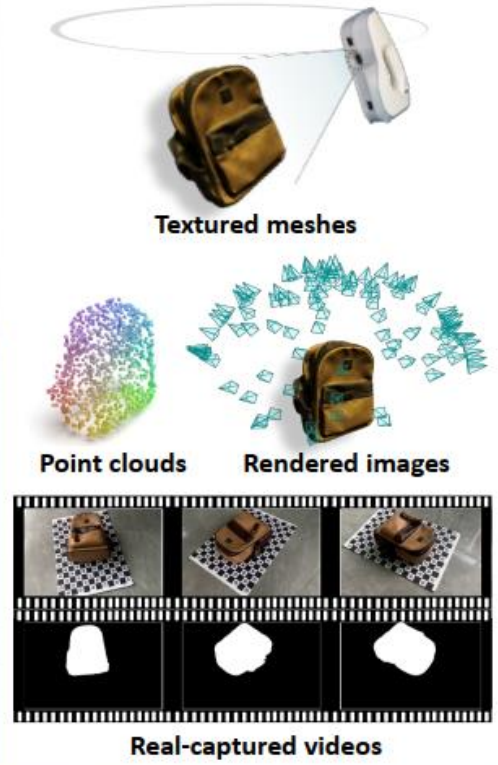
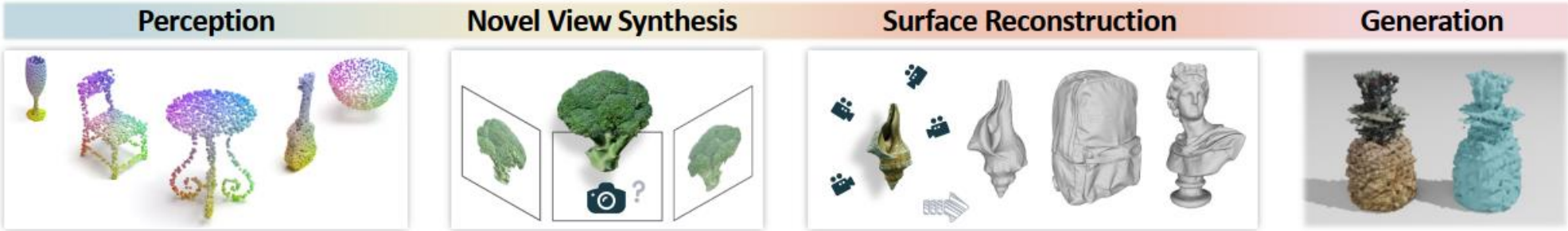
Part.01

OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation

OmniObject3D,
这是一个包含大量高质量真实扫描 3D 对象的大词汇量 3D 对象数据集，以促进现实世界中 3D 感知、重建和生成的发展。





1）大词汇量：它包括**190个**日常类别中的**6,000个**扫描对象，与流行的2D数据集（例如ImageNet和LVIS）共享公共类别，有助于追求可推广的3D表示。

