VTP: Towards Scalable Pre-training of Visual Tokenizers for Generation

Table of Contents

Background: 预训练扩展性问题

问题背景

Latent Diffusion Models (LDMs) 使用 visual tokenizer（如 VAE）将图像压缩到潜空间。Tokenizer 通常通过重建目标单独预训练。

核心悖论 (Scaling Paradox)

- 更好的重建 ≠ 更好的生成，重建与生成存在明显的 trade-off
- 对于 VAE，扩大预训练计算量虽改善重建，却可能损害生成

根本原因

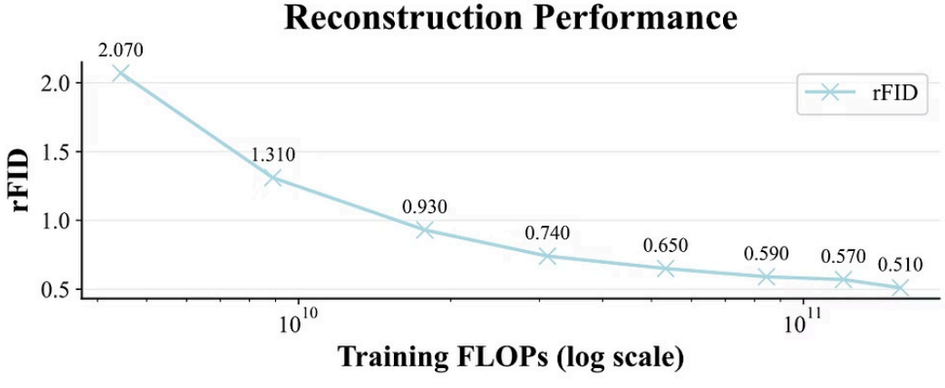
- 重建目标偏向低级信息（像素细节）
- 随训练扩大，VAE Latent 偏离生成所需的结构化语义空间
- **→ 这就是「预训练扩展性问题」(Pre-training Scaling Problem)**

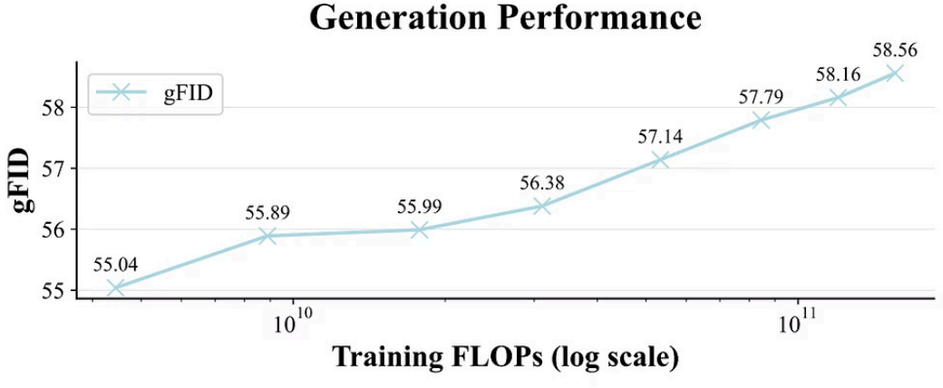
FID 指标计算

- 使用预训练的 **Inception-v3 网络**（去掉最后分类层），将输入图像映射为 **2048 维特征向量**（来自网络最后一层池化层的输出）
- FID 本质是两个高斯分布之间的 Fréchet 距离，公式为：

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

- 扩大训练计算量：rFID 2.0 → 0.5（重建持续改善）
- 但 gFID 55.04 → 58.56（生成反而恶化！）
- **重建任务本身对下游生成不具备可扩展性**





现有方案及其局限性

从「语义」的角度，DiT + VAE 三类已有的方案

1. DiT 侧语义对齐（如 REPA）

- 用辅助 loss，对齐 DiT 内部表征与 VFM 特征空间
- 局限：在 DiT 侧对齐，并未直接优化 AE 隐空间

2. VAE 侧语义对齐（如 VA-VAE）

- 在 AE 训练阶段引入监督信号，增强隐空间判别性
- 将 VAE 与视觉基础模型的语义特征进行对齐
- 缺点：依赖于视觉基础模型，能力受限

3. 直接用 VFM 表征 + Reconstruction（如 RAE）

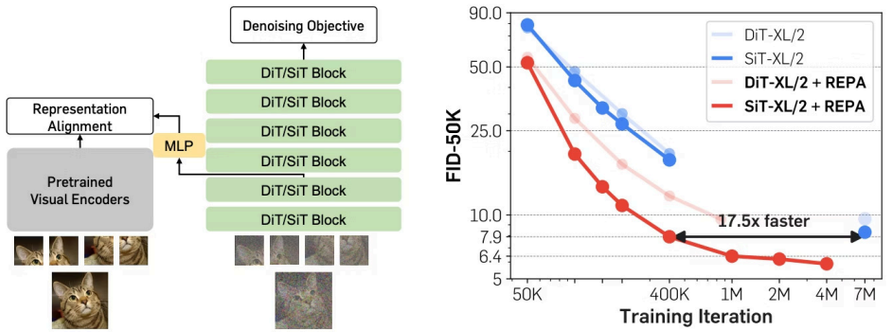
- 赋予 VFM 重建能力，继承其语义先验
- 比如 RAE: 直接使用 DINOv2 特征，训练额外 pixel decoder 做重建
- 局限：受限于已有基础模型，性能天花板低或重建损失大

现有方案及其局限性

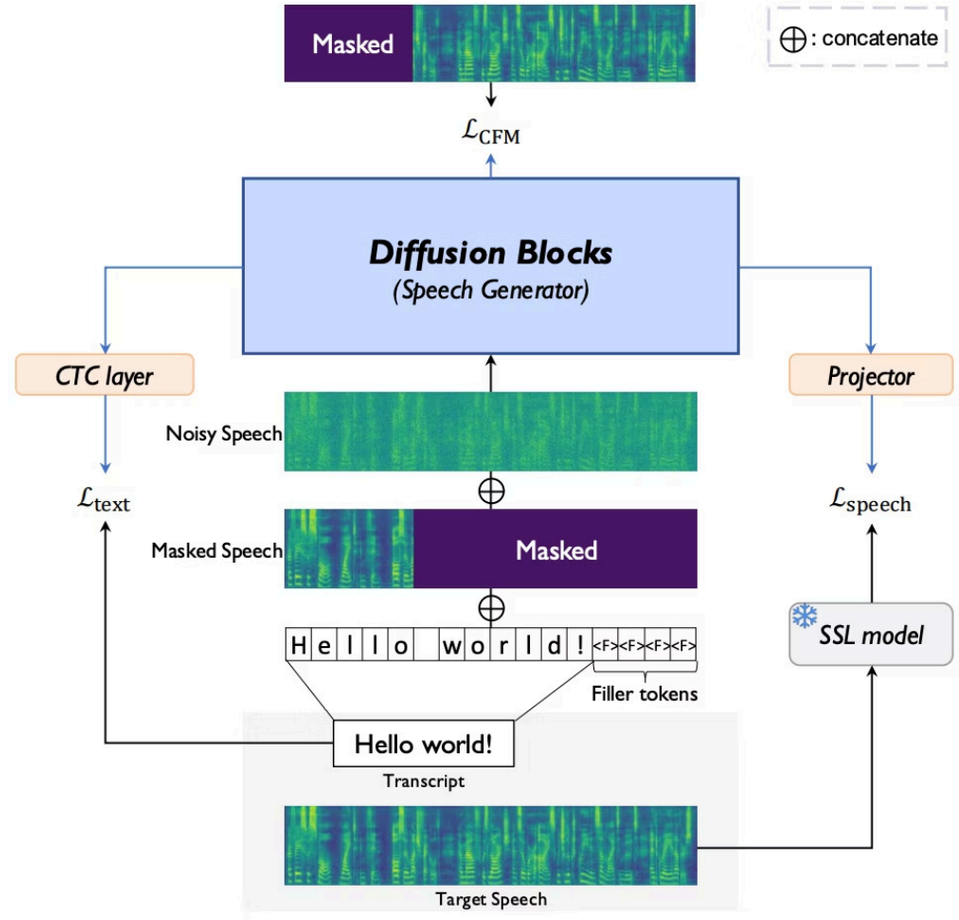
从「语义」的角度，DiT + VAE 三类已有的方案

1. DiT 侧语义对齐（如 REPA）

- 用辅助 loss，对齐 DiT 内部表征与 VFM 特征空间
- 作用：加速模型的收敛速度，原论文达到 17.5x 收敛速度
- 局限：在 DiT 侧对齐，并未直接优化 AE 隐空间



