《UniWorld: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation》

阅 读 报 告

学　　校：南开大学

学　　院：网络空间安全学院

姓　　名：田晋宇

专　　业：物联网工程

二〇二五年六月

1 研究背景

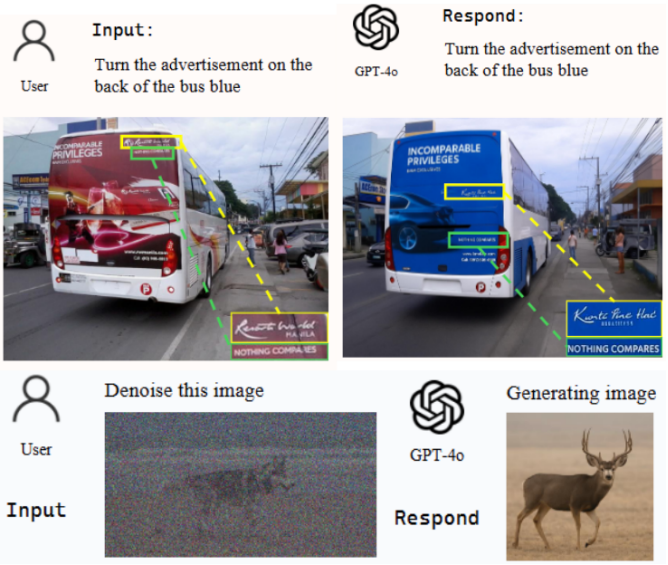

图 1: GPT-4o-Image 底层架构探究

近年来，多模态大模型的发展带来了视觉理解与图像生成任务的显著突破。统一架构的模型在视觉-语言理解和文本到图像生成方面已表现出强大的性能。

然而现有的一些工作通常聚焦于图像理解或文本生成任务，而较少涵盖图像到图像的感知与编辑任务，这在实际应用中尤为重要。部分已有方法试图使用 VAE 提取图像特征以实现编辑，但其生成内容中保留了大量低频信息，不利于表达语义级变化，导致在更复杂、统一的任务下表现不佳。

受到 OpenAI 发布的 GPT-4o-Image 模型的启发，这篇工作进行了两个关键实验，目的是推断其视觉特征注入机制：

1. **局部编辑实验**：给模型指令“将公交车背面的广告变成蓝色”，若模型使用 VAE 提取的低频特征，则图像中的文字等非编辑区域应几乎保持不变。然而实验结果表明，GPT-4o-Image 编辑后的广告文字位置发生明显变化，表明其更可能依赖由语义编码器提取的高层语义特征进行生成。

2. **图像去噪实验**：向一张狗的图像加入不同程度的噪声，要求模型恢复原图。在高噪声条件下，GPT-4o-Image 错误地将狗还原为鹿。进一步使用 Qwen2.5-VL 等多模态理解模型对噪声图像进行识别，结果同样识别为鹿。这表明 GPT-4o-Image 更可能基于语义模型的理解结果进行推理和生成，而非通过保留局部细节的 VAE 进行低频还原。

基于上述观察，这篇工作指出强大的统一模型如 GPT-4o-Image 并非依赖 VAE，而是使用语义编码器作为视觉特征来源。基于以上假设作者构建了一个具备图像理解、生成和编辑能力的统一模型 UniWorld，并探索如何通过高分辨率语义特征替代 VAE 以提供更具语义控制力的图像生成能力。

