GSV3D: Gaussian Splatting-based Geometric Distillation with Stable Video Diffusion for Single-Image 3D Object Generation

Ye Tao[1], Jiawei Zhang[2], Yahao Shi[1], Dongqing Zou[2,3], and Bin Zhou[1,3]*
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]SenseTime Research [3]Peace-Bay Academy of Virtual Reality

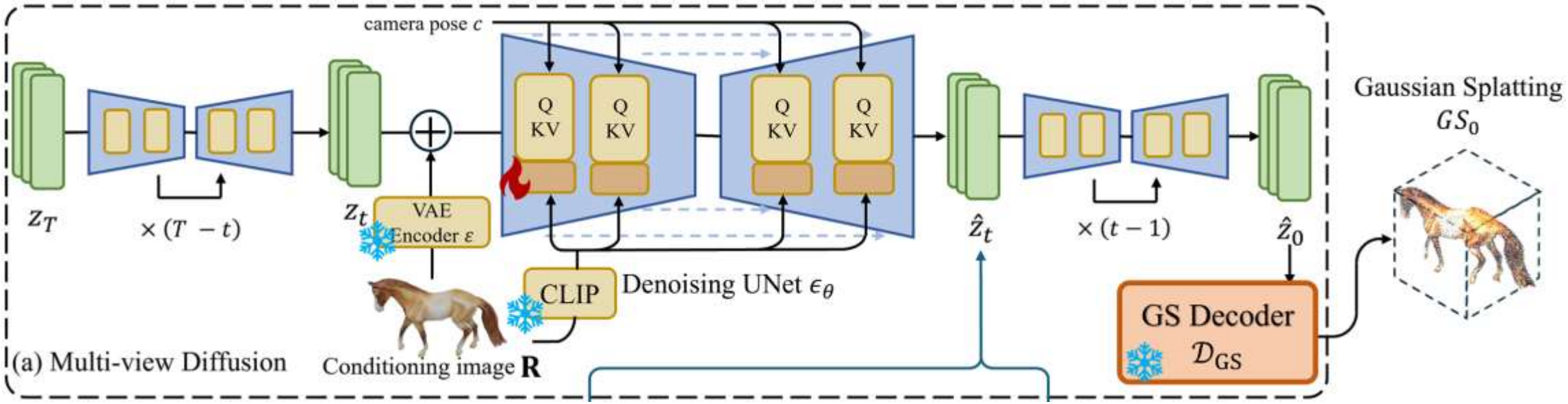
Accepted by ICCV 2025

一、介绍 - 任务：给定单张图像，生成一个3D模型

背景问题

- 在现有的单张图片生成3D模型的任务上
  - 3D diffusion模型：直接建模 3D，但数据量小、多样性差
    - 3D 扩散模型受到数据集稀缺和缺乏强预训练先验的限制
  - 2D diffusion模型：生成的纹理细节好，但生成的多视图存在多视角几何不一致的问题
    - 现有方法主要应用隐式约束，这些隐式约束缺乏显式的一致几何对齐