REPA 论文：https://arxiv.org/pdf/2410.06940



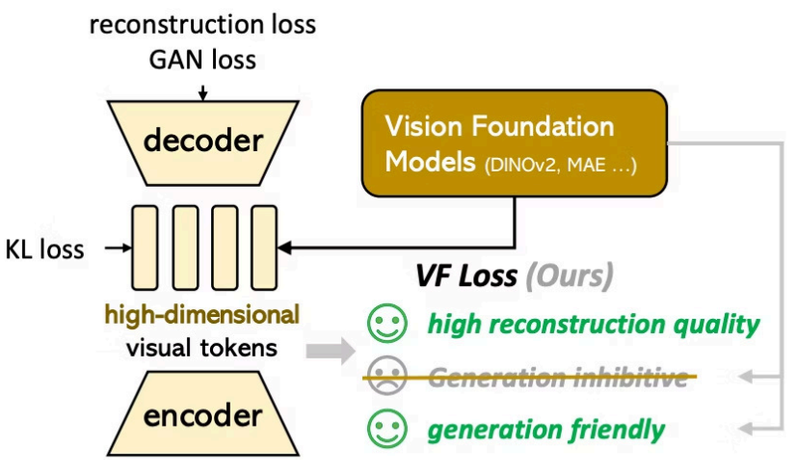
参考论文：https://arxiv.org/pdf/2505.19595

现有方案及其局限性

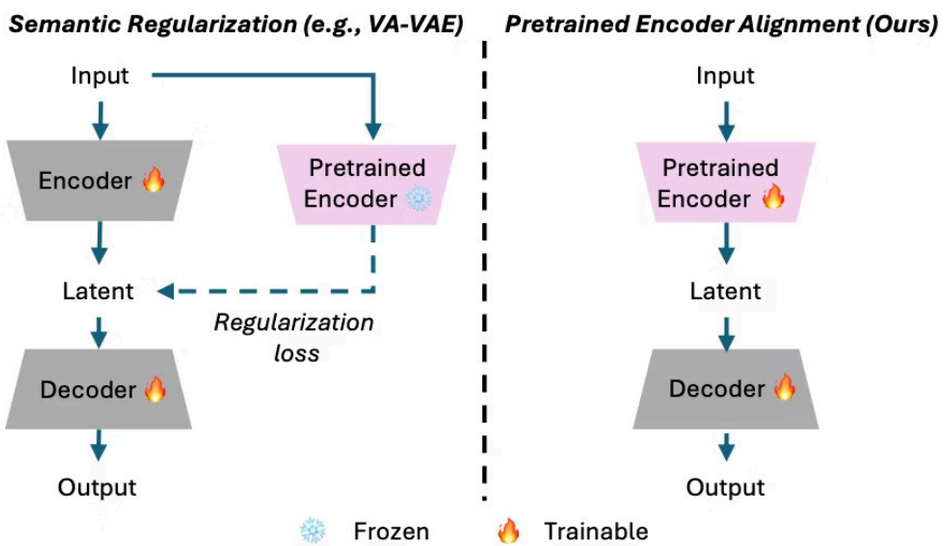
从「语义」的角度，DiT + VAE 三类已有的方案

2. VAE 侧语义对齐（如 VA-VAE, Semantic-VAE）

- 在 AE 训练阶段引入监督信号，增强隐空间判别性
- 将 VAE 与视觉基础模型的语义特征进行对齐
- 缺点：依赖于视觉基础模型，能力受限



论文：https://arxiv.org/pdf/2501.01423



论文：https://arxiv.org/pdf/2509.25162

现有方案及其局限性

从「语义」的角度，DiT + VAE 三类已有的方案

2. VAE 侧语义对齐（如 VA-VAE, Semantic-VAE）

- 在 AE 训练阶段引入监督信号，增强隐空间判别性
- 将 VAE 与视觉基础模型的语义特征进行对齐
- 缺点：依赖于视觉基础模型，能力受限

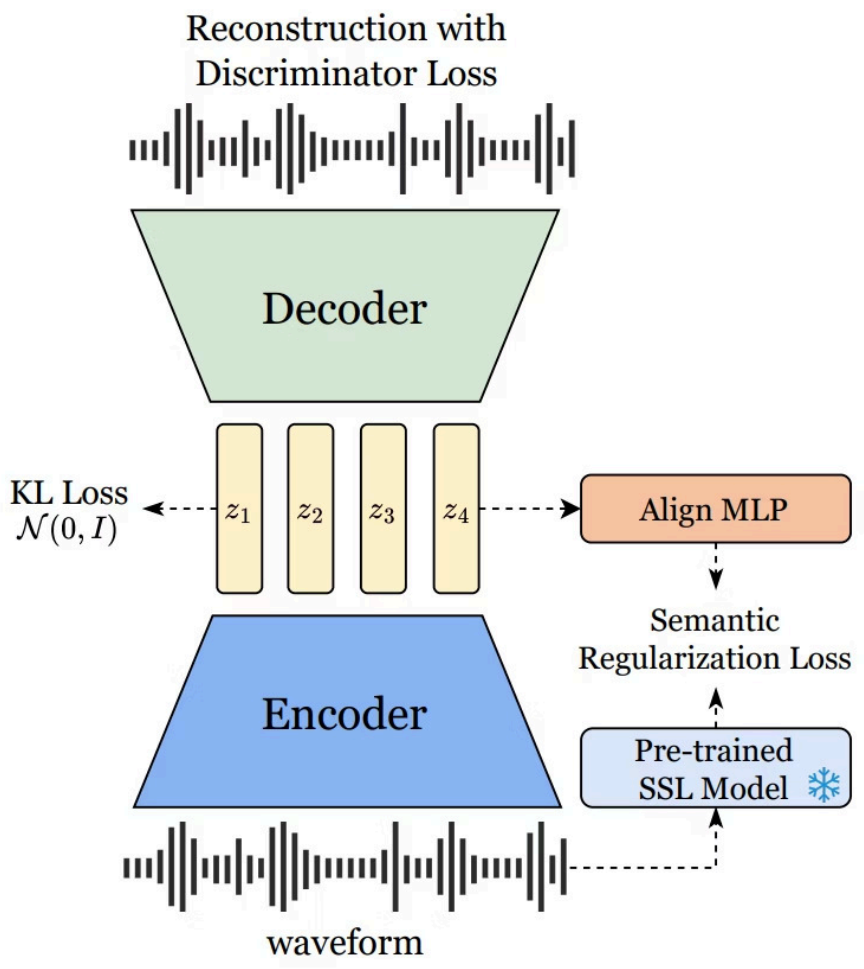


Table 1. Zero-shot TTS results on the LibriSpeech-PC test-clean. The best results in the low-resource setting are marked in **bold**. High-resource results are quoted from the F5-TTS for comparison.

Method	# Param.	Data	WER(%)↓	SIM↑
Ground Truth			2.23	0.69
High-resource				
CosyVoice [37]	300M	170kh	3.59	0.66
FireRedTTS [38]	580M	248kh	2.69	0.47
E2 TTS [15]	333M	100kh	2.95	0.69
F5-TTS [9]	336M	100kh	2.42	0.66
Low-resource				
USLM [25]	361M	0.6kh	6.11	0.43
E2 TTS [15]	157M	0.6kh	3.51	0.61
+ **Semantic-VAE**	157M	0.6kh	2.31	0.62
F5-TTS [9]	159M	0.6kh	2.23	0.60
+ **Vanilla VAE**	159M	0.6kh	2.65	0.60
+ **Semantic-VAE**	159M	0.6kh	**1.95**	**0.64**

Table 3. Effect of different SSL models, layer, and alignment methods on zero-shot TTS performance on LibriSpeech-PC test-clean. "Avg." = average of all layers; "Last" = final layer output.