2 相关工作

2.1 BEGAL

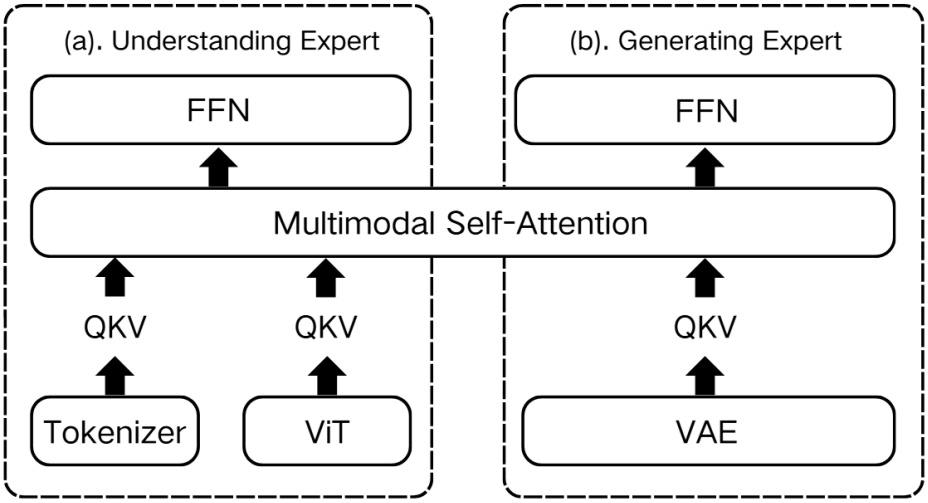

图 2: BAGEL Pipeline

在架构层面上 BAGEL 使用了 MoT 架构来同时处理图像理解与生成任务，理解 Expert 处理 ViT 编码后的视觉 token + 文本 token，用于抽取语义；生成 Expert 处理 VAE 编码的潜在表示，进行图像重建或编辑。不同专家之间通过共享的 Self-Attention 模块交互，使得理解得到的高层语义能自然传导到图像生成过程中，而不通过信息瓶颈或独立推理路径，促进跨模态理解与生成之间的无损语义信息融合。

在数据层面上 BAGEL 使用了大量跨模态交错数据，来自视频的时序图文对，捕捉物理世界中的概念变化来自网页文档的图文教程/百科，具备结构化、语义密集的信息流。

BAGEL 的训练策略分为四个阶段，通过不同的阶段逐步提升模型的多模态理解、生成、以及高级推理能力：

- **Stage 1:** 将视觉编码器与语言模型进行对齐，确保两者在表示空间上的兼容性。此阶段中仅训练中间连接模块。
- **Stage 2:** 执行大规模的多模态预训练，模型在包含文本、图文对、交错网页与视频等数据上学习基本的跨模态理解与生成能力。
- **Stage 3:** 提高视觉输入的分辨率，引入更多的交错数据，从而增强模型在高分辨率图像与复杂上下文下的表现，特别是在推理与图像编辑方面的能力。
- **Stage 4:** 使用高质量任务数据集对模型进行有监督微调，进一步优化生成与理解性能，提升在多模态任务中的实用性与泛化能力。

2.2 Janus-Pro

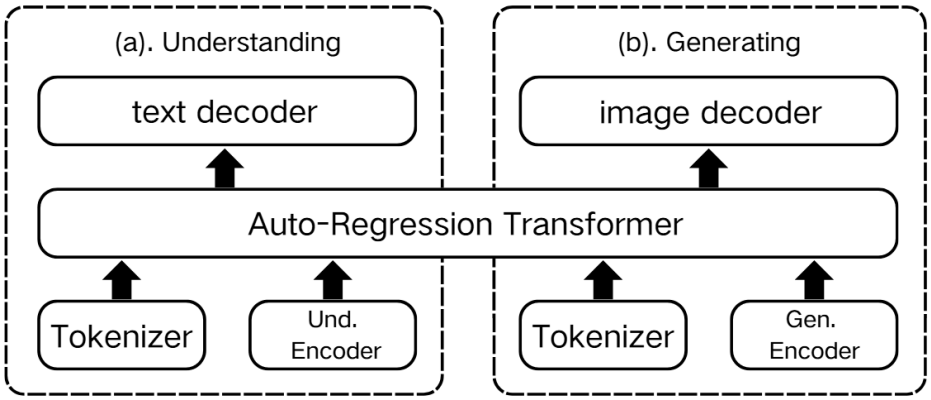

图 3: Janus-Pro Pipeline

Janus-Pro 在架构设计上简单来说就是解耦生成与理解任务，对于不同的任务采用不同的 Image encoder 和输出头，更像一个多头专家系统而非 BEGAL 中的 MoT 设计。理解 Encoder 使用 SigLIP 对图像提取语义特征，通过适配器转换为语言模型输入，结合文本 token 用于视觉问答、图文理解等任务；生成 Encoder 使用 VQ 编码器将图像离散为 token 序列，并映射为语言模型输入，指导图像生成与编辑。