2）丰富注释：每个3D对象都使用2D和3D传感器捕获，提供**纹理网格、点云、多视角渲染图像和多个真实捕获视频**。

3）逼真扫描：专业扫描仪支持高质量物体扫描，具有精确形状和逼真外观。

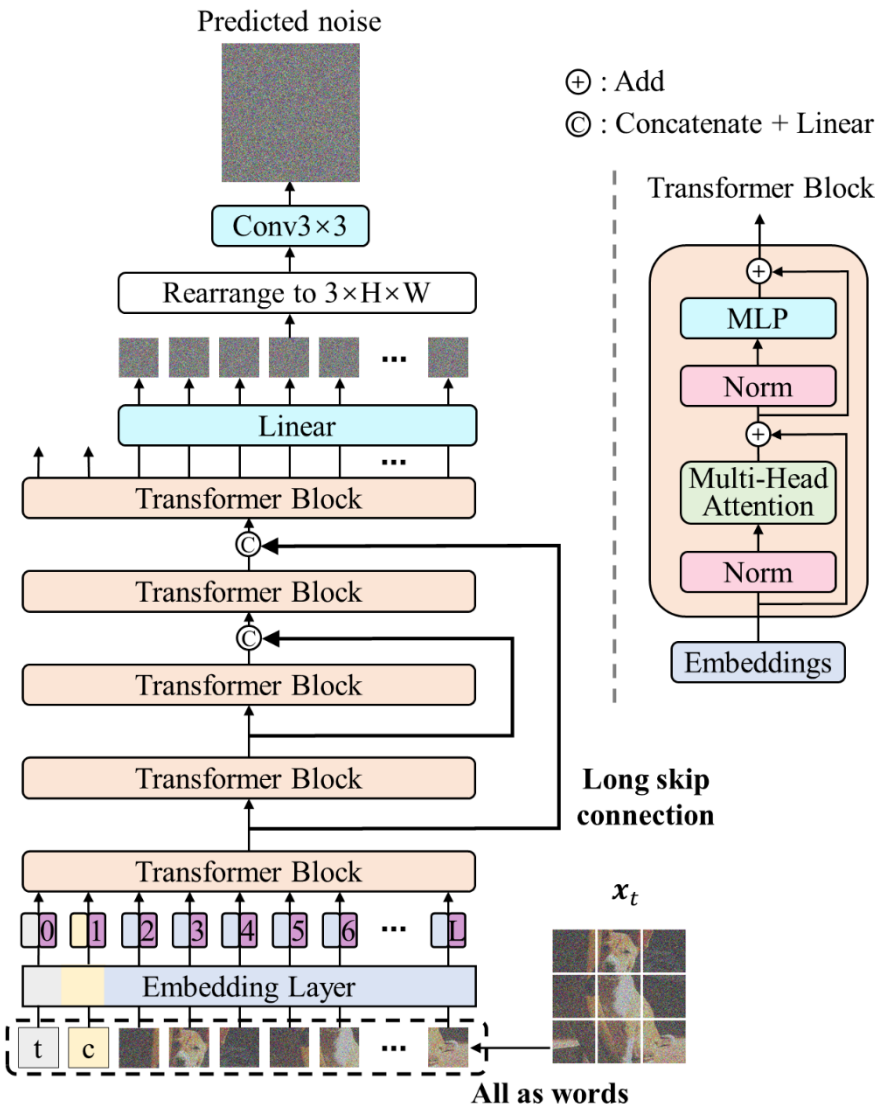
Statistics and Distribution

Dataset	Real	Full 3D	Video	#Objects	#Classes	R^{LVIS}(%)
ShapeNet		✓		**51k**	55	4.1
ModelNet		✓		12k	40	2.4
3D-Future		✓		16k	34	1.3
ABO		✓		8k	63	3.5
Toys4K		✓		4k	105	7.7
CO3D	✓		✓	19K	50	4.2
DTU	✓	✓		124	-	0
ScanObjectNN	✓			15k	15	1.3
GSO	✓	✓		1k	17	0.9
AKB-48	✓	✓		2k	48	1.8
Ours	✓	✓	✓	6k	**190**	**10.8**