Align method	Layer	WER (%)↓	SIM↑	UTMOS↑
Baseline	-	2.65	0.59	4.42
$\lvert\Phi_n - f(x)\rvert$	WavLM 23rd	3.12	0.47	3.29
$\lVert\Phi_n - f(x)\rVert_2$		4.37	0.48	4.16
$-\cos(\Phi_n, f(x))$		**2.10**	**0.64**	**4.39**
SSL Models	**Layer**	**WER (%)↓**	**SIM↑**	**UTMOS↑**
HuBERT	23	2.28	0.63	4.38
	Last	2.33	0.50	4.41
	Avg.	2.29	0.61	4.39
WavLM	23	**2.10**	**0.64**	4.39
	Last	2.62	0.58	4.34
	Avg.	2.31	0.63	**4.42**

现有方案及其局限性

从「语义」的角度，DiT + VAE 三类已有的方案

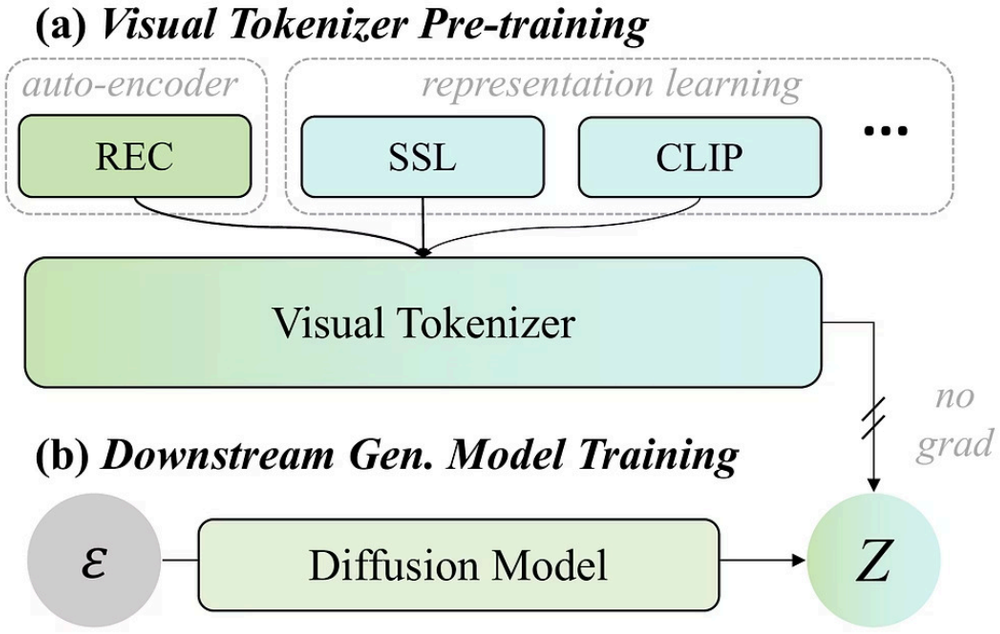
3. 直接用 VFM 表征 + Reconstruction（如 RAE）

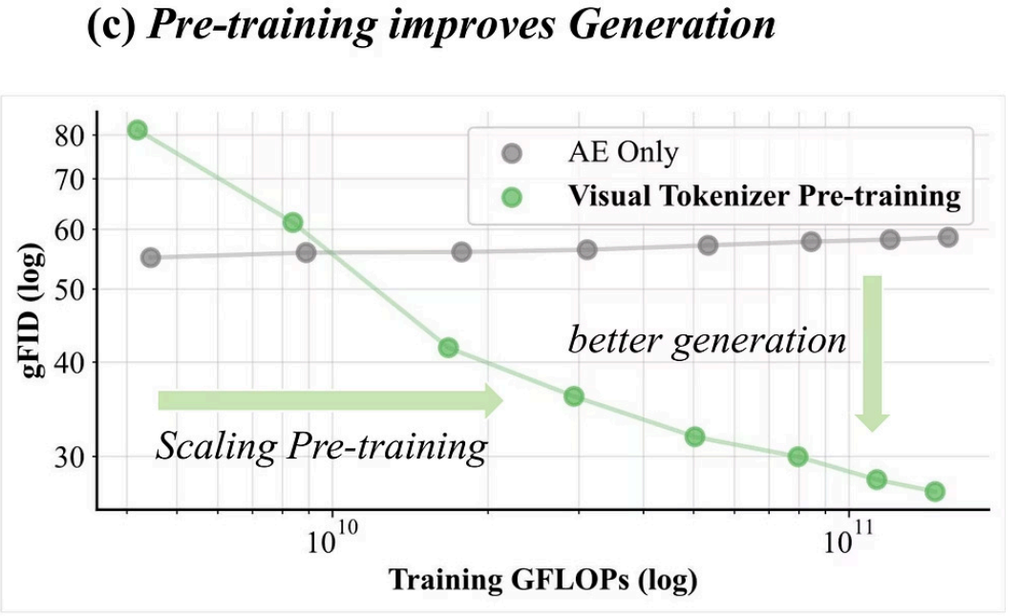
- 赋予 VFM 重建能力，继承其语义先验
- 比如 RAE: 直接使用 DINOv2 特征，训练额外 pixel decoder 做重建
- 局限：性能天花板低或重建损失大

VTP 的出发点

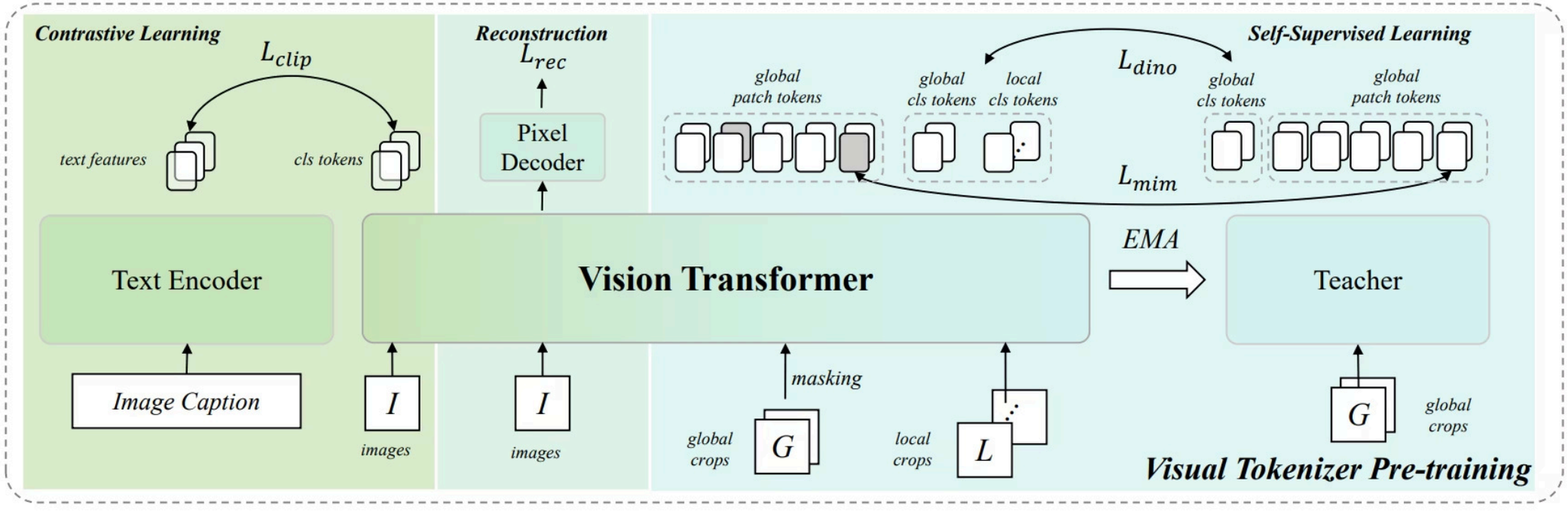
Tokenizer 采用 VFM 的预训练任务 + 重建任务

- 不需要依赖任何预训练的 VFM，同时 Encoder 也会受到重建任务的监督，整体天花板会更高
- 统一的 tokenizer 预训练框架，联合优化 CLIP + SSL + Reconstruction，首次证明 tokenizer 预训练的可扩展性