二、框架（推理部分）

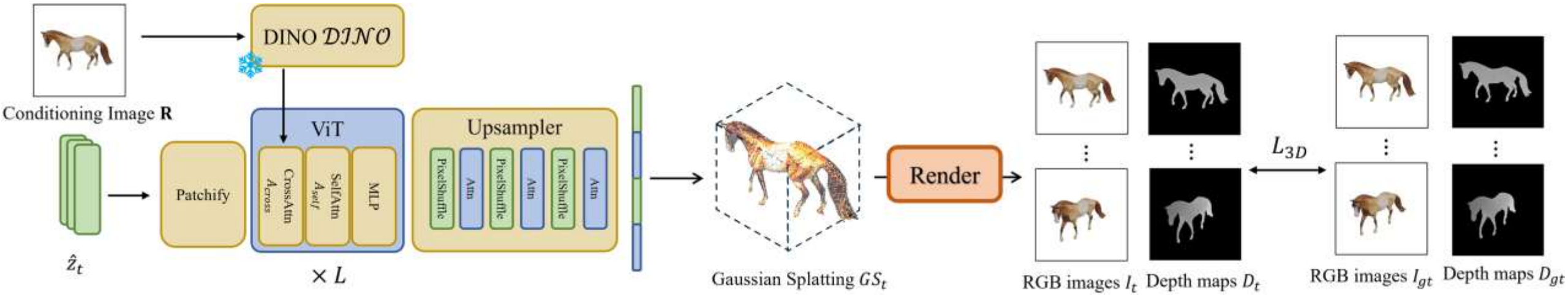


- 输入单图 → SV3D (2D diffusion) 生成多视角 latent
- 最后，多视角latent → GS Decoder → 输出 3D Gaussian Splatting

论文核心

- 提出GSV3D，结合 2D diffusion 的丰富纹理 + 3D 方法显式约束几何一致性
  - 先用 2D diffusion 生成多视角 latents
  - 提出 GS Decoder 利用 Gaussian Splatting 强制几何一致性
  - 利用几何蒸馏损失+lora 微调 diffusion，增强3D结构

二、框架 – GS Decoder



DINO：注入全局语义结构