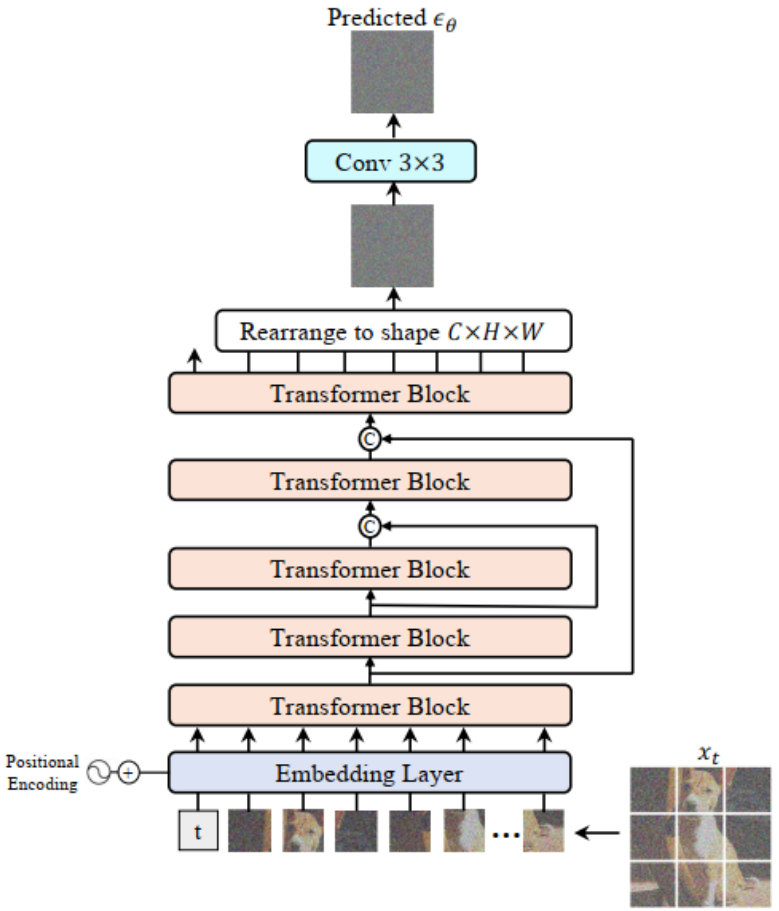
Part.02

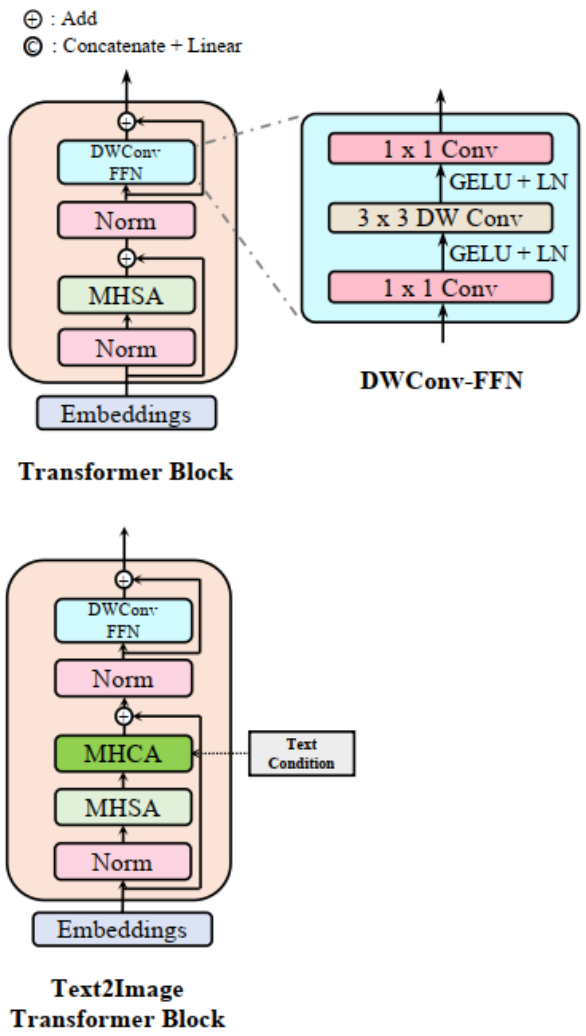
Exploring Vision Transformers as Diffusion Learners

U-ViT


Predicted noise
⊕ : Add
© : Concatenate + Linear
Conv3×3
Rearrange to 3×H×W
Linear
Transformer Block
Transformer Block
Transformer Block
Transformer Block
Transformer Block
Long skip connection
0 1 2 3 4 5 6 ... L
Embedding Layer
t c
x_t
All as words
Transformer Block
MLP
Norm
Multi-Head Attention
Norm
Embeddings

IU-ViT


Predicted ϵ_θ
Conv 3×3
Rearrange to shape $C \times H \times W$
Transformer Block
Transformer Block
Transformer Block
Transformer Block
Transformer Block
Positional Encoding
Embedding Layer
t
x_t


⊕ : Add
© : Concatenate + Linear
DWConv FFN
Norm
MHSA
Norm
Embeddings
Transformer Block
1 x 1 Conv
GELU + LN
3 x 3 DW Conv
GELU + LN
1 x 1 Conv
DWConv-FFN
DWConv FFN
Norm
MHCA
Text Condition
MHSA
Norm
Embeddings
Text2Image Transformer Block

ASymmetriC ENcoder Decoder (ASCEND)

