Structure-Guided Image Generation of *HunyuanDiT*

Contributor: Dejie Yang，Guanyu Jiang，Xufei Guo，Yuhao Chen，Hanxiao Wei，Kefan Wu

Advisor：Zhiming Li，Xinlu Lai

2024/10/25

Directory

Background

研究背景

在工作开始之前我们对当前现有的比较好的文生图模型进行了调研，并采纳尝试了几种比较好的模型，以下列举我们的部分调研结果：

LLM-Blueprint[ICLR2024]：

- 利用 LLM 生成 k 个布局，然后将其插值到单个布局
- 查询 LLM 以生成对象描述以及简明的背景提示，总结场景的本质
- layout-to-Image 模型将布局转换为初始图像
- 使用基于框掩码、框的（生成）参考图像和源图像的扩散模型细化框提议的内容

研究背景

LLM-grounded Diffusion：

- 阶段 1: 预训练的 LLM 根据用户的提示prompt生成图像布局，包括带有标签的边界框
- 阶段 2: 使用布局指导控制器（layout-grounded controller）引导预训练的扩散模型生成最终图像。

LayoutLLM-T2I：

- 这个模型使用了La-UNet架构, 将布局信息加入到图像生成中来：
- Layout Induction: 使用ChatGPT来生成所给文本prompt的layout布局。
- Prompt Encoding: 使用prompt encoder 分别处理文本提示、从文本中提取的关系三元组（主语、关系、宾语）以及生成的layout。
- Layout Integration:在 UNet 中引入 Layout-aware Spatial Transformer，以有效地将布局信息整合到扩散模型中，从而生成最终图像。

相关工作

布局到图像的生成

- 该任务的目标是通过给定的布局（通常为边界框、物体类别信息等）生成图像。具体方法可以采用 LostGAN 和 Layout2Im 等模型。 **LostGAN**，它通过输入对象的边界框（bounding boxes）及对应的类别信息生成高质量的图像。LostGAN 引入了基于生成对抗网络（GANs）的框架，将布局信息与图像生成过程紧密结合，使得生成的图像更加符合场景布局的预期。此外，**Layout2Im** 是另一个典型的布局到图像生成模型，它通过同时输入图像布局和类别信息，结合条件GAN架构生成符合指定布局的高质量图像。这些模型能够生成复杂的场景图像，且显著提升了在场景理解和对象排列中的表现 。
- **方法描述：**

  输入：物体的布局信息（bbox或热力图）

  模型：LostGAN、Layout2Im

  输出：符合布局约束的图像

  步骤：利用GAN架构，通过接收布局信息逐步生成图像，最后使用多尺度判别器优化生成效果。

相关工作

文本和布局到图像的生成

该任务旨在结合用户输入的文本描述和布局信息生成图像。常用的模型可参考ControlNet和DALL-E 2。**DALL-E 2**是基于CLIP模型的扩展，它能够根据输入的文本提示生成模型。而**ControlNet**在diffusion model基础上集成了用户提供的额外的控制信息，例如bbox和其他几何信息，允许用户更精准地控制生成图像中对象的排列。例如，用户可以通过指定多个物体的相应位置和大小生成复杂的多物体场景，避免了单纯文本生成时对位置、尺寸等信息理解不准确。

方法描述：

- 输入：文本提示和物体的布局信息（bbox等）
- 模型：ControlNet、DALL-E 2
- 输出：符合文本描述和布局的图像
- 步骤：先基于文本生成初步图像，再使用布局信息细化图像中对象的排列，保证场景的一致性。

• 相关工作

文本编辑图像：

该任务的核心是通过文本提示对现有图像进行编辑。可以使用Prompt2Prompt和InstructPix2Pix等模型。其中**Prompt2Prompt**是一种基于diffusion model的文本引导图像的编辑方法，它通过调整生成过程中的注意力机制，控制图像中特定对象或场景的变化。另一个例子是**InstructPix2Pix**，这是一个将文本编辑应用到现有图像的模型，用户提供新的文本提示，并通过模型引导对图像进行局部修改。