- 提取输入图像 R 的高层语义特征：$\mathbf{F}_{\text{DINO}} = \mathcal{DINO}(\mathbf{R})$.
- 语义锚点作用，语义特征送入 ViT 的 cross-attention

$$\mathbf{F}_{\text{ViT}} \leftarrow \mathcal{A}_{\text{cross}}(\mathbf{F}_{\text{ViT}}, \mathbf{F}_{\text{DINO}}),$$

GSDecoder 训练：

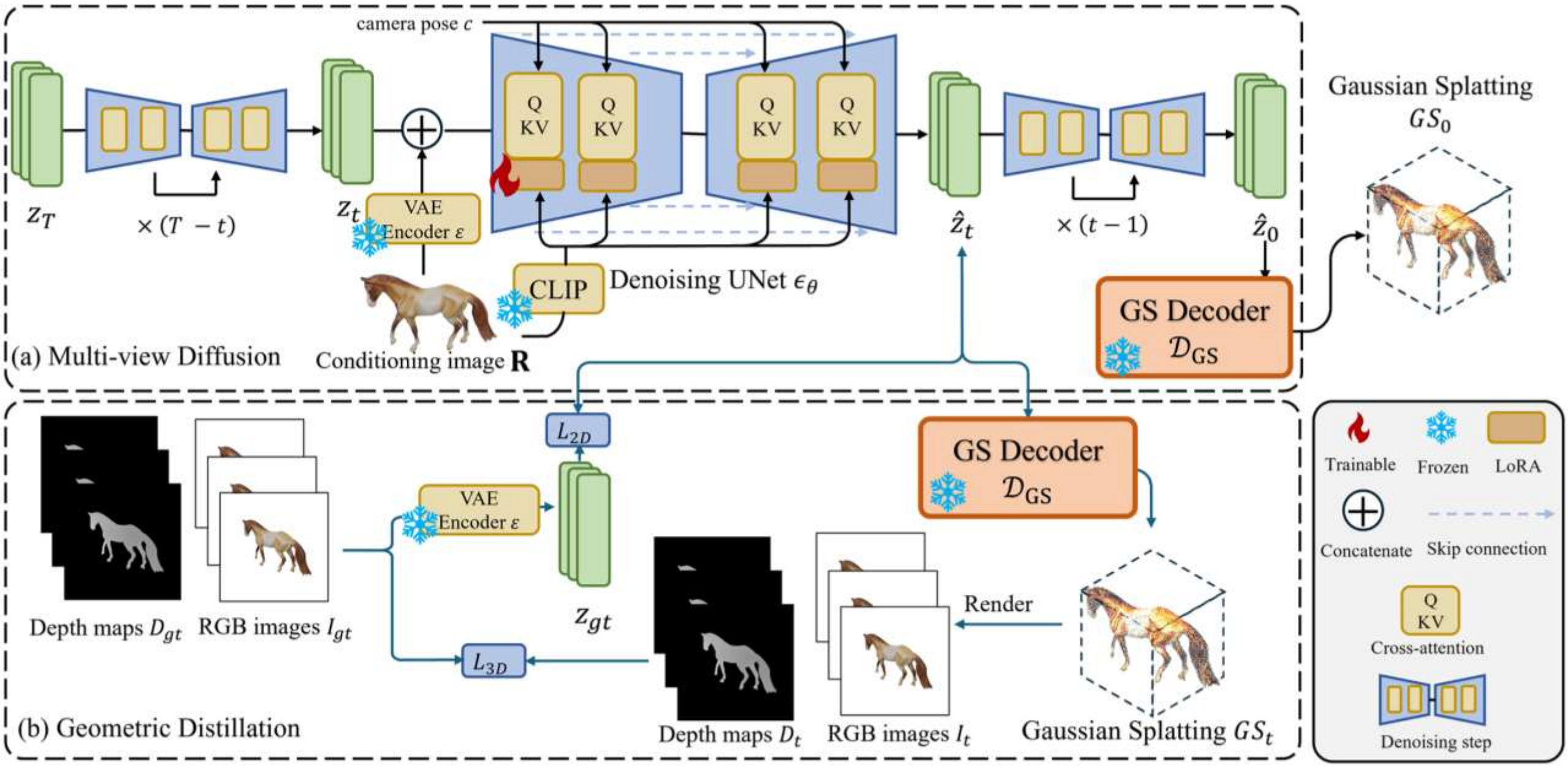
- 数据集由 P 个3D模型，每个模型渲染 Q 张视角图片
- 从 Q 张视角图里随机挑选 N 张图片作为输入
- 获取最终的 GS 输出并渲染出 Q 张多视角结果图