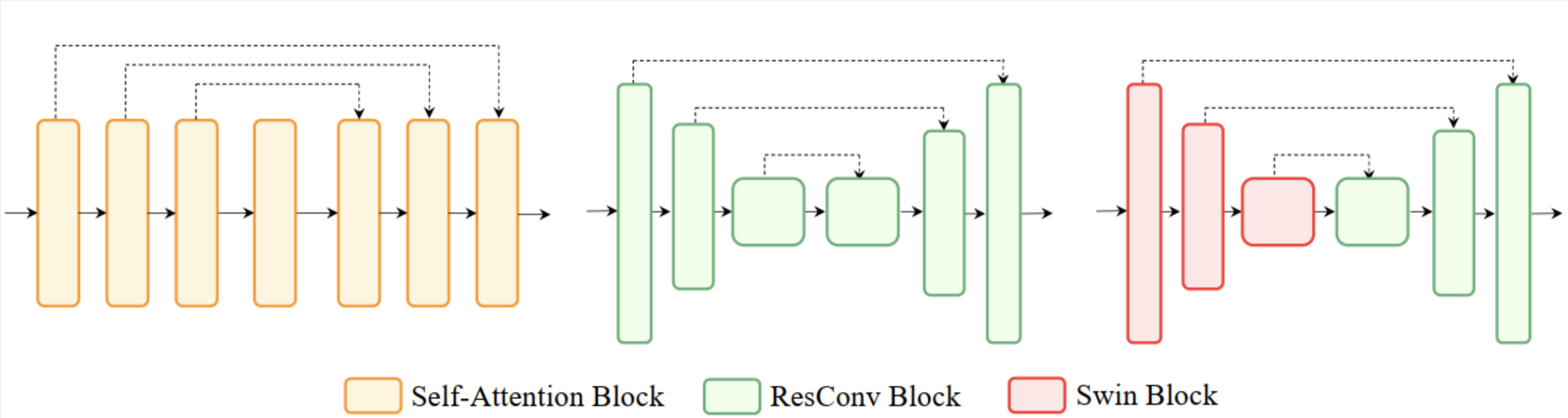
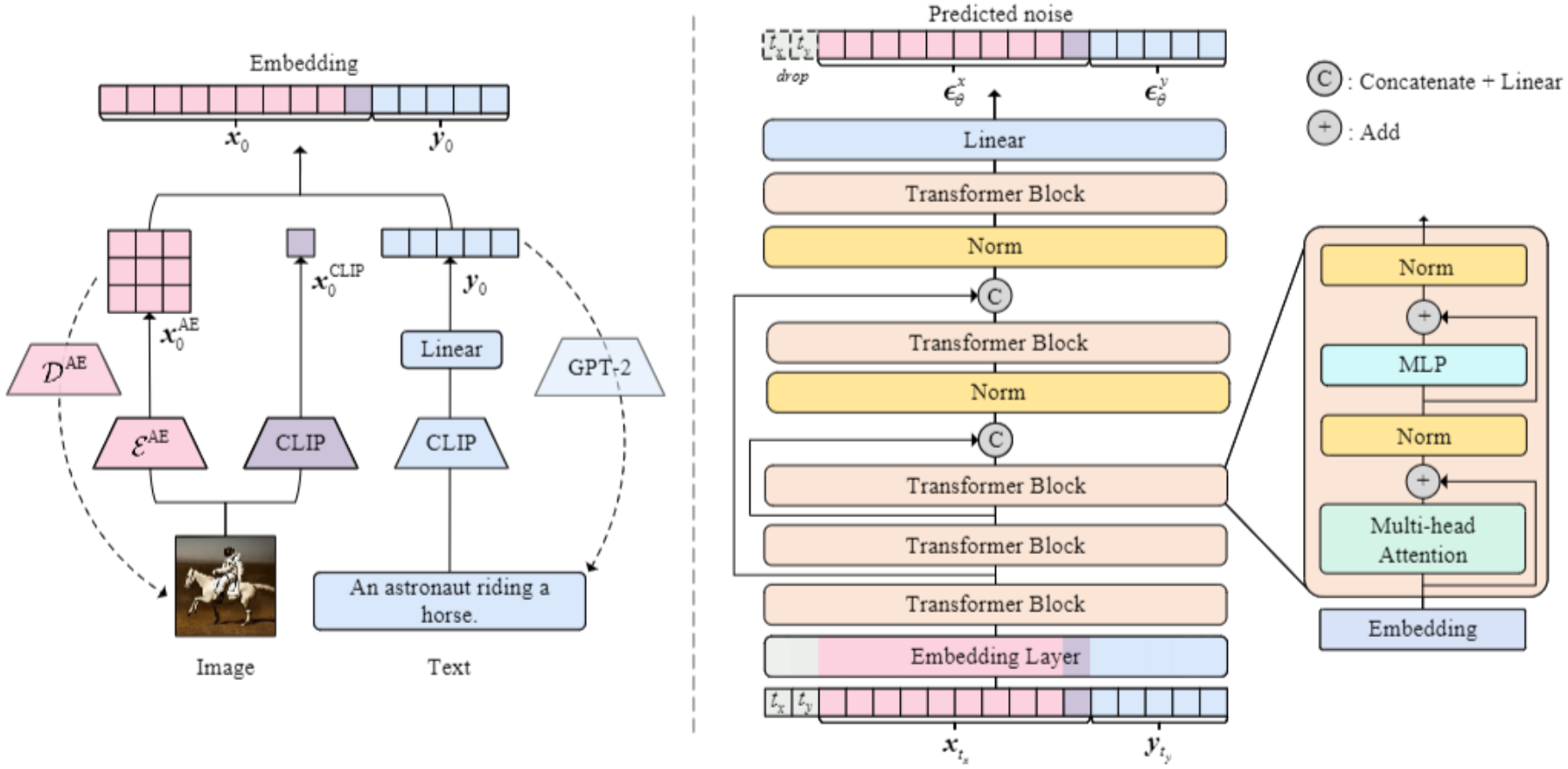


Figure 3. Illustrative architecture of U-ViT (Left), U-Net (Middle) and ASCEND (Right).

UnidifFuser 能够通过设置适当的时间步长生成各种类型的数据，例如图像、文本、文本到图像、图像到文本和图像-文本对。
UnidifFuser 在FID和CLIP分数的定量结果中不仅优于现有的通用模型，而且在文本到图像生成等代表性任务中还可以与Stable Difsion和DALL•E 2等定制模型相媲美



