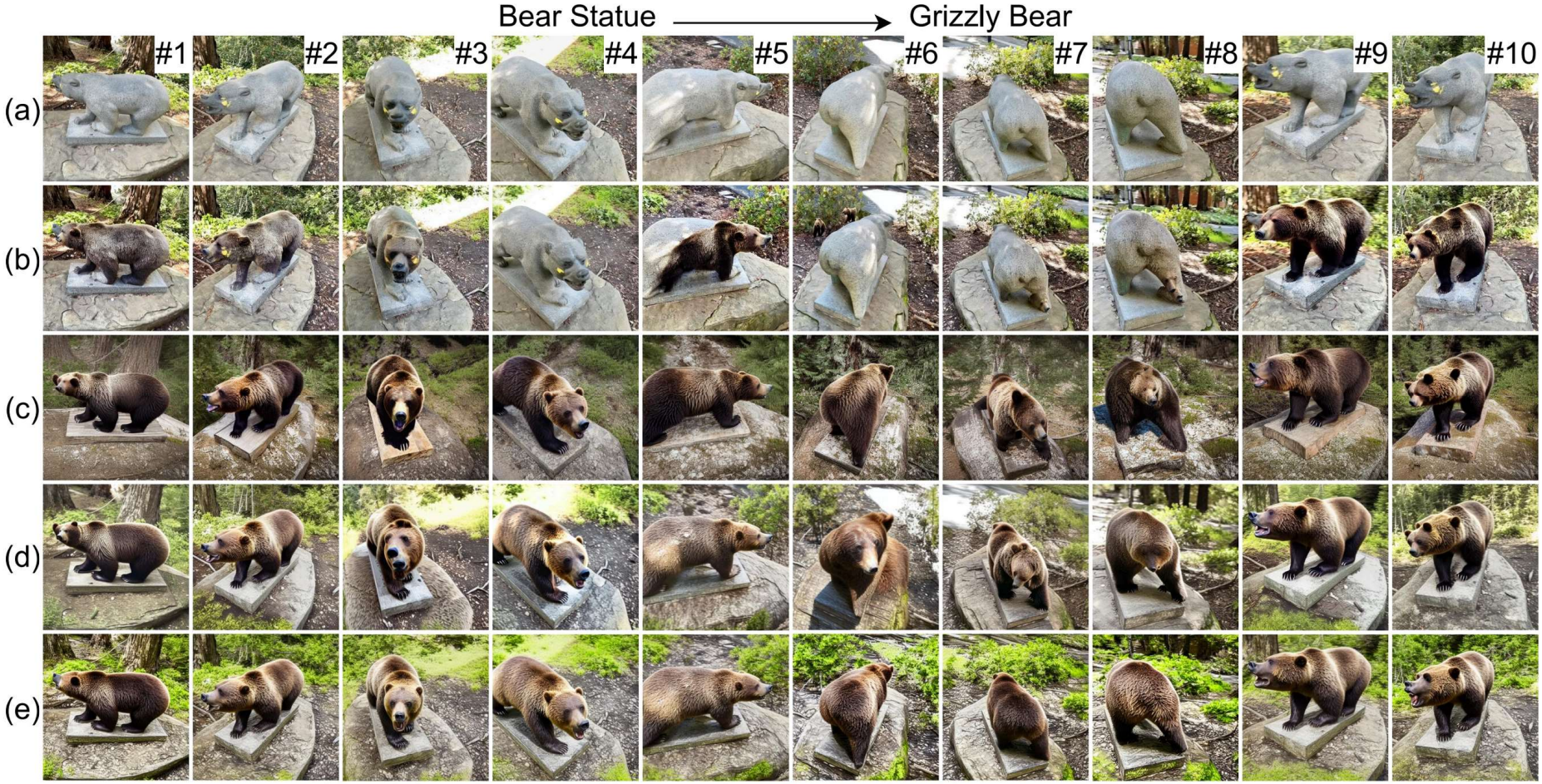
四、Quantitative Comparisons

Table 1: Quantitative Evaluation. CLIP_{dir}: CLIP Text-Image Direction Similarity