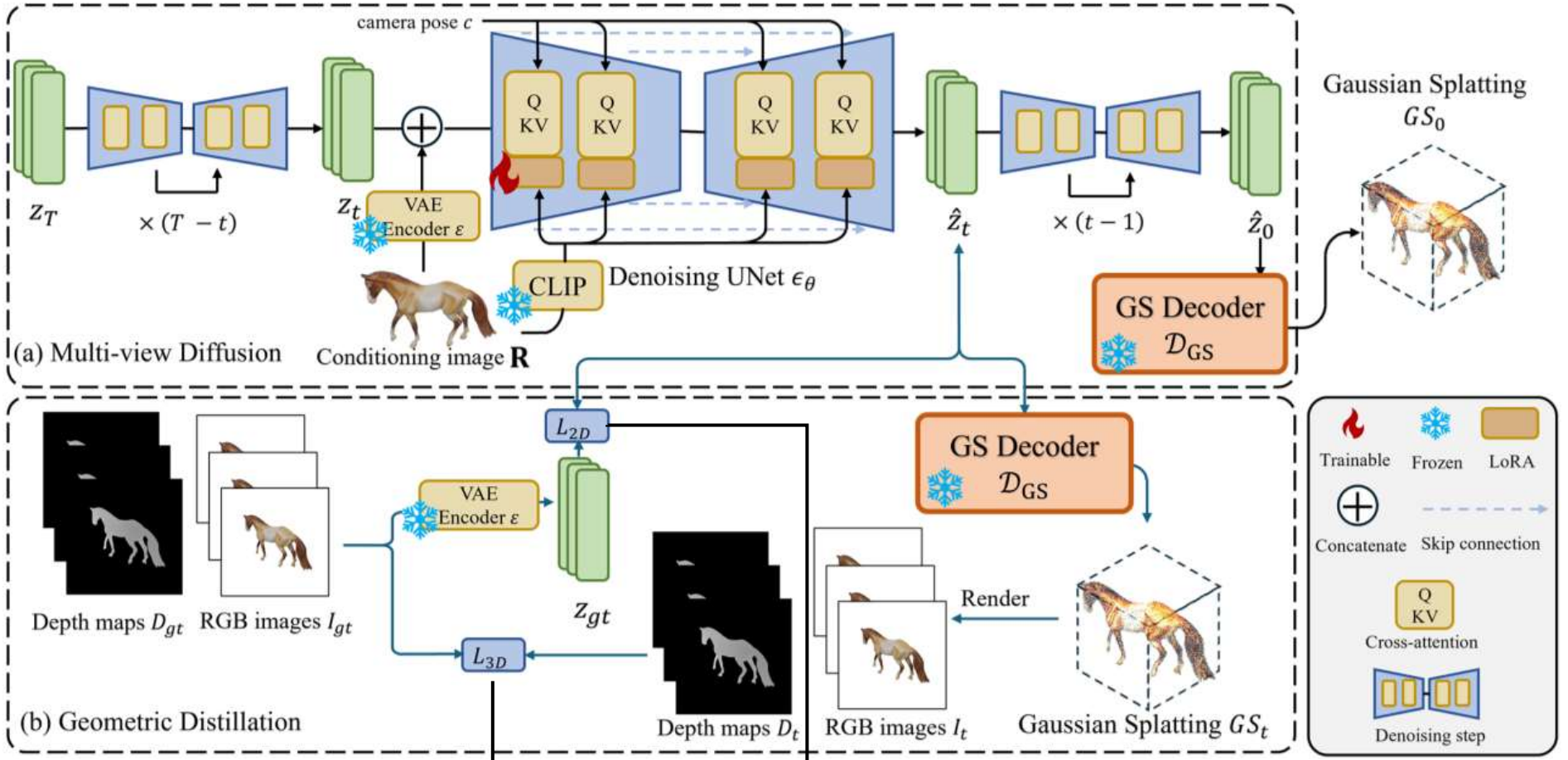
$$\mathcal{L}_{\text{rgb}} = \frac{1}{P \cdot Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \left\| I_{\text{output}}^{(p,q)} - I_{\text{gt}}^{(p,q)} \right\|_2^2.$$