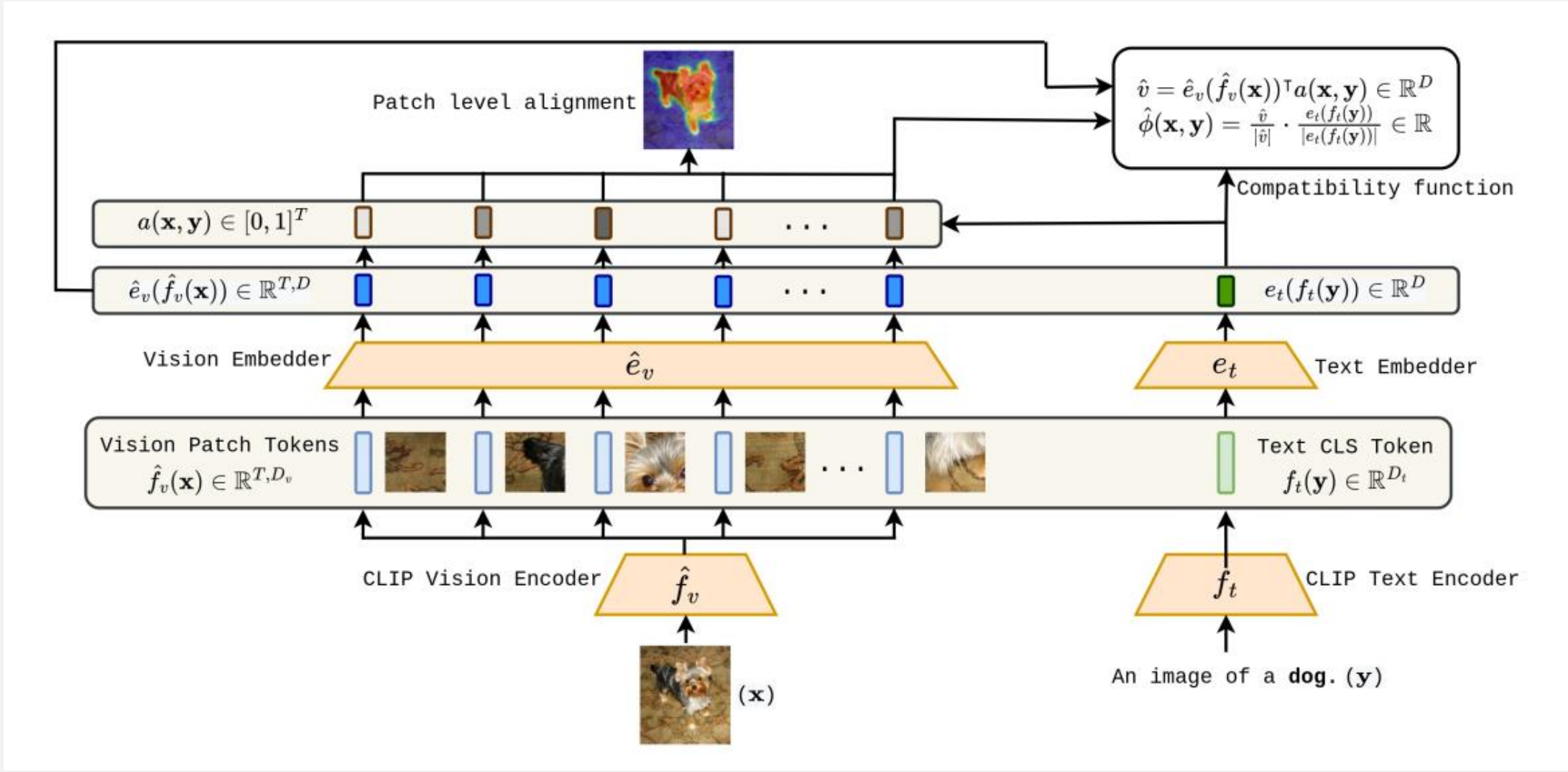
(a) Encode images & texts into latent space

(b) The U-ViT backbone of the joint noise prediction network

Figure 4. **Implementation of UniDiffuser on image-text data.** (a) First, we encode images and texts into latent space. (b) Second, we train UniDiffuser parameterized by a transformer (Bao et al., 2022a) in the way illustrated in Figure 2 on the latent embeddings.