VTP 框架总览



VTP 在 ViT-based Auto-Encoder 上整合四种学习目标：

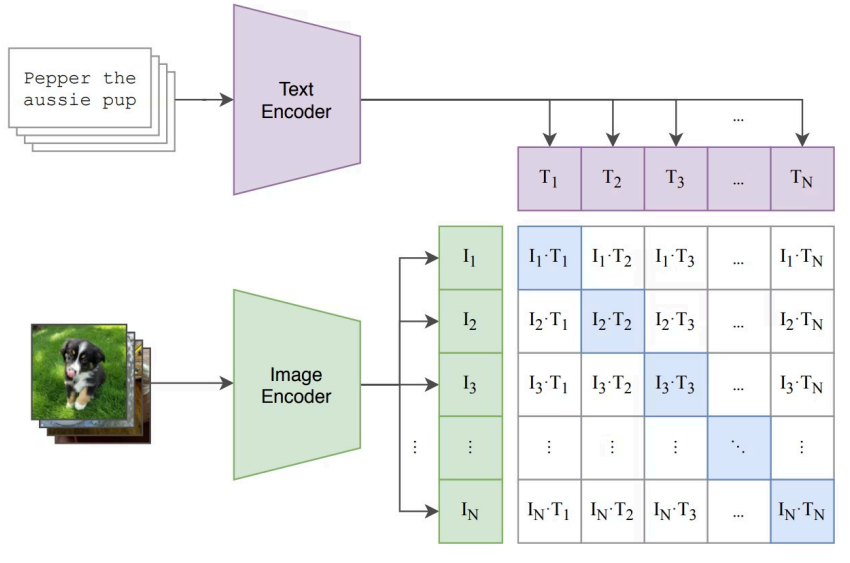
- Cross-modal Alignment: Image-text contrastive learning (CLIP) → 全局语义对齐
- Pixel Reconstruction: L1 + Perceptual loss → 保持细粒度视觉细节
- SSL-任务A：Self-distillation (DINO): Global / local view 一致性 → 全局语义理解
- SSL-任务B：Masked Image Modeling (MIM): 补全被遮蔽的 patch → 局部空间感知

训练目标：对比学习 & 重建任务

1. Contrastive Learning (CLIP)

给定 image-text pairs，visual tokenizer 编码图像，text encoder 编码文本：

- 最大化正样本对（matched image-text）的相似度
- 最小化负样本对的相似度
- 使潜空间具备跨模态语义对齐能力



Clip 示意图

2. Visual Reconstruction

图像 I ∈ R^(3×H×W) → 潜空间 R^(d×H/16×W/16) → pixel decoder 重建 I'

- 两阶段训练：预训练用 L1 + perceptual loss
- 微调时冻结 Encoder，GAN loss 提升保真度
- GAN loss 与 ViT 兼容性差（梯度大、不稳定），故分阶段处理

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_1 + \mathcal{L}_{\text{perceptual}}$$

Batch Sampling 策略

不同任务最优 batch size 差异大：

- CLIP: 极大 batch (16k~32k)
- SSL: 中等 batch (4k)
- Reconstruction: 较小 batch (2k)

解决方案：

给定 B 个 image-caption pairs：

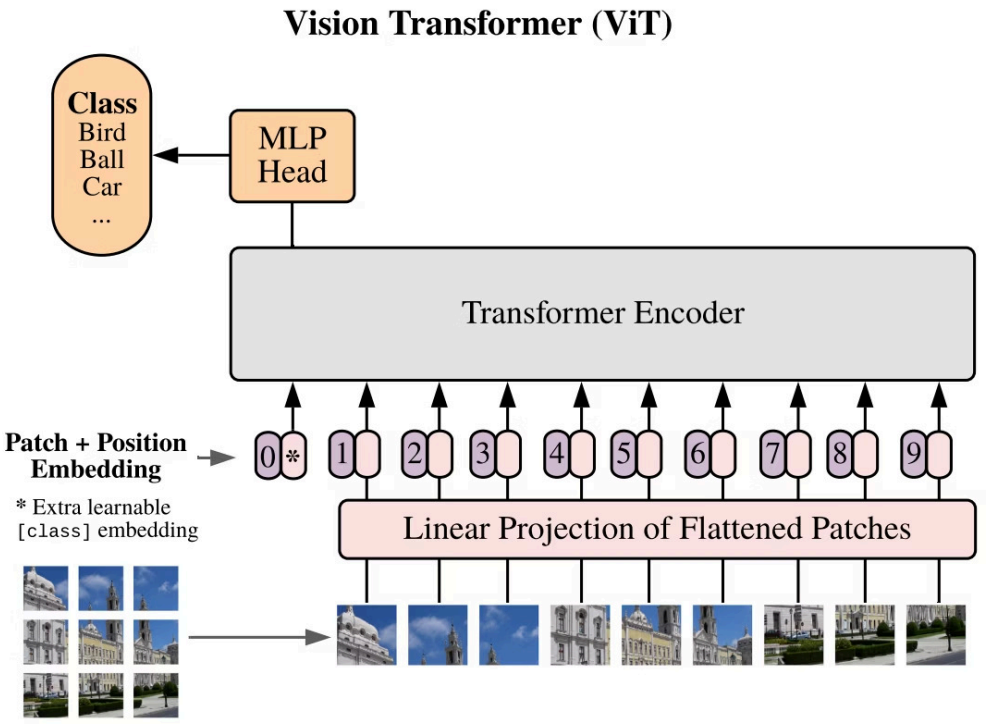
- B_clip = B（全部用于 CLIP）
- B_ssl 从 B 中随机采样
- B_rec 从 B 中随机采样

训练目标：DINO-v2 自监督学习

3.1 SSL 任务 A: MIM

patch-level 任务（MIM Loss 又称为 iBOT Loss）

- 学生网络（Student）：对输入的 global crop 随机遮掩一部分 patch，只看图片的局部
- 老师网络（EMA Teacher）：输入完整的、未被遮掩的原图，看到所有 patch
- 每个网络共享一个 iBOT MLP head，将 VIT 每个 patch 映射到「虚拟类别个数」的维度
- 找到学生被遮掩的 patch 位置，与老师对应位置的 patch 特征做对齐
- 对学生和老师的输出分别做 softmax（转为概率分布）并中心化
- 用交叉熵损失让学生的预测分布尽可能接近老师的伪标签分布



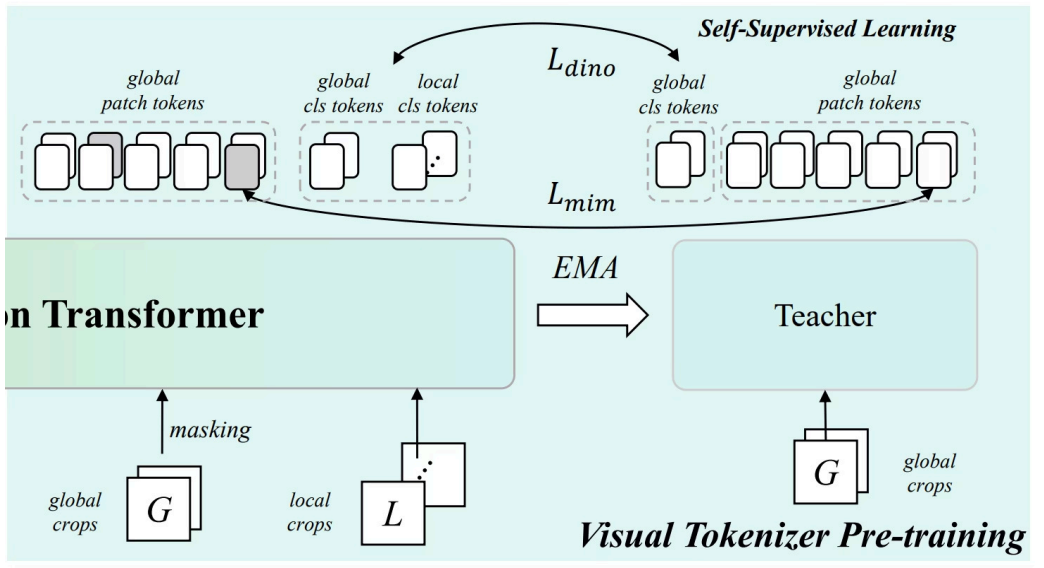
ViT 架构示意图

$$\mathcal{L}_{iBOT} = -\sum_i p_{ti} \log p_{si}$$

3.2 SSL 任务 B：DINO

image-level 任务

- 改为用学生和老师网络的 CLS token 计算损失
- 同一张图做不同裁剪（global/local），学生和老师分别提取特征
  - 注意：student 也会输入 local 特征来学习 teacher 的 global 特征
- 学生和老师各用一个 DINO MLP head 把特征映射为分布
- 分别经过 softmax 转换成概率分布
- 用交叉熵让学生分布去拟合老师分布