$$\mathcal{L}_{\text{depth}} = \frac{1}{P \cdot Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \| D_{\text{output}}^{(p,q)} - D_{\text{gt}}^{(p,q)} \|_2^2.$$

$$\mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}},$$

二、框架（训练部分）


camera pose c
z_T
$\times (T - t)$
z_t
VAE Encoder ε
Q KV
CLIP
Denoising UNet ϵ_θ
Conditioning image $\mathbf{R}$
$\hat{z}_t$
$\times (t - 1)$
$\hat{z}_0$
GS Decoder $\mathcal{D}_{GS}$
Gaussian Splatting GS_0
(a) Multi-view Diffusion
Depth maps D_{gt}
RGB images I_{gt}
L_{2D}
VAE Encoder ε
z_{gt}
L_{3D}
Depth maps D_t
RGB images I_t
Render
GS Decoder $\mathcal{D}_{GS}$
Gaussian Splatting GS_t
(b) Geometric Distillation
Trainable
Frozen
LoRA
Concatenate
Skip connection
Q KV
Cross-attention
Denoising step



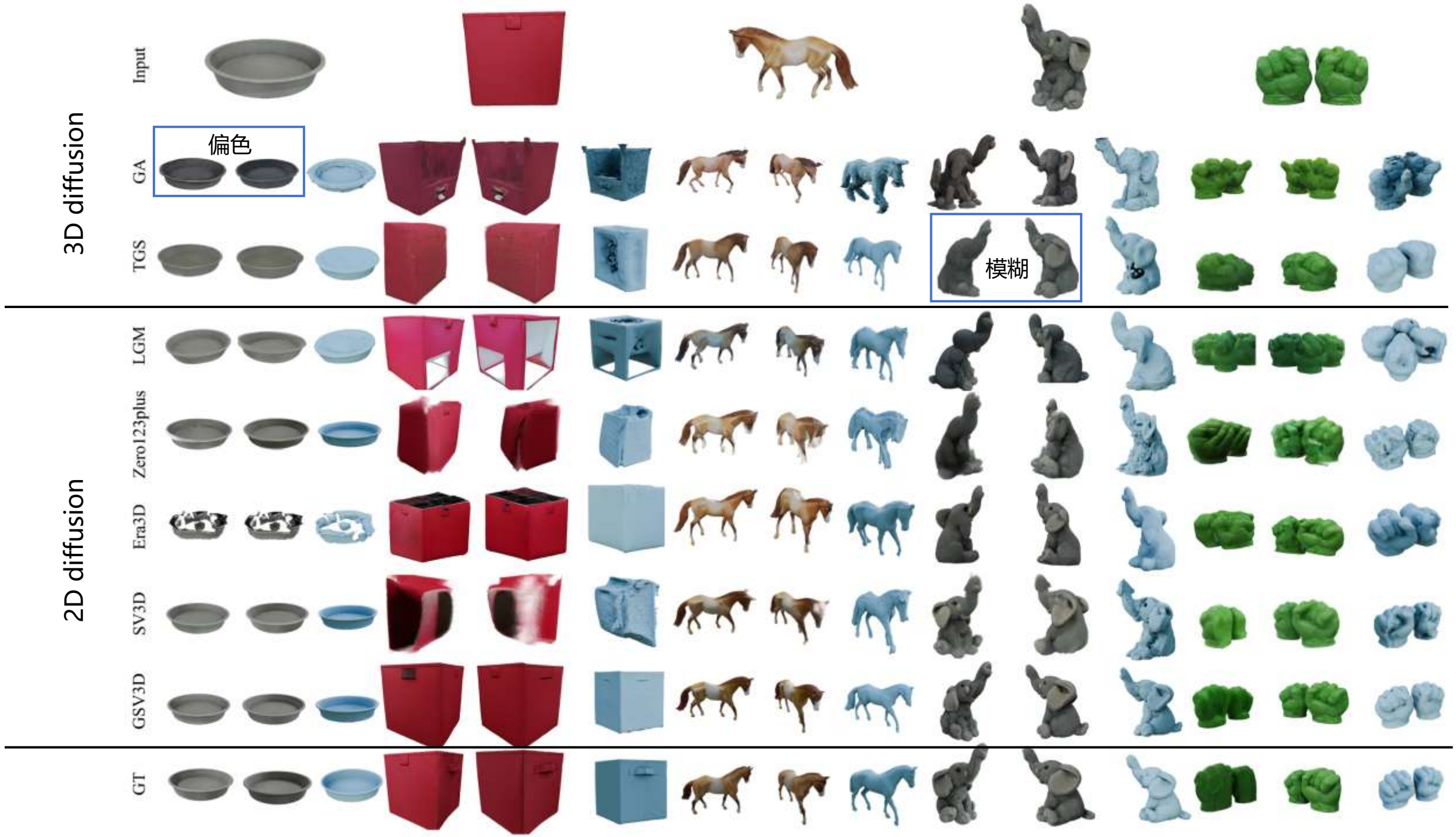
$$\mathcal{L}_{\text{rgb}} = \frac{1}{P \cdot Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \left\| I_{\text{output}}^{(p,q)} - I_{\text{gt}}^{(p,q)} \right\|_2^2.$$

$$\mathcal{L}_{\text{depth}} = \frac{1}{P \cdot Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \| D_{\text{output}}^{(p,q)} - D_{\text{gt}}^{(p,q)} \|_2^2.$$

$$\mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{depth}} \mathcal{L}_{\text{depth}},$$