Part.03

Open Vocabulary Semantic Segmentation with Patch Aligned Contrastive Learning

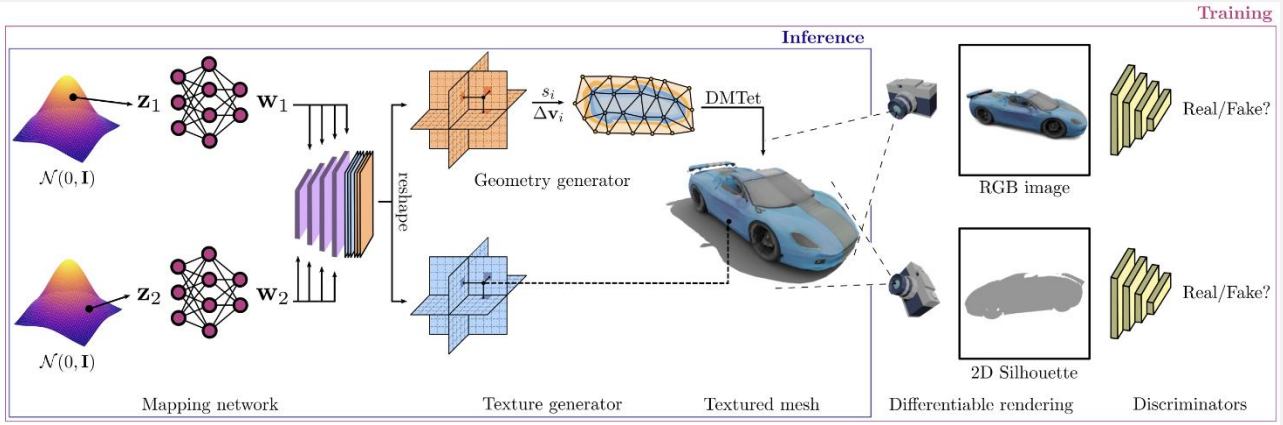

Patch level alignment
$\hat{v} = \hat{e}_v(\hat{f}_v(\mathbf{x}))^\intercal a(\mathbf{x},\mathbf{y}) \in \mathbb{R}^D$
$\hat{\phi}(\mathbf{x},\mathbf{y}) = \frac{\hat{v}}{|\hat{v}|} \cdot \frac{e_t(f_t(\mathbf{y}))}{|e_t(f_t(\mathbf{y}))|} \in \mathbb{R}$
Compatibility function
$a(\mathbf{x},\mathbf{y}) \in [0,1]^T$
$\hat{e}_v(\hat{f}_v(\mathbf{x})) \in \mathbb{R}^{T,D}$
$e_t(f_t(\mathbf{y})) \in \mathbb{R}^D$
Vision Embedder
$\hat{e}_v$
e_t
Text Embedder
Vision Patch Tokens
$\hat{f}_v(\mathbf{x}) \in \mathbb{R}^{T,D_v}$
Text CLS Token
$f_t(\mathbf{y}) \in \mathbb{R}^{D_t}$
CLIP Vision Encoder
$\hat{f}_v$
$\hat{f}_t$
CLIP Text Encoder
$(\mathbf{x})$
An image of a **dog**. $(\mathbf{y})$