	Scene	IN2N		IN2N(GS)		ViCA-NeRF		**Ours**	
		CLIP_{dir}	Time	CLIP_{dir}	Time	CLIP_{dir}	Time	CLIP_{dir}	Time
360	Bear Statue	0.1019	~1.5h	0.1165	~13.5min	0.1104	~38.5min	**0.1388**	~9min
	Dinosaur	0.1466		0.1490		0.0723		**0.1584**	
	Garden	**0.3027**		0.1663		0.2903		0.2891	
	Stone Horse	0.1654		0.1947		0.1926		**0.2268**	
Forward	Fangzhou	0.1598		**0.2032**		0.1809		0.1887	
	Face	0.1332		0.1357		0.1119		**0.1503**	

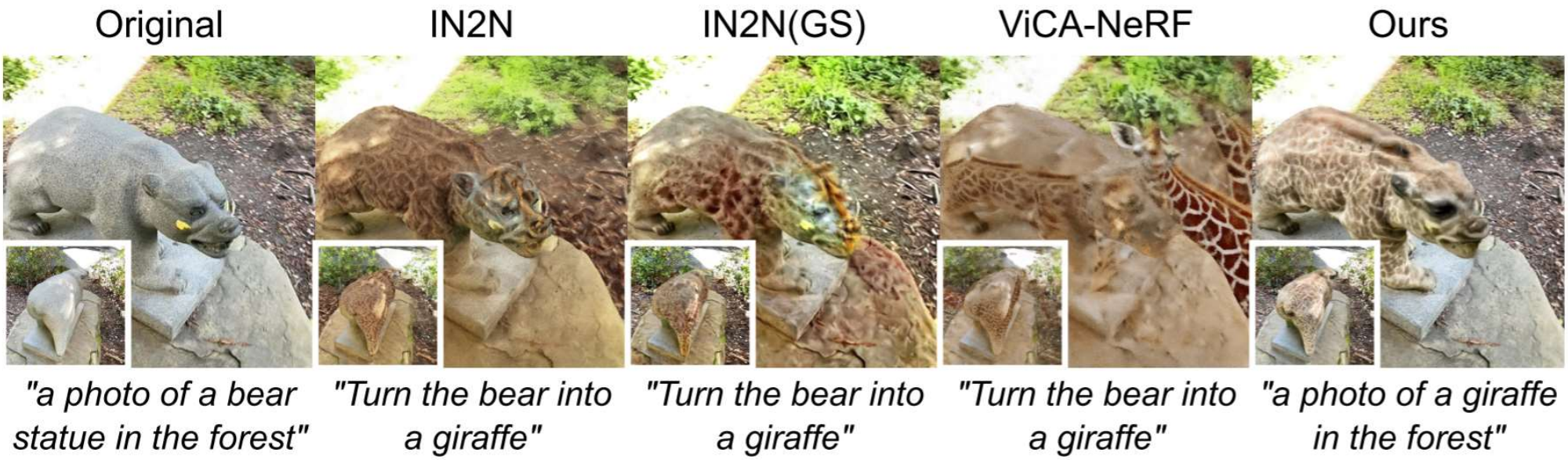
- 该方法相比以前的方法，编辑效率更高所需时间更少
- 在CLIP指数上也能达到良好的编辑效果

五、Ablation Study--Consistency Modules



- (a)：从三维模型中渲染出来的多视图。(b)：使用InstructPix2Pix编辑结果。
- (c)：仅使用该方法提出的深度图引导编辑，它使用ControlNet和随机初始化的latent code。
- (d)：在(c)基础上采用DDIM反演，得到一致的initial latent code。(e)：在(d)的基础上添加基于Attention的特征对齐模块。

Failure cases



- 由于使用深度引导，当需要重大的几何变化时，该方法不能很好地工作。
- 而且，他们发现即使不使用深度，也不能很好地在这种情况下工作。

Thank You