方法描述：

●输入：原始图像+修改文本提示

●模型：Prompt2Prompt、InstructPix2Pix

●输出：经过修改的图像

●步骤：通过模型在生成图像中捕捉特定对象的注意力图，然后根据新的文本提示对目标进行调整，最终输出更新的图像。

Data Synthesis

Case Study-1

◆ **检测缺失：**

书柜中的每一层书未能都检测出来，图片中其他物体也未能都检测出来。

◆ **边界框不准确：**

一些物体的边界框（bounding box）没有完全覆盖目标物体，或者包含了多余的背景。例如，图片中床头柜的边界框没有精确地包围柜子。

◆ **遮挡物处理问题：**

图片中某些物体（如书架上摆放的物品）被部分遮挡，检测模型可能未能准确识别出这些遮挡的物体或给出合适的边界框。

数据生成-模型1：Florence2

Florence-2 - Advancing a Unified Representation for a Variety of Vision Tasks

Xiao B, Wu H, Xu W, et al. Florence-2: Advancing a unified representation for a variety of vision tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 4818-4829.

Case Study-2

◆ **边界框不清晰：**
图片中只有两个较大的框，但这种大的边界框过于宽泛，无法精确识别出不同物体。房间内有多种物体，如床、书柜、窗户、植物等，需要更细粒度的检测。

◆ **边界框覆盖过大：**
边界框在某些情况下覆盖了较大的背景区域，例如左边框覆盖了房间的一部分和窗外的自然景色，这样的设置可能导致物体检测器将背景错误地视为目标物体。

◆ **检测缺失：**
房间里有多种物体，例如书架上的书、床上的枕头和毯子、桌面上的物品等。理想情况下，边界框应该分别针对这些物体进行检测，而不仅仅是对房间整体或大块区域。

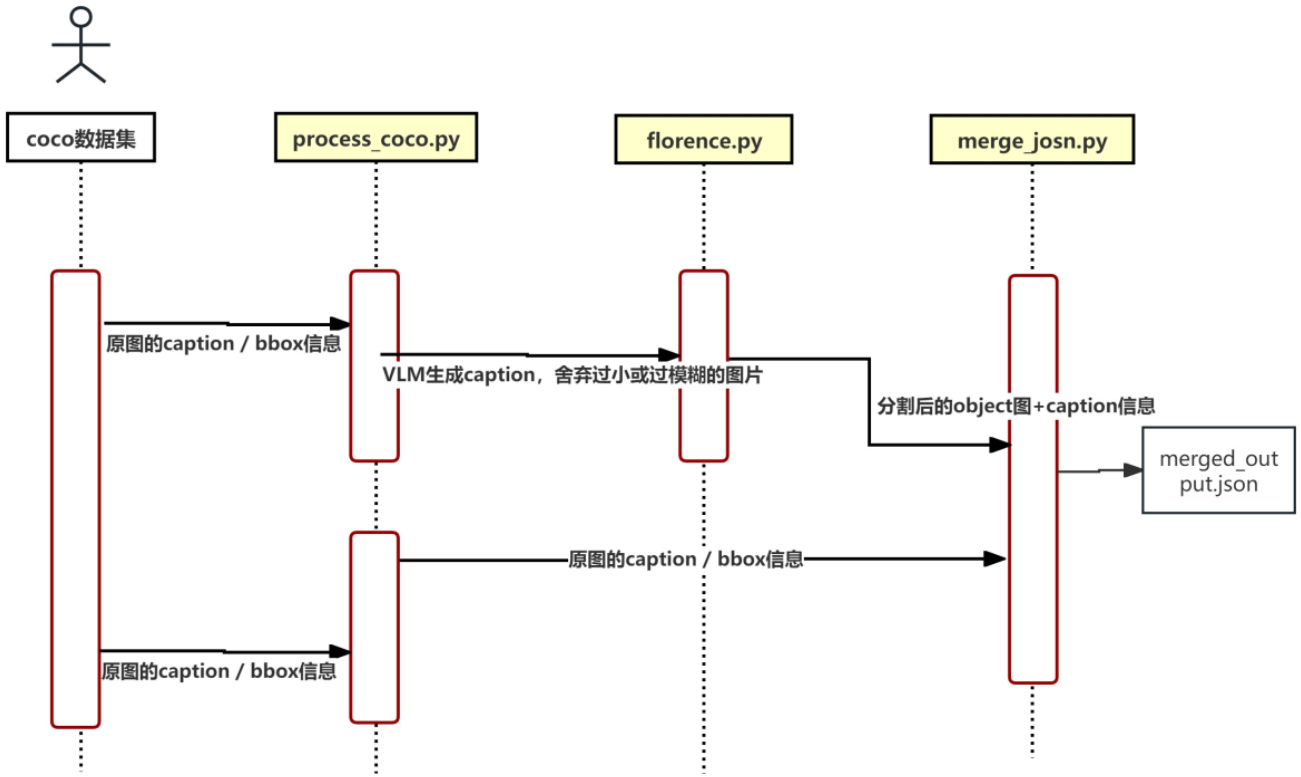
◆ **景中的复杂性：**
房间中有很多细节和遮挡物，尤其是植被与房间物品相交的地方。如果目标检测模型无法精细区分这些细节和遮挡，可能会导致检测不到某些物体，或者检测边界框不准确。