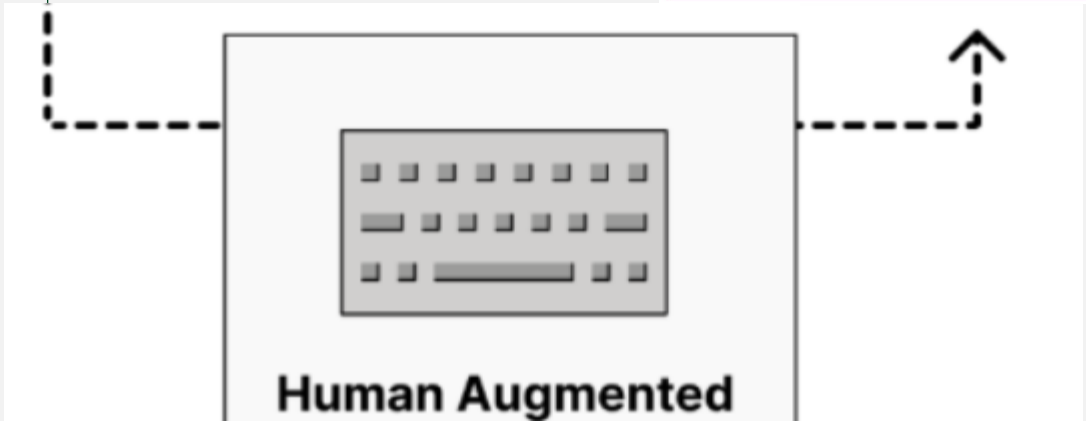
Part.04

RLHF

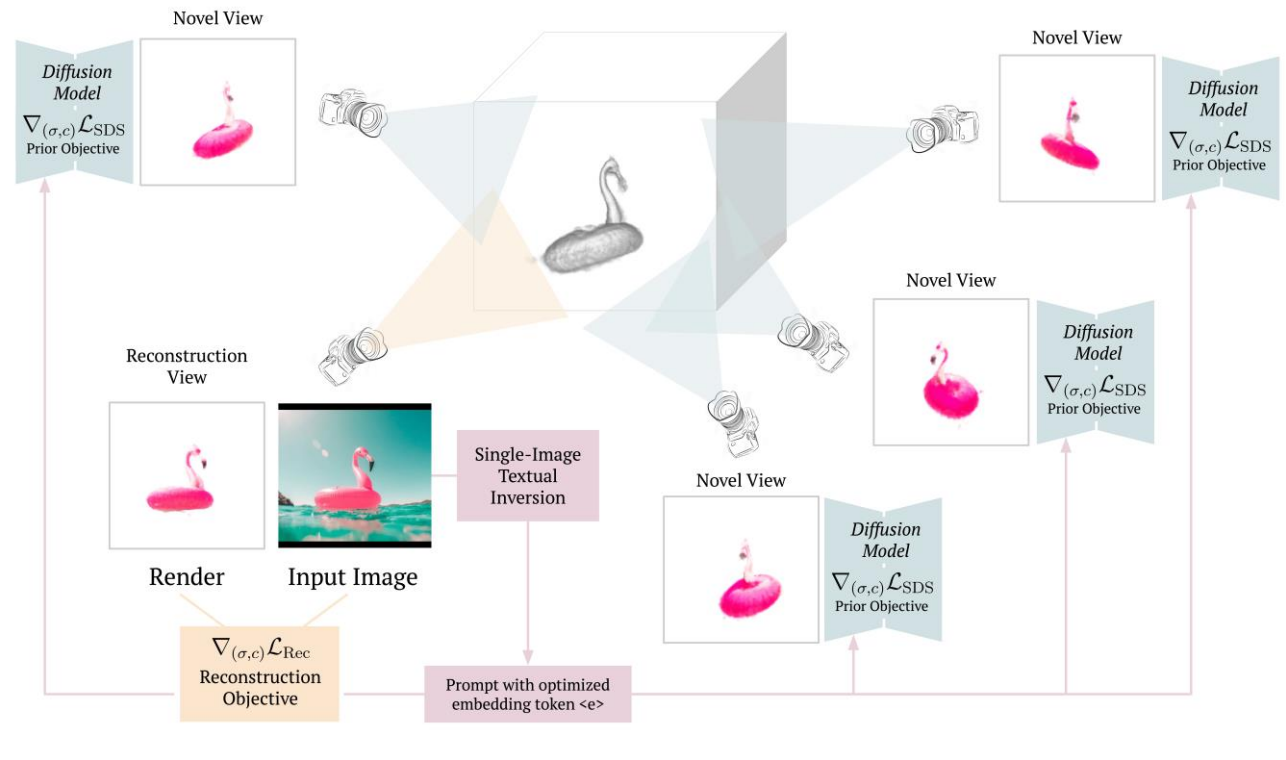
预训练3D模型

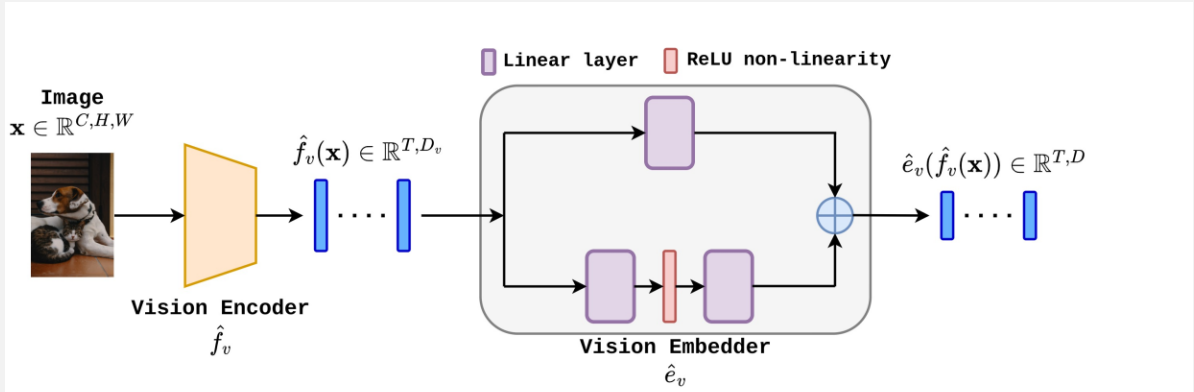
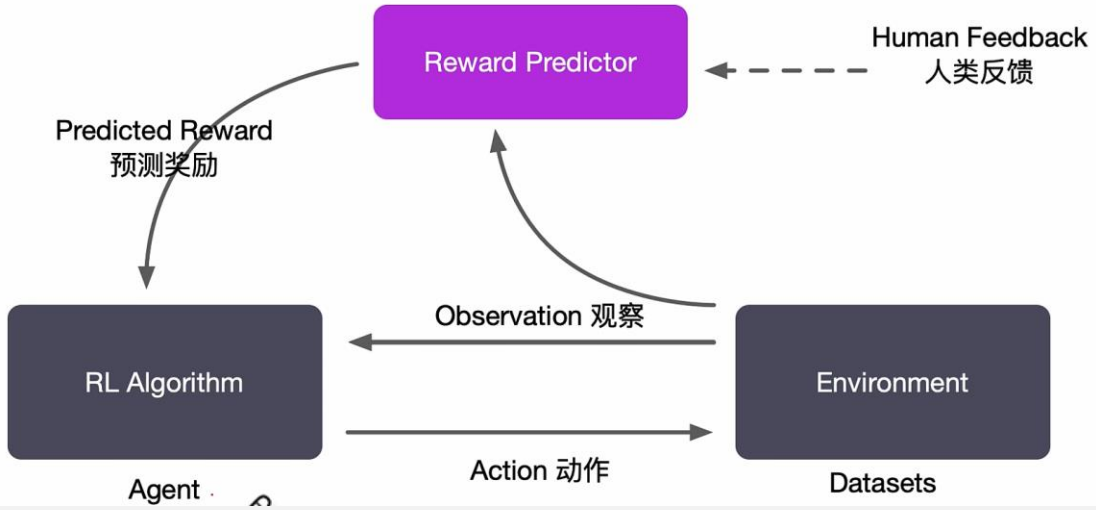
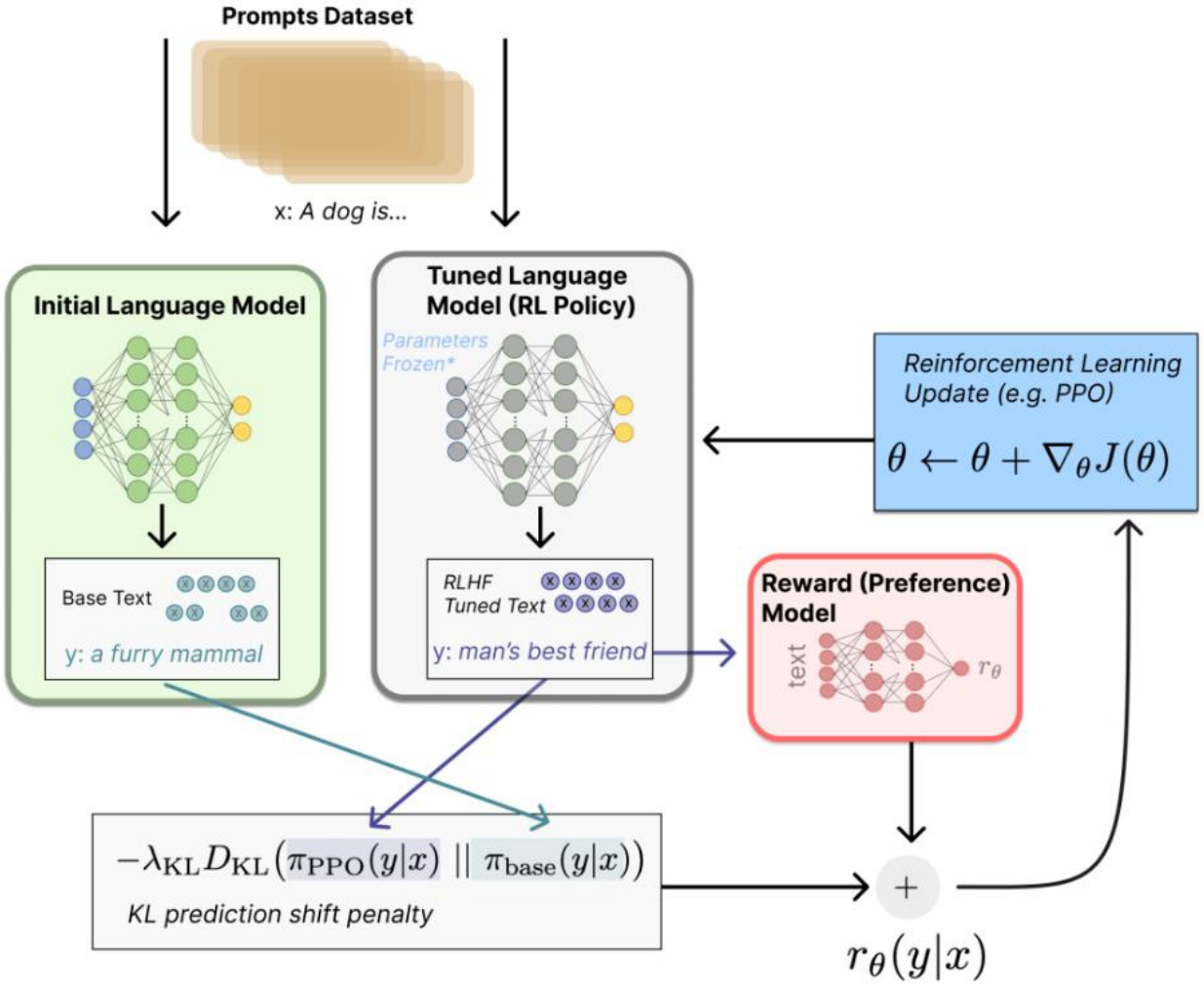
Omni
3D
Object