$$\mathcal{L}_{ssl} = \mathcal{L}_{mim} + \mathcal{L}_{dino}$$

$$\mathcal{L}_{DINO} = -\sum p_t \log p_s$$

总训练目标

Overall Objective

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{ssl}}\mathcal{L}_{\text{ssl}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}}$$

- λ_rec = 0.1, λ_clip ∈ {0, 1}, λ_ssl ∈ {0, 1}
- **较小的重建权重，更有利于生成性能**

可以认为是：从头训练了一个**具有重建能**的 DINO-v2 模型

ViT Auto-Encoder 的重建能力

重建能力到底怎么样?

ViT 作为 Visual Tokenizer

构建对称 encoder-decoder 的 ViT visual tokenizer (下采样 16 倍，64 维 latent)

Arch.	FLOPs	#Params	*r*PSNR	*g*FID
CNN (Rombach et al., 2022)	389.4G	70.3M	30.63	59.53
ViT-B (Dosovitskiy, 2020)	87.7G	171.2M	30.72	58.40
ViT-L (Dosovitskiy, 2020)	311.1G	607.2M	**31.28**	**53.51**

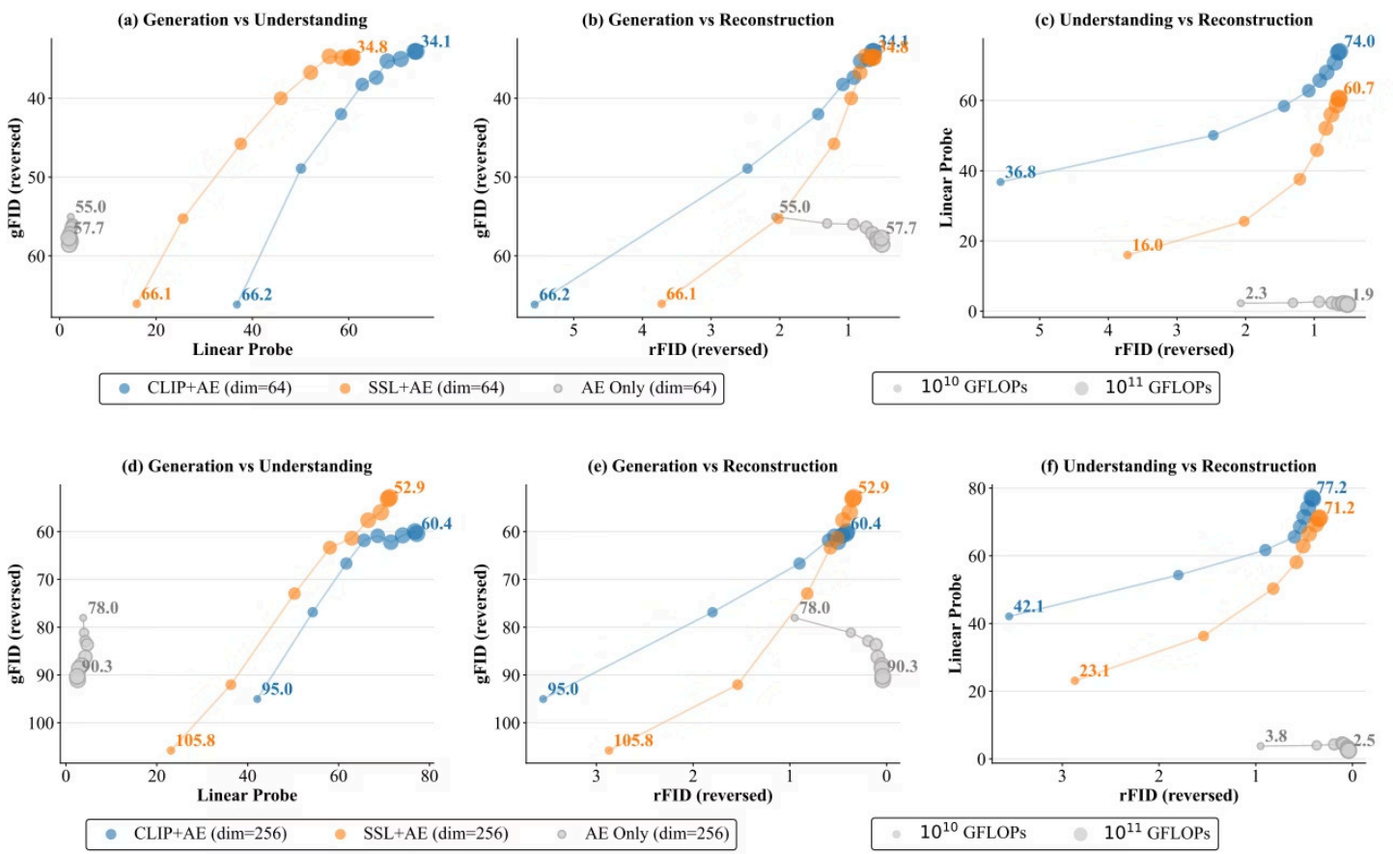
与传统 CNN-based Tokenizer 对比：

- ViT-L: 重建 PSNR=31.28, gFID=53.51，性能与 CNN 持平, ViT 也可以是有效的 tokenizer 架构

Scaling 实验一：语义任务的作用

实验配置：

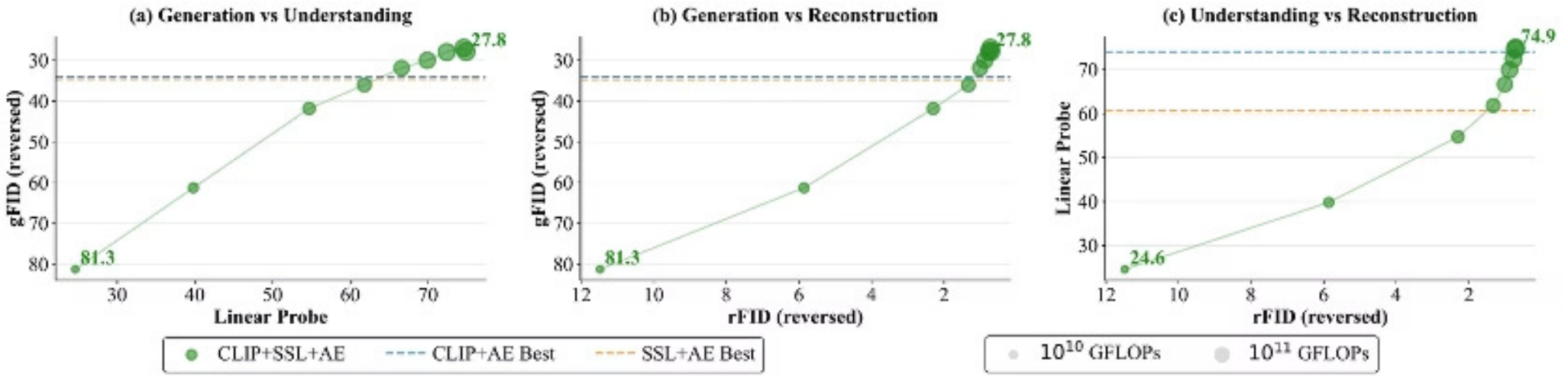
- 2.77 亿数据集，相同的 ViT 结构。下游任务是基于 DiT 的 Image 类别生成任务
- 重建任务 AE，CLIP/SSL 表示语义任务