- **Stage 1:** 训练理解与生成路径的适配器模块与图像头，延长训练步数以增强类别语义与图像表示之间的对齐效果。
- **Stage 2:** 在统一的自回归框架下进行大规模图文预训练。与 Janus 不同，Janus-Pro 直接采用自然语言描述的图像对，而不再使用 ImageNet 类别数据，提升了生成任务的训练效率。
- **Stage 3:** 在监督微调中调整数据比例（从 7:3:10 调整为 5:1:4），在保持生成能力的同时强化理解性能。

2.3 MetaQueries

传统统一多模态模型通常通过一个大型自回归 Transformer 来同时处理文本输出和图像生成，这会导致训练上非常复杂。MetaQueries 提出理解归 LLM，生成归 Diffusion 的策略，保持 MLLM 冻结，仅训练 Connector 模块和生成模型，简化训练流程，同时提升生成能力。为了避免直接使用 LLM 输出 token embedding 造成的生成质量差、不稳定，引入了 MetaQueries 主动引导 MLLM 激活其内部深层表示，提取理解后的知识表示而不是生成输出，拿到 MLLM 的中间条件特征 C 送入 Connector 编码器，将特征映射到 Stable Diffusion 输入空间，作为扩散时的 condition 生成图像。

MetaQueries 的训练策略主要分为两阶段：

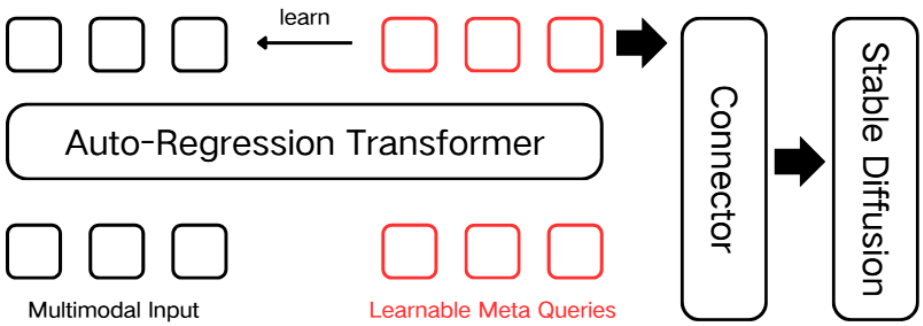

图 4: MetaQueries Pipeline

- **Stage 1:** 预训练阶段，在 2500 万对图文对上仅训练 MetaQueries、Connector 和 Diffusion 模块。
- **Stage 2:** 指令微调，构造自然图像对利用 MLLM 生成转换指令，学习图像编辑、风格转化、logo 设计等复杂任务。