Training
Inference
$\mathbf{z}_1$
$\mathcal{N}(0,\mathbf{I})$
$\mathbf{w}_1$
$\mathbf{z}_2$
$\mathcal{N}(0,\mathbf{I})$
$\mathbf{w}_2$
reshape
s_i
$\Delta\mathbf{v}_i$
DMTet
Geometry generator
RGB image
Real/Fake?
2D Silhouette
Real/Fake?
Mapping network
Texture generator
Textured mesh
Differentiable rendering
Discriminators


Human Augmented

奖励模型的训练


Novel View
Diffusion Model
$\nabla_{(\sigma,c)}\mathcal{L}_{\text{SDS}}$
Prior Objective
Novel View
Diffusion Model
$\nabla_{(\sigma,c)}\mathcal{L}_{\text{SDS}}$
Prior Objective
Novel View
Diffusion Model
$\nabla_{(\sigma,c)}\mathcal{L}_{\text{SDS}}$
Prior Objective
Novel View
Diffusion Model
$\nabla_{(\sigma,c)}\mathcal{L}_{\text{SDS}}$
Prior Objective
Reconstruction View
Render
Input Image
Single-Image Textual Inversion
$\nabla_{(\sigma,c)}\mathcal{L}_{\text{Rec}}$
Reconstruction Objective
Prompt with optimized embedding token <e>


Linear layer
ReLU non-linearity
Image
$\mathbf{x} \in \mathbb{R}^{C,H,W}$
$\hat{f}_v(\mathbf{x}) \in \mathbb{R}^{T,D_v}$
$\hat{e}_v(\hat{f}_v(\mathbf{x})) \in \mathbb{R}^{T,D}$
Vision Encoder
$\hat{f}_v$
Vision Embedder
$\hat{e}_v$

基于 RL 进行语言模型优化