四个关键结论：

- (c-f) 多任务混合训练：表征学习 + 重建同步进行，理解与重建指标稳步提升，相比于 AE，未出现明显饱和现象
- (b-e) 重建悖论：随计算量增加，重建改善但生成恶化（AE-only 曲线）
- (a-d) **理解是关键驱动力**：引入语义理解任务后，生成性能随 FLOPs 持续改善
- CLIP 与 SSL 虽范式不同，但都能通过丰富语义来提升生成 → 表征学习的通用性

Scaling 实验一：语义任务的作用



联合 Contrastive + Self-Supervised + Reconstruction 进行 tokenizer 预训练

- 训练稳定可行，多目标框架使 tokenizer 捕获多尺度特征
- f16d64 配置：gFID 达到 27.8，理解性能 74.9%，显著超越双目标 (CLIP+AE / SSL+AE)
- **多种表征学习范式可无缝集成 → 未来更多新范式可以进一步提升性能上限**

Scaling 实验二：参数量 Scaling

Encoder 扩展

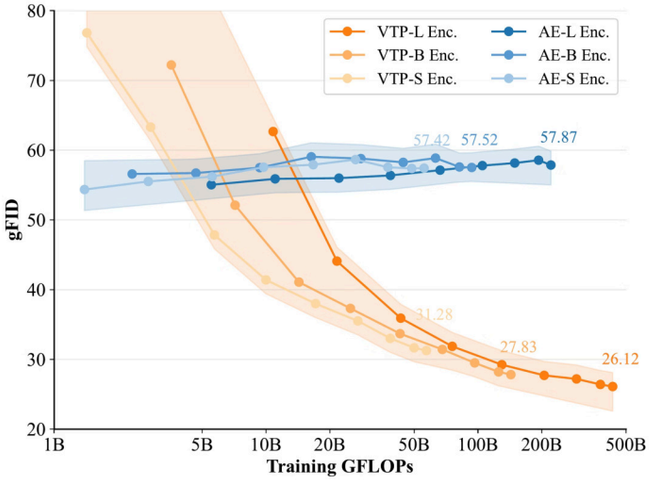
- AE baseline: gFID 在 ~57 停滞，不随模型参数量扩大而增长
- VTP: gFID 从 31.28 → 26.12（20M → 300M 参数）
- 形成很明确的参数量 scaling curve

Decoder 扩展

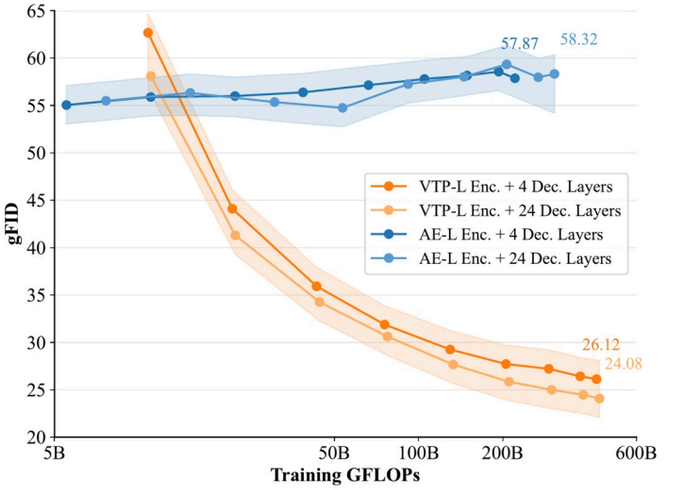
- 扩大 pixel decoder 同样能带来生成的改善
- gFID 从 26.12 → 24.08

VTP 是首个展现参数扩展性的 visual tokenizer

→ 传统普通 AE 的性能天花板被彻底打破



(a) Encoder Scaling



(b) Decoder Scaling

Scaling 实验三：数据量 Scaling

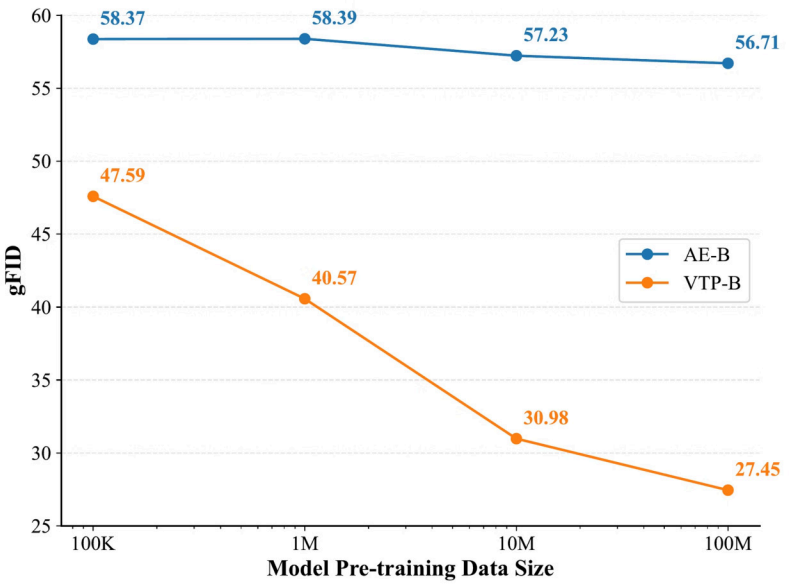
实验设置

- 从 DataComp-1B 随机采样 100K / 1M / 10M / 100M 四个子集
- 训练 VTP-ViT-L 和 AE-ViT-L
- 保持下游 DiT 结构/参数量完全相同，只观察 Tokenizer 的影响

关键结果

- VTP 在所有数据规模上均优于传统 AE
- AE: FID 仅从 58.37 → 56.71（数据量的增加几乎无改善）
- VTP: FID 从 47.59 → 27.45（数据越多越好，效果一直稳定）

