模型能力简介

语义理解能力

Kolors 从模型和数据两方面来提升文生图模型的语义理解能力。模型方面，Kolors 选用了强大的中文 LLM ChatGLM 来作为文本编码器；数据方面，Kolors 使用 MLLM 机造细节丰富的高质量文本描述。

更强的文本编码器

文本编码器的能力是文生图模型的语义理解能力的关键。一开始，大部分文生图模型使用 CLIP（如 SD、DALL-E 2）来作为文本编码器。CLIP 目标通过图文对比学习训练，来建模视觉语言联合表征空间，是多模态学习领域里程碑式的工作，用于文生图模型的文本编码是很自然的想法。然而，受到训练目标监督粒度的限制，CLIP 对于一张图中有多个物体，有不同属性、位置的复杂情况通常理解得较差。表现在生图模型上，就会出现生图结果属性绑定错乱的问题。另外，原始 CLIP 模型文本编码的最大长度也很有限，只有 77。

为了提升文生图模型的语义理解能力，Imagen 首先提出了使用 T5 作为文本编码器，并指出了 scaling 文本编码器比 scaling 生图 UNet 带来的提升要显著得多。自此，新的文生图模型纷纷优化文本编码器的能力，有的引入更大更强的 T5-XXL（如 Imagen、Pixart），有的将多个 CLIP 或 T5 的特征结合起来（如 SDXL、SD3）作为文本条件。然而，现有的开源模型在中文生图方面的能力还比较一般。直到最近腾讯开源出的 Hunyuan-DiT，才有了一个比较可用的中文生图模型。Hunyuan-DiT 使用了双语的 CLIP 模型加多语言的 T5 模型来作为文本编码器。但是由于多语言 T5 的训练预料中中文占比太少（只有 2%），而 CLIP 又受限于本身训练目标，细粒度的文本理解能力较差。因此，目前开源界中文、细粒度文生图模型的文本编码器仍存在较大的优化空间。

Kolors 针对这一问题，选择直接使用大语言模型进行文本编码。具体来说，Kolors 使用了 ChatGLM-6B-Base 模型，这是一个中英双语的大语言基座模型。这里没有选择其 SFT 版本 ChatGLM-6B 是因为作者认为未经对齐人类偏好的基座模型反而更适合文本特征的提取。在最大编码长度方面，ChatGLM 也更高，达到了 256。与 SDXL 一样，Kolors 取文本编码器的倒数第二层特征作为文本条件。

下表对比了主流开源文生图模型所选用的文本编码器和支持的语言。

使用 MLLM 生成细节更丰富的文本描述

要增强文生图模型的语义理解能力，除了更强大的模型结构之外，更重要的是训练数据。一开始，文生图模型都是使用网络上的图文对数据进行训练，而这些图文对不可避免地存在不准确、细节不够丰富的问题，因此用于文生图模型的训练质量不够高。DALL-E 3 首先提出使用 image captioner 模型机造生成细节丰富的高质量文本描述。最近的文生图模型中，recaption 机造数据几乎已经是标配。

那么，开源 MLLM 那么多，用哪个呢？Kolors 通过 5 个指标（长度、完整度、正确性、幻觉、主观性）来评估 MLLM 生成 caption 的质量。经综合比较，GPT4V 和 CogVLM 生成的 caption 质量较高，再考虑成本，Kolors 最终选用了 CogVLM 来生成 caption。并且，除了 GPT4-V 和 InternVL-

XComposer 之外，作者发现直接生成的中文 caption 质量明显差于英文 caption，因此作者选择先生成英文 caption 然后翻译为中文。

MLLM 机造 caption 的质量虽然比原 caption 高，但是有些原 caption 包含的专有名词（如故宫、月饼）却是无法由 MLLM 生成出来。因此，Kolors 选择原始 caption 和机造 caption 50:50 的比例来进行训练，这与 SD3 的策略相同。

有了强大的文本编码器模型和高质量的 caption 数据，Kolors 展现出了强大的复杂中文提示词理解能力。从下图可以看到，使用 GLM 作为文本编码器的 Kolors 对提示词内容的生成更完整，且属性绑定正确。