3 方法路线

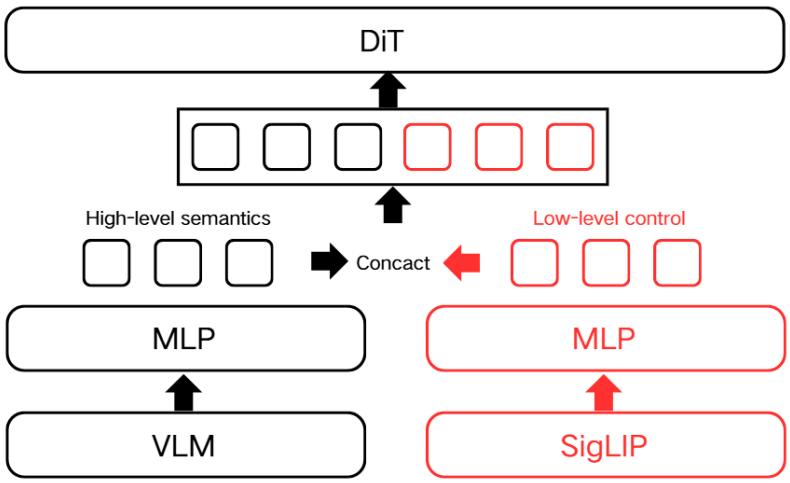

图 5: UniWorld Pipeline

3.1 模型架构

UniWorld 采用统一的生成架构，结合不同的子模块实现对图像感知与编辑任务的支持：High-level 的语义特征使用预训练多模态大模型 Qwen2.5-VL-7B 提取，自动回归的高层语义信息，其参数在训练中被冻结，用于保持语义理解能力；Low-level 的像素信息引入高分辨率语义编码器 SigLIP2-so400m/14，用于提取图像的低层细节和全局语义，替代传统的 VAE。生成模型 DiT 采

用 FLUX 架构下的 Flow Matching 机制生成图像。理解模块输出的 token 与 SigLIP 的图像特征在 MLP 层连接后输入至 DiT 生成器。

FLUX 原始使用 T5 作为语言条件，UniWorld 中其为可选项，但作者不建议在训练早期使用，以防陷入局部最优。

3.2 训练策略

UniWorld 训练流程分为两个阶段：

- **Stage 1:** 第一阶段用于对齐 VLM 输出与 DiT 文本分支的语义表示，在此过程中仅使用 VLM 的 token，不使用 SigLIP 图像特征，而且仅对 VLMDiT 的 MLP 进行训练，其他部分参数保持冻结。在第一阶段的训练之后模型能初步实现基于文字的图像生成，但缺乏图像参照一致性。
- **Stage 2:** 第二阶段是图像生成一致性微调，引导模型利用 SigLIP 图像特征完成参照一致的图像生成，此过程中只解冻 SigLIP 与 DiT 相连的 MLP。在 5k 1w 个 steps 后，模型开始有效利用图像特征实现编辑指令。

3.3 ZeRO-3 EMA

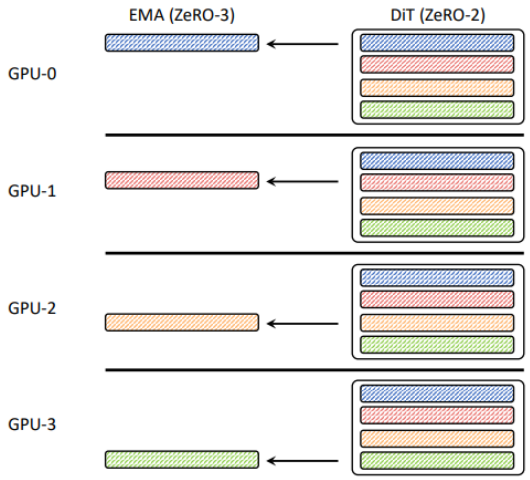

图 6: ZeRO-3 EMA

为了稳定模型训练，UniWorld 引入了 ZeRO-3 结构的 EMA 策略。训练模型采用 ZeRO-2，并行优化高效。EMA 模型使用 ZeRO-3 跨 GPU 分片存储，每个 GPU 仅保存部分 FP32 权重副本，有效降低内存占用。这样做的优势在于能够平滑训练过程中的权重波动同时允许更大的 batch size；EMA 每步更新一次，每 GPU 计算成本为原来的 1/N（N 为 GPU 数量），具有高度可扩展性。