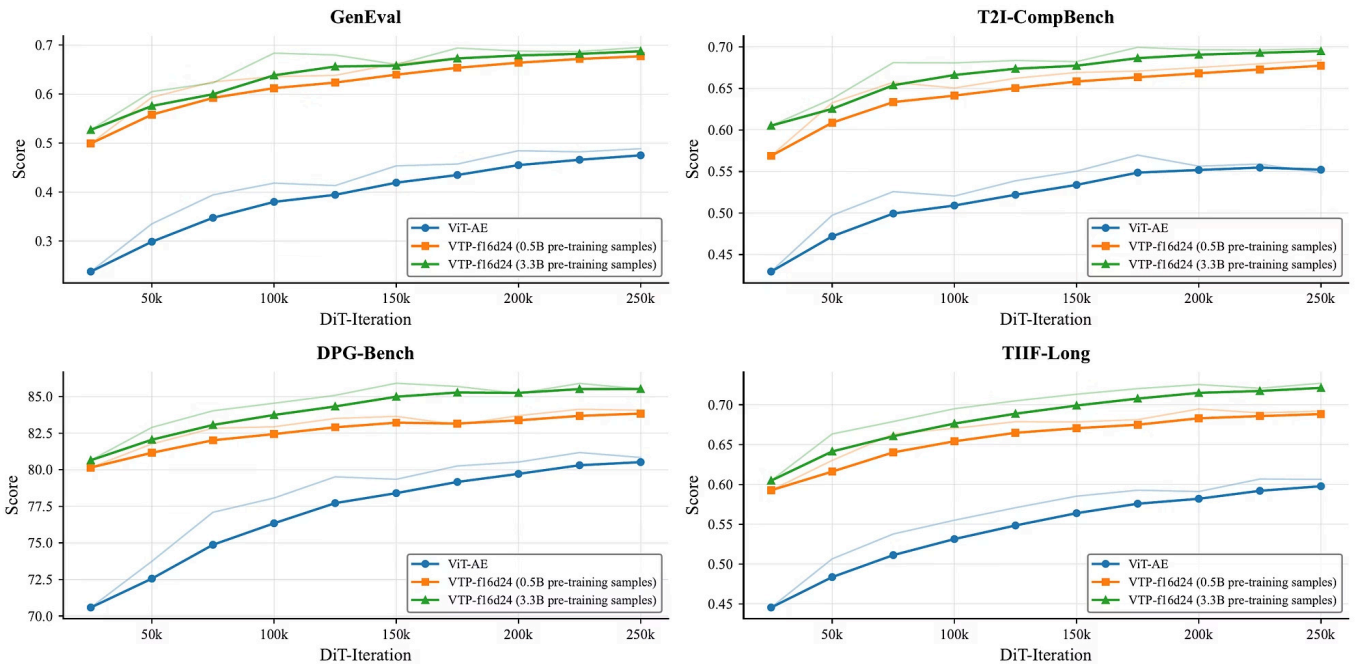
验证了 tokenizer 对于生成任务的数据 scaling 能力



Scaling 实验四：文生图 T2I 任务

实验设置

在 LAION 数据集上训练 VTP tokenizer，下游使用 DiT-XL 做 Text-to-Image 生成



关键发现

- 带表征学习目标的 tokenizer 收敛速度远快于 AE baseline
- 语义感知的预训练大幅提升下游 T2I 训练效率
- 随 tokenizer 预训练计算量增加，下游生成持续改善
- VTP 的 **scaling** 性质从 ImageNet 类别生成，推广到更具挑战的 T2I 场景

T2I: CLIP Loss 改善文本渲染

不同语义损失对 T2I 的影响

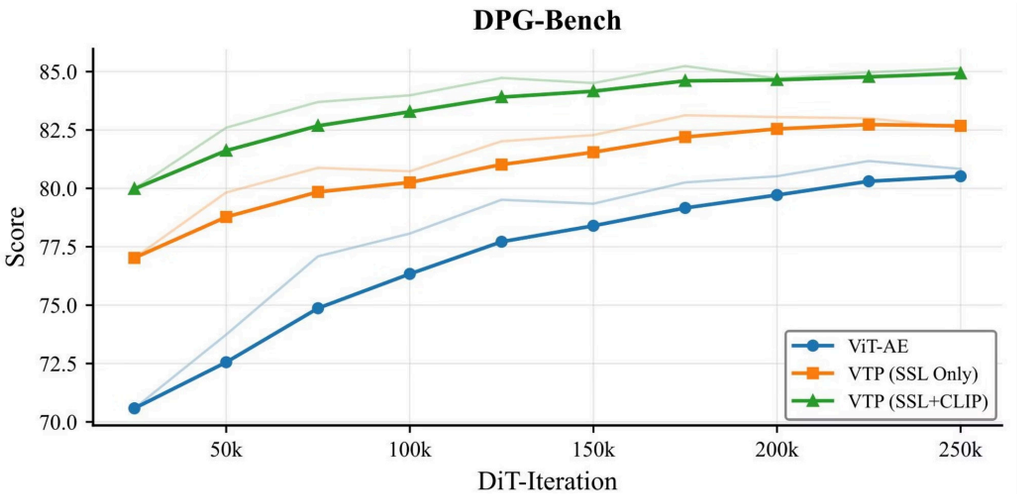
- 与 ImageNet 上类似，VTP 预训练融入的感知损失越多，T2I 生成性能越好

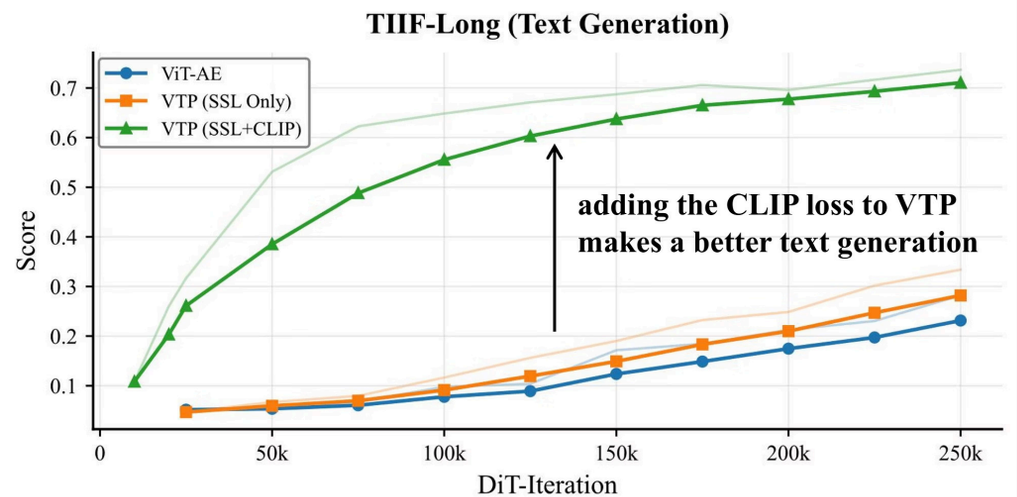
CLIP Loss 的独特优势

- 带 CLIP loss 的 tokenizer 在文本渲染能力上有显著优势
- 大幅超越 AE-only 和 SSL+AE tokenizer
- CLIP 的跨模态对齐使 latent 更好地编码了文本语义信息

渐进增加目标的效果

- AE < SSL+AE < CLIP+SSL+AE
- 每增加一种感知目标，T2I 质量稳步提升





VTP vs RAE: 扩展性对比

RAE 的方案与局限

RAE 直接使用预训练 DINOv2 特征做生成，训练独立 pixel decoder 做重建。受限于固定基础模型，扩展性受限

Scalability 对比（相同 DiT 训练配置，80 epochs w/o guidance）

Tokenizer	RAE	VTP
Small	3.50	5.46
Base	4.28	3.88
Large	6.09	**2.81**

Table 3 | **Scalability comparison.** LightningDiT gFID↓ at 80 epochs (w/o guidance) with identical DiT training. We take RAE's results from its original paper.

VTP 的核心优势

- VTP 始终包含重建目标 → 保留细粒度细节
- RAE 依赖额外训练 pixel decoder → 信息损失大
- **VTP 展现正向 scaling，RAE 在大模型时性能退化**

VTP vs RAE: 重建与生成质量可视化

(a) 重建对比：VTP 比 RAE 保留更多细粒度细节

(b) 生成对比：VTP 语义更准确

最终性能：ImageNet 256×256 Generation

VTP 核心指标

- Reconstruction: rFID = 0.36
- Understanding:
  - Zero-shot Acc = 78.2%, Linear Probe = 85.7%

Convergence (80 epochs, w/o guidance)

- VTP gFID = 2.62，大幅超越 REPA (7.90), RAE (4.28)
- VTP-1B: gFID = 2.03，超越 VA-VAE (4.29)

Long Period Training (w/ guidance)

- VTP: gFID = 1.11 (SOTA)，600 epochs
- 4.1× 更快收敛：80 epochs 达到 2.03，无需 guidance tricks