中文文本渲染能力

准确生成文字的能力一直是文生图模型的一大难题。DALL-E 3 和 SD3 已经有了很强的英文文字生成能力。但是，目前还未有模型具有中文文字的生成能力。中文文字的生成有两点困难：一是相比于英文呢，中文汉字的集合太大，而且纹理结构更复杂；二是缺少中文文字的图文对数据。

为了提升中文文字的生成能力，Kolors 从两个方面准备数据。一是选择 50000 个最常用的汉字，机造生成了一个千万级的中文文字图文对数据集。但是机造数据毕竟真实性不足。因此，第二方面又实用 OCR 和 MLLM 生成了海报、场景文字等真实中文文字数据集，大概有百万量级。

作者观察到，虽然使用机造数据一开始中文文字的生成能力的真实性比较差，但是在结合高质量真实数据之后，真实性大大提升，而且即使是真实数据中不存在的汉字的真实性也得到了提升。

图片视觉质量

作为一个生图模型，好不好看，自然才是最关键的指标。Kolors 从数据和训练方法两方面入手，提升图片视觉质量。在网络结构方面，Kolors 没有进行改动，仍旧使用与 SDXL 一致的 UNet 结构。

高质量数据

Kolors 的训练分为概念学习和质量提升两个阶段。在概念学习阶段，Kolors 在十亿级的图文对数据集上进行训练，该阶段的数据包括开源的 LAION、DataComp、JourneyDB 等，也包括一些公司的内部数据。通过平衡各个类的样本数量，确保在该阶段能够学习到广泛的视觉概念。这是所有文生图模型训练都包含的过程。

重点在第二阶段，质量提升阶段，训练重点转移到提升图像的细节和美学质量，以及高分辨率的训练。笔者印象中，视觉质量提升阶段的重要性应当是 Meta 的 Emu 首先指出的，后来的 Playground V2.5 等模型也都加入了视觉质量提升阶段的训练。实际上，可以将第一阶段概念学习阶段类别为 LLM 的纯自回归的预训练阶段，将第二阶段的质量提升阶段类比为 LLM 对齐人类偏好的 SFT 阶段。类比 LLM，文生图模型的两阶段训练也可以看作是第一阶段学习大量的知识和概念，第二阶段对齐人类的美学偏好。

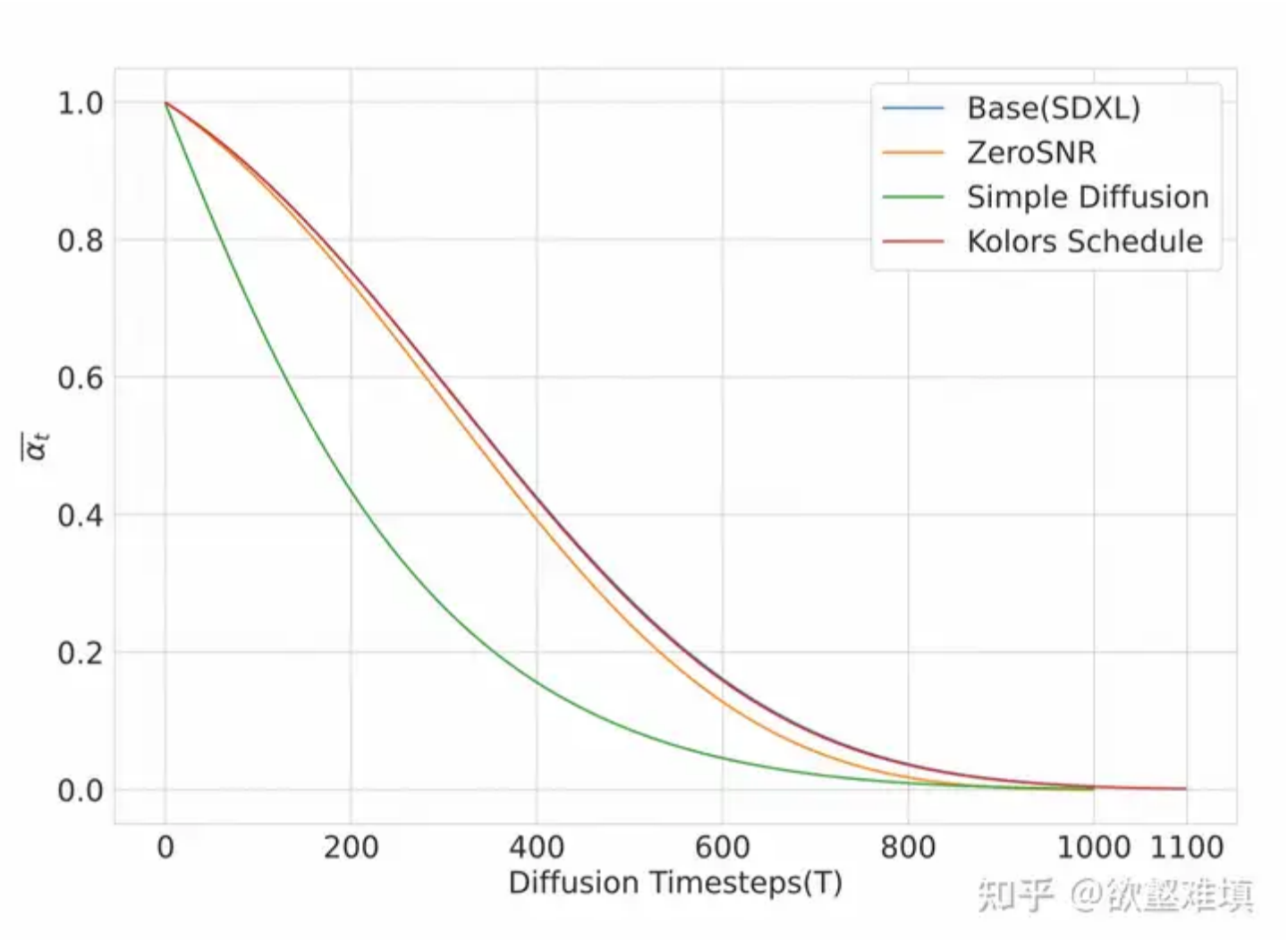
为了获取高质量的数据集，Kolors 首先使用常规的过滤策略（如分辨率、OCR 准确率、人脸数量、清晰度、美学分等）将图文对数据量级过滤到千万级。然后，将图片的质量划分为 5 个等级，人标！最终得到了百万量级的高质量图片。