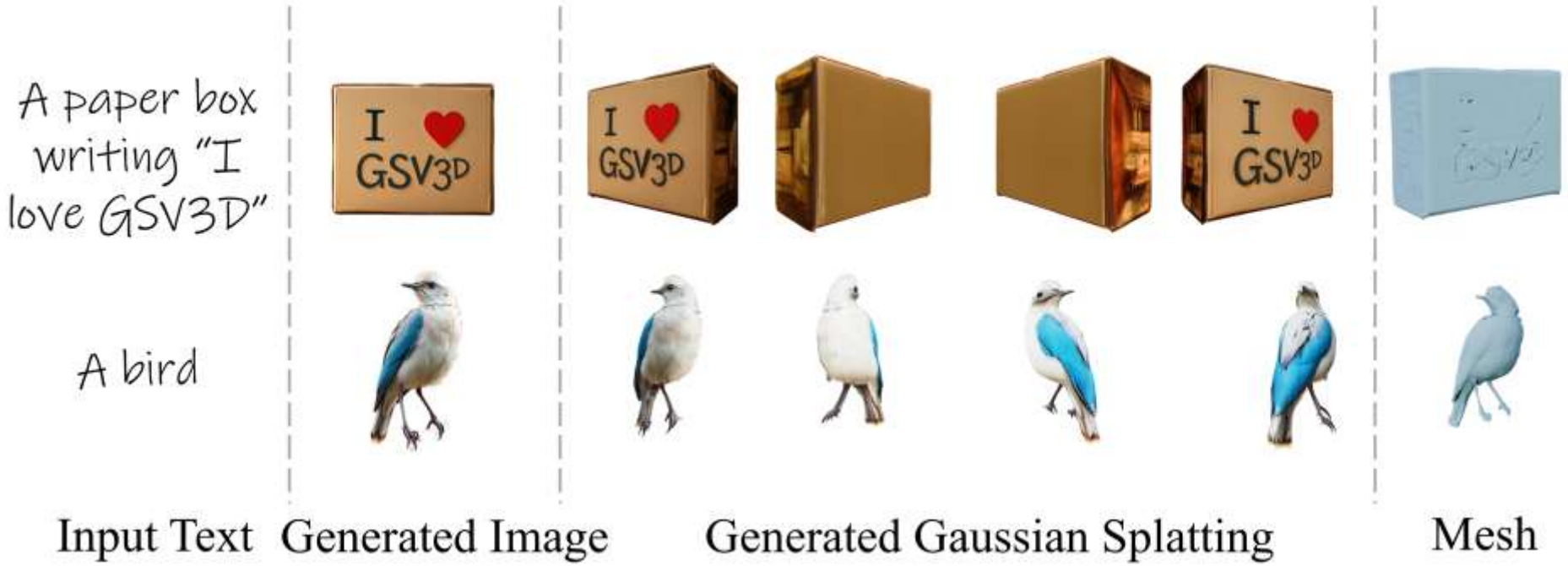
$$\mathcal{L}_{\text{2D}} = \mathbb{E}_{\mathbf{R}, z_t, z_{gt}} \left[\| \epsilon_\theta(z_t; \mathbf{R}, t) - z_{gt} \|_2^2 \right],$$

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{2D}} + \lambda_{\text{3D}} \mathcal{L}_{\text{3D}},$$