只通过 Visual Tokenizer 的升级，实现 SOTA 结果

Method	Tokenizer			Gen Model Params	Epochs	Generation w/o guidance				Generation w/ guidance			
	rFID↓	Zero-Shot↑	Linear Probe↑			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Perception or Unified Tokenizer Baselines													
SigLIP (Zhai et al., 2023)	-	80.5	-	-	-	-	-	-	-	-	-	-	-
MAE (He et al., 2022)	-	-	85.9	-	-	-	-	-	-	-	-	-	-
DINOv2 (Oquab et al., 2023)	-	-	86.7	-	-	-	-	-	-	-	-	-	-
VILA-U (Wu et al., 2024)	1.80	73.3	-	-	-	-	-	-	-	-	-	-	-
UniTok (Ma et al., 2025)	0.41	70.8	-	1.4B	-	2.51	216.7	0.82	0.57	2.77	227.5	0.81	0.57
Convergence Efficiency													
REPA (Yu et al., 2024)	0.61	-	-	675M	80	7.90	122.6	0.70	0.65	-	-	-	-
DDT (Wang et al., 2025)	0.61	-	-	675M	80	6.62	135.2	0.69	**0.67**	1.52	263.7	0.78	0.63
VA-VAE (Yao et al., 2025)	**0.28**	-	-	675M	80	4.29	-	-	-	-	-	-	-
REPA-E (Leng et al., 2025)	**0.28**	-	-	675M	80	3.46	159.8	0.77	0.63	1.67	**266.3**	**0.80**	**0.63**
RAE (Zheng et al., 2025)	0.57	-	84.5	675M	80	4.28	-	-	-	-	-	-	-
RAE (Zheng et al., 2025)	0.57	-	84.5	835M	80	2.16	214.8	0.82	0.59	-	-	-	-
VTP (Ours)	0.36	**78.2**	**85.7**	675M	80	**2.62**	**197.8**	**0.79**	0.62	**1.44**	238.2	**0.80**	**0.63**
VTP (Ours)	0.36	78.2	85.7	1.0B	80	2.03	219.4	0.80	0.62	-	-	-	-
Long Period Training													
DiT (Peebles and Xie, 2023)	-	-	-	675M	1400	9.62	121.5	0.67	0.67	2.27	278.2	**0.83**	0.57
SiT (Ma et al., 2024)	-	-	-	675M	1400	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
VA-VAE (Yao et al., 2025)	**0.28**	-	-	675M	800	2.17	205.6	0.77	0.65	1.35	295.3	0.79	0.65
REPA (Yu et al., 2024)	0.61	-	-	675M	800	5.78	158.3	0.70	0.68	1.29	306.3	0.79	0.64
DDT (Wang et al., 2025)	0.61	-	-	675M	400	6.27	154.7	0.68	**0.69**	1.26	**310.6**	0.79	0.65
REPA-E (Leng et al., 2025)	**0.28**	-	-	675M	800	**1.70**	217.3	0.77	0.66	1.15	304.0	0.79	0.66
RAE (Zheng et al., 2025)	0.57	-	84.5	676M	800	1.87	209.7	**0.80**	0.63	1.41	309.4	0.80	0.63
RAE* (Zheng et al., 2025)	0.57	-	84.5	839M	800	1.51	242.9	0.79	0.63	1.13	262.6	0.78	0.67
VTP (Ours)	0.36	**78.2**	**85.7**	675M	600	1.85	**232.3**	0.79	0.63	**1.11**	279.5	0.79	**0.67**

Summary & Conclusion

VTP 重新设计了 visual tokenizer 的可扩展预训练范式

核心发现

- 语义理解是生成质量的核心驱动力 (Understanding drives generation)
- 融入理解任务后，visual tokenizer 首次展现 parameter / data 的 scaling law

核心结果

- ImageNet 256×256: 1.11 gFID (w/ guidance, SOTA)
- 重建：0.36 rFID｜理解：78.2% zero-shot, 85.7% linear probe
- 收敛：80 epochs → 2.03 gFID, 4.1× 加速
- Scaling: 10× compute → 65.8% FID 改善

对比传统 AE：纯重建预训练在 1/10 计算量即饱和，性能天花板低；VTP 则持续受益于更大规模的预训练。

代码与模型已开源: github.com/MiniMax-AI/VTP

Takeaways: 技术贡献与可迁移方案

1 联合多目标预训练范式

- CLIP + SSL + Reconstruction 的联合优化，使 tokenizer 同时具备语义理解和像素重建能力
- 对于语音生成任务，**训练 AE/VAE latent 时增加 CLIP/SSL 的任务，可能比 Semantic-VAE 上限更高**

2 Tokenizer Scaling Law

- 证明 visual tokenizer 的预训练存在 scaling law
- DiT 下游任务不变，扩展 Tokenizer 训练的参数量/数据量，就能带来稳定性的提升

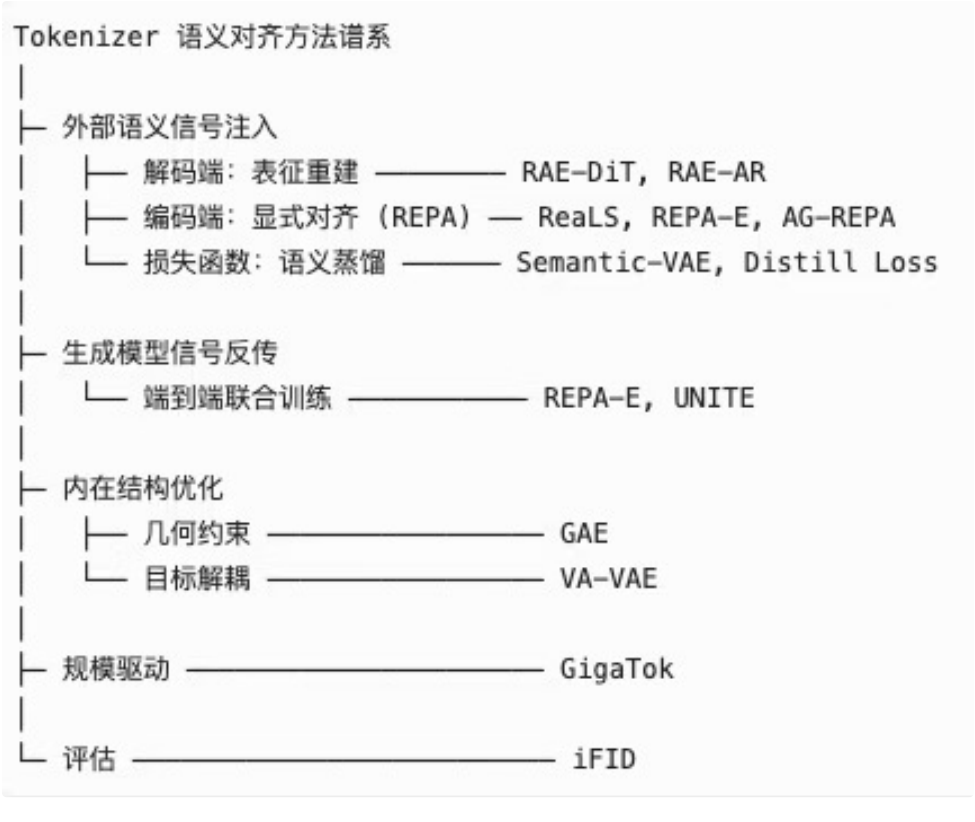
3 Batch Sampling 策略

针对不同任务最优 batch size 差异大的问题，提出差异化采样策略。对多任务联合训练有借鉴意义。

4 两阶段训练（解耦 GAN）

- ViT 结构是 Tokenizer 具备 scaling 能力的基础要求
- 先联合预训练（无 GAN），再冻结 tokenizer 微调 decoder（加 GAN），解决 GAN 与 ViT 兼容性问题

其他推荐论文



1. **REPA**: Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think — **2410.06940**
2. **VA-VAE**: Reconstruction vs. Generation: Taming Optimization Dilemma in Latent Diffusion Models — **2501.01423**
3. **RAE (RAE-DiT)**: Diffusion Transformers with Representation Autoencoders — **2510.11690**
4. **GAE**: Geometric Autoencoder for Diffusion Models — **2603.10365**
5. **RAE-AR**: Taming Autoregressive Models with Representation Autoencoders — **2604.01545**
6. **REPA-E**: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers — **2504.10483**
7. **UNITE**: End-to-End Training for Unified Tokenization and Latent Denoising — **2603.22283**
8. **GigaTok**: Scaling Visual Tokenizers to 3 Billion Parameters for Autoregressive Image Generation — **2504.08736**
9. **iFID**: Making Reconstruction FID Predictive of Diffusion Generation FID — **2603.05630**
10. **Semantic-VAE**: Semantic-Alignment Latent Representation for Better Speech Synthesis — **2509.22167**
11. **Distill Loss**: On the Distillation Loss Functions of Speech VAE — **2604.12383**
12. **AG-REPA**: Causal Layer Selection for Representation Alignment in Audio Flow Matching — **2603.01006**
13. **ReaLS**: Exploring Representation-Aligned Latent Space for Better Generation — **2502.00359**