3.4 训练数据

训练数据主要涵盖三个方面，包括图像感知任务（1.4M），图像编辑任务（1M），文本生成图像任务（30w）。文生图的训练数据来自于最新开源的 BLIP3-o 和 Open-Sora Plan 的高质量训练数据。BLIP3-o 是最新开源统一多模态模型系列，旨在统一图像理解与生成任务，其训练集包括预训练数据和指令调优数据；Open-Sora Plan 是由作者团队发起的一项开源且高质量的工作，在技术创新、性能表现和社区生态等方面展现出显著优势，成为开源视频生成领域的重要代表，其训练集包含高质量、无水印视频，总时长约 274 小时以及 300 万张来自 LAION 数据集的高审美评分图像。

作者团队并没有在数据量上下特别大的功夫，仅用少量数据即可在多个 BenchMark 上达到 SOTA，相信在未来一定会有更大的潜力。

3.5 图像编辑任务优化

大多数图像编辑仅影响图像的一小部分区域，而统一的损失函数可能掩盖了编辑区域的训练信号。为解决图像编辑中编辑区域小、背景大的损失不平衡问题，UniWorld 设计了自适应编辑区域加权策略。

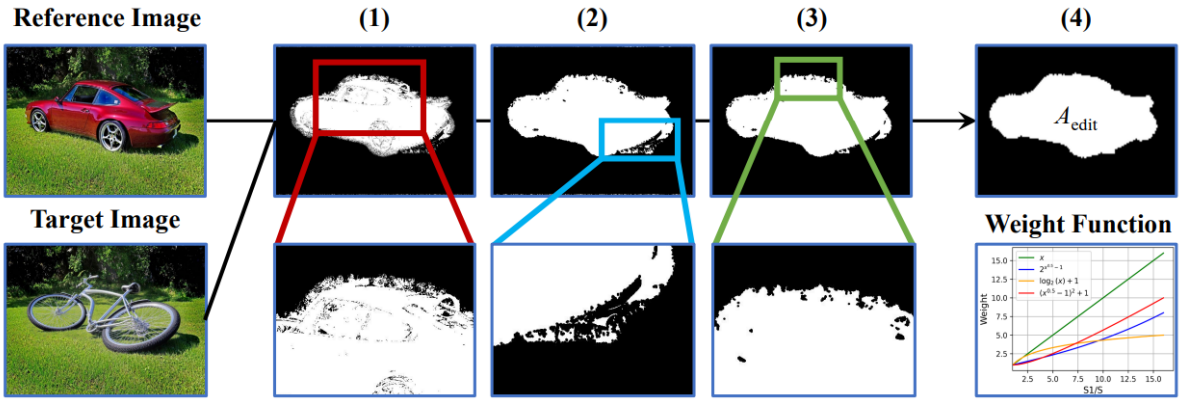

图 7: 自动掩码生成 Pipeline

由于许多公开图像编辑数据没有提供显式的编辑掩码，UniWorld 设计了一个四步流程来自动生成掩码:

- **像素差分**：计算参考图像与目标图像之间的像素差异，并设定阈值初步标记编辑区域。
- **膨胀**：扩大标记区域，缓解微小位移对编辑识别的影响。
- **连通域过滤**：移除过小的噪点区域。
- **最大池化**：增强掩码区域内部的一致性并抑制干扰。

令 $x = A_{\text{total}}/A_{\text{edit}}$，定义像素加权函数 $w(x)$。作者希望加权函数 $w(x)$ 满足以下性质：当编辑区域较小时（即 x 较大），对其赋予更大的权重；当编辑区域为整图时，退化为统一权重，即 $w(1) = 1$。为此，作者设计了以下四种候选函数：

1. **Linear:** $w(x) = x$
2. **Exponential Root:** $w(x) = 2\sqrt{x-1}$
3. **Logarithmic:** $w(x) = \log_2(x) + 1$
4. **Quadratic Root:** $w(x) = (\sqrt{x}-1)^2 + 1$