3D diffusion
2D diffusion
Input
GA
TGS
LGM
Zero123plus
Era3D
SV3D
GSV3D
GT
偏色
模糊

Type	Method	Appearance Quality						Geometry Quality			User Study	
		PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	CLIP-IQA↑	CD↓	IoU ↑	F-Score↑	App. Score↑	Geo. Score↑
3D	GA [15]	15.201	0.834	0.039	95.47	1.17	0.805	0.197	0.502	0.303	3.154	5.000
	TGS [39]	18.874	0.872	0.032	85.25	1.28	0.812	0.272	0.415	0.232	4.000	3.846
2D	LGM [30]	17.013	0.845	0.033	61.66	0.45	0.819	0.238	0.431	0.308	3.538	4.000
	Zero123Plus [28]	15.787	0.827	0.037	82.80	0.82	0.825	0.165	0.525	0.409	2.088	2.154
	Era3D [17]	14.993	0.830	0.040	90.35	0.92	0.811	0.229	0.436	0.256	3.846	2.418
	SV3D [31]	17.772	0.863	0.034	74.86	0.80	0.816	0.200	0.418	0.311	4.231	2.923
2D-3D	GSV3D	20.390	0.884	0.023	50.79	0.21	0.839	0.100	0.721	0.661	6.308	6.692

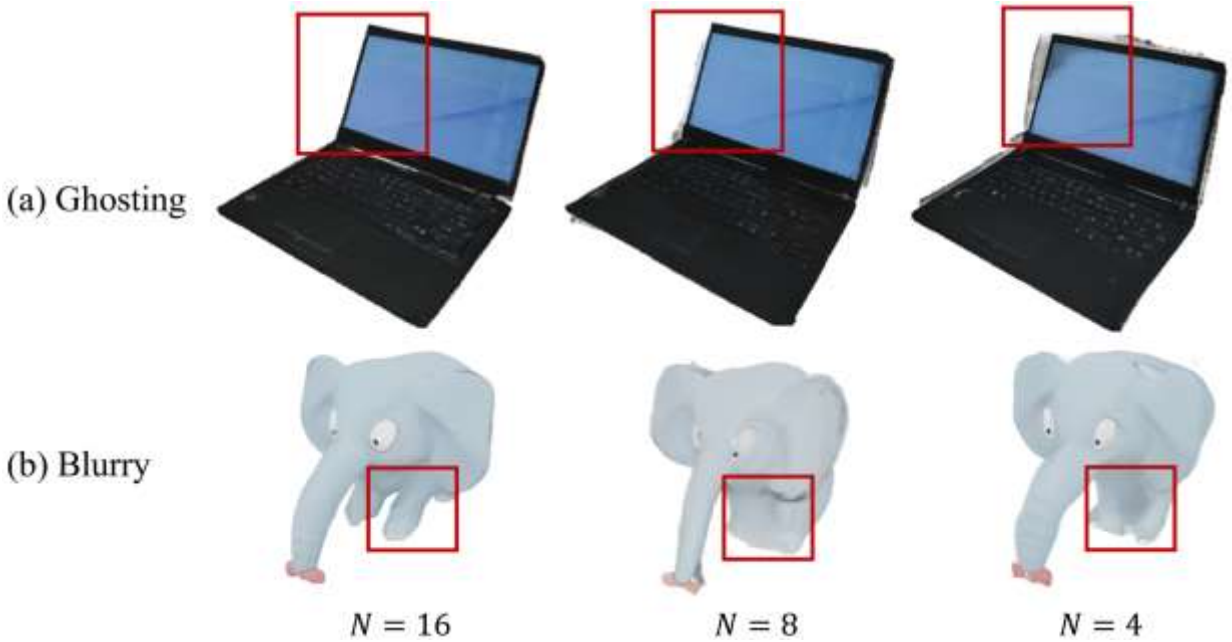
红色：第一 ； 橙色：第二 ； 黄色：第三;



使用 GSV3D 进行 纯文本生成的例子

四、消融实验

Setting	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID(%)↓
Geometric Distillation					
w/o RGB loss	19.551	0.872	0.027	53.89	0.27
w/o Depth loss	19.826	0.877	0.025	53.20	0.32
Number of Frames					
N = 8	19.733	0.877	0.026	55.49	0.34
N = 4	19.548	0.874	0.027	57.85	0.42
DINO encoder					
w/o DINO encoder	19.508	0.876	0.027	55.20	0.30
Full Model (GSV3D)	**20.390**	**0.884**	**0.023**	**50.79**	**0.21**



减少推理帧数会导致伪影和模糊问题

Thank You