高分辨率训练

Diffusion 训练时还有一个问题，那就是标准的 noise schedule 在给高分辨图像进行前向加噪声时，加的噪声不够。下图是以 SDXL 的 noise schedule 为例，给分辨率分别为 512 和 1024 的图像加噪声。明显可以看到，512 的图像在 1000 步基本已经是纯噪声了，看不出原图的样貌，而 1024 的图片在 1000 步还可以隐约看到原图。而在采样生成时，我们都是采样纯高斯噪声作为起始噪声，这就出现了训练、推理的差异，导致高分辨率的图像生成质量不高。之前有工作 Zero Terminal SNR 和 Simple Diffusion 提出了针对这个问题修改的 noise schedule。

而 Kolors 则提出了另一种方法。在训练的第一阶段，即概念学习阶段，一般是低分辨率图片，此时就采用 SDXL 中的 noise schedule。而在第二阶段，质量提升阶段时，将原来 1000 步的最大步数延长至 1100 步。这样就能保证在加噪的最后一步，噪声图的信噪比足够低，基本成为纯噪声图。同时，调整超参 ☒ 的值，从而保持 ☒¯☒ 随时间 ☒ 的曲线形状与原来一致。这样，从低分辨率训练阶段迁移到高分辨率训练阶段就会更稳定平滑。

下图对比了几种优化高分辨率图像训练的 noise schedule。可以看到 Kolors 与 SDXL 在前 1000 步是完全一致的，同时也能保持最终的信噪比足够低，达到纯噪声。这使得 Kolors 从低分辨率训练到高分辨率训练的过度更平滑。



此外，为了更好地支持不同分辨率生图，Kolors 训练时采用了 NovelAI 提出的 bucketed sampling 的方法，但仅在高分辨率训练时采用，以降低训练成本。

总结

Kolors 可以说是最近开源的文生图模型中最给力的一个了。从技术报告来看，改进也是很全面的，更强的中文文本编码器、机造的高质量文本描述、人标的高质量图片、强大的中文渲染能力，以及巧妙的 noise schedule 解决高分辨率图加噪不彻底的问题。可以说是目前主流的文生图训练技巧都用上了，实测效果也确实很不错。在看到 Kling 视频生成的强大表现，不得不让人赞叹快手的技术实力。