最终选用对数函数 $w(x) = \log_2(x) + 1$，因其增长平滑、稳定性好，适应性强。

这一优化策略提升了模型对于小区域编辑的关注度和执行精度，也是在图像编辑任务中能够超越其他开源大型模型的关键所在。

4 实验结果

Model	Understanding				Image Generation		Image Editing						
	MMB^V	MMB^I	MMMU	MM-Vet	GenEval	WISE	Overall	Add	Adjust	Extract	Replace	Remove	Hybird
Image Understanding													
LLaVA-1.5 [21]	×	36.4	67.8	36.3	×	×	×	×	×	×	×	×	×
LLaVA-NeXT [51]	×	79.3	51.1	57.4	×	×	×	×	×	×	×	×	×
Image & Video Understanding													
Video-LLaVA [18]	1.05	60.9	32.8	32.0	×	×	×	×	×	×	×	×	×
LLaVA-OV [15]	0.94	80.8	48.8	57.5	×	×	×	×	×	×	×	×	×
Text-to-Image Generation													
SDXL [29]	×	×	×	×	0.55	0.55	×	×	×	×	×	×	×
FLUX.1 Dev [14]	×	×	×	×	0.66	0.50	×	×	×	×	×	×	×
Image Editing													
MagicBrush [50]	×	×	×	×	×	×	1.85	2.72	1.47	1.31	1.89	1.57	1.80
Instruct-P2P [3]	×	×	×	×	×	×	1.89	2.29	1.79	1.33	1.93	1.49	1.48
AnyEdit [43]	×	×	×	×	×	×	2.63	3.12	2.66	1.82	2.71	2.34	2.07
Step1-Edit [23]	×	×	×	×	×	×	3.17	**3.90**	3.13	1.87	3.45	2.61	2.52
Unified Understanding & Generation													
Show-o [40]	×	-	27.4	-	0.68	0.35	×	×	×	×	×	×	×
Janus [39]	×	69.4	30.5	34.3	0.61	0.18	×	×	×	×	×	×	×
Janus-Pro [6]	×	75.5	36.3	39.8	0.80	0.35	×	×	×	×	×	×	×
Emu3 [38]	-	58.5	31.6	37.2	0.66†	0.39	-	-	-	-	-	-	-
MetaQuery-XL [27]	-	83.5	**58.6**	66.6	0.80†	**0.55**	-	-	-	-	-	-	-
BAGEL [8]	-	**85.0**	55.3	**67.2**	**0.88**†	0.52	3.17	3.55	3.30	1.56	3.38	2.44	2.55
UniWorld-V1	**1.79**	83.5	**58.6**	67.1	0.84†	**0.55**	**3.37**	3.86	**3.70**	**2.23**	**3.49**	**3.54**	**3.13**

图 8: UniWorld 在不同 Benchmark 下的性能对比

论文所提供的实验数据中 Uniworld 在图像理解、图像生成、图像编辑三大核心任务上均展现出领先的统一能力。在开源模型这一行列中在大部分的 BenchMark 下都取得了 SOTA 或者接近 SOTA 的表现，证明了该架构的多模态统一性与数据效率。

4.1 文生图能力

GenEval 的原始得分为 0.80，使用 BLIP3-o 的重写提示后提升至 0.84，接近 BAGEL（0.88）。UniWorld-V1 仅使用 2.7M 的训练数据，远低于 BAGEL（2665M），展现出极强的数据利用效率。

值得注意的是在最新提出的专门用于评估文生图世界知识整合能力的 WISE 总得分 0.55，在统一模型中表现优异，在空间知识类别中得分 0.73，是除 GPT-4o-Image 外的最佳表现，表明其在空间推理方面尤为突出。

4.2 图像编辑能力

在基于 ImgEdit-Bench 的比较中，UniWorld-V1 以 3.37 的总分成为所有开源模型中表现最强者。在 Adjust、Remove、Extract、Replace、Hybrid 五个子任务中均取得开源最高分。总体能力仅次于 GPT-4o-Image（4.31），是开源中最接近行业顶级的图像编辑模型。

4.3 图像理解能力

在 MMBench、MMBI、MMMU、MM-Vet 四个基准上均取得了 SOTA 或者接近 SOTA 的表现, 得益于冻结 vlm 的策略，减少了训练资源消耗，同时避免理解能力退化。

4.4 图像感知能力

目前暂无可完全评估全面感知能力的 BenchMark，因此通过与 GPT-4o-Image 的质性案例对比进行分析，UniWorld-V1 在多个感知任务中表现出高度竞争力。在边缘检测，法线图生成，语义分割，草图生成任务中表现更优。

5 个人思考

在本项工作的开篇部分, 一个精妙的对比实验设计令我印象深刻。由于无法直接获取 GPT-4o 的内部实现细节，作者巧妙地设计了一系列行为对比实验，尝试从其外在表现中反向推理其视觉特征提取机制。这一策略不仅提升了论文的可读性和说服力，也为理解封闭模型提供了新的思路。此外，UniWorld 在仅使用不足 0.01 倍的训练数据量的情况下，在图像编辑任务上全面超越了 BAGEL，展现出在未来与 GPT-4o-Image 竞争的巨大潜力。

在我看来，UniWorld 成功的关键在于其 VLM 和生成模型的融合方式：把不同层次的语义信息直接传递给生成模型。以往的工作中，DiT 架构通常将高层语义信息作为条件向量嵌入，用以指导生成过程。而 UniWorld 则通过 MLP 将高层语义与 DiT 的输入空间对齐后，直接与低层像素级特征拼接。这一做法不仅保留了像素级细节信息，还有效引入了抽象语义，从而实现了显著的生成性能提升。然而，这种拼接方式不会影响 Decoder 对低层信息的建模能力，反而取得了卓越的生成质量，这一现象背后的深层机制值得我进一步探究。

在深入理解 UniWorld 的整体框架之后，我也产生了一些思考：

1. **生成是否反哺理解？**当前大多数统一模型的研究重点仍集中在理解促进生成这一方向，通过捕捉图文间隐式语义来提升生成质量。然而，尚缺乏系统性的研究来探讨生成是否反哺理解的可能性。我注意到作者团队提出的 WISE Benchmark 也主要用于评估理解对生成的贡献。在 UniWorld 的训练策略中，VLM 模块的权重是冻结的。若将 DiT 的梯度回传至 VLM，是否会反向促进模型的理解能力？VLM 中包含多个编码器和解码器，具体应回传至哪些子模块更有助于提升整体理解性能？这些问题可能构成下一阶段探索的关键方向。

2. **Moe 融入的可能性？**在 UniWorld 框架中，生成模型的输入由先前工作的 VAE 提取的图像特征 + 文本嵌入扩展为多层级的语义输入，包含来自 VLM 的语义特征。这种多源输入为生成模型在去噪过程中提供了更丰富的指导信息，有助于提升生成质量与语义一致性。然而，引入多种类型的输入特征也对生成模型提出了更高的适配要求。不同来源的信息具有异质性，直接融合可能导致训练过程变得不稳定，甚至影响模型的收敛速度。未来可以考虑在 UniWorld 中引入 Moe 模块，来增强模型对多种信息的自适应处理能力。MoE 机制能够根据当前任务的语义特征动态激活最合适的专家子网络，从而有效适配文本生成图像、图像感知、图像编辑等多种任务，有望在保持模型统一性的同时提升参数效率和任务泛化能力。

3. **如何组织高效的后训练？**目前，UniWorld-V1 的训练流程主要侧重于预训练阶段，对于后续训练阶段尚未展开深入设计。未来的工作可引入一系列后训练策略，以进一步增强模型对复杂语义的理解能力。例如，BEGAL 在训练中引入了大量图文交错样本，促使模型对图像与文本之间的时间顺序和空间位置信息进行推理。这种在数据构建与训练策略上的设计体现了较高的创新性与有效性。在 UniWorld 的后续版本中，可借鉴类似 CoT 式的推理范式，以提升模型在多模态语境中的理解与生成能力。如何构建适用于图文推理的有效推理流程，可以成为未来研究中值得重点关注的方向。