数据生成-模型2：KOSMOS2

Microsoft
Kosmos-2: Grounding Multimodal Large Language Models to the World paper

Peng Z, Wang W, Dong L, et al. Kosmos-2: Grounding multimodal large language models to the world[J]. arXiv preprint arXiv:2306.14824, 2023.

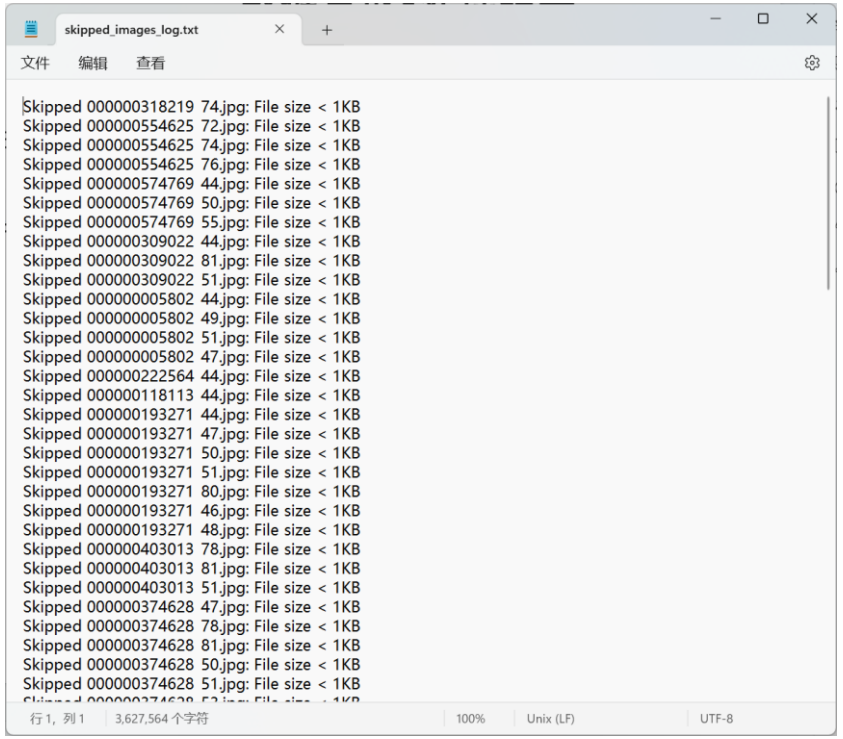
- **训练数据合成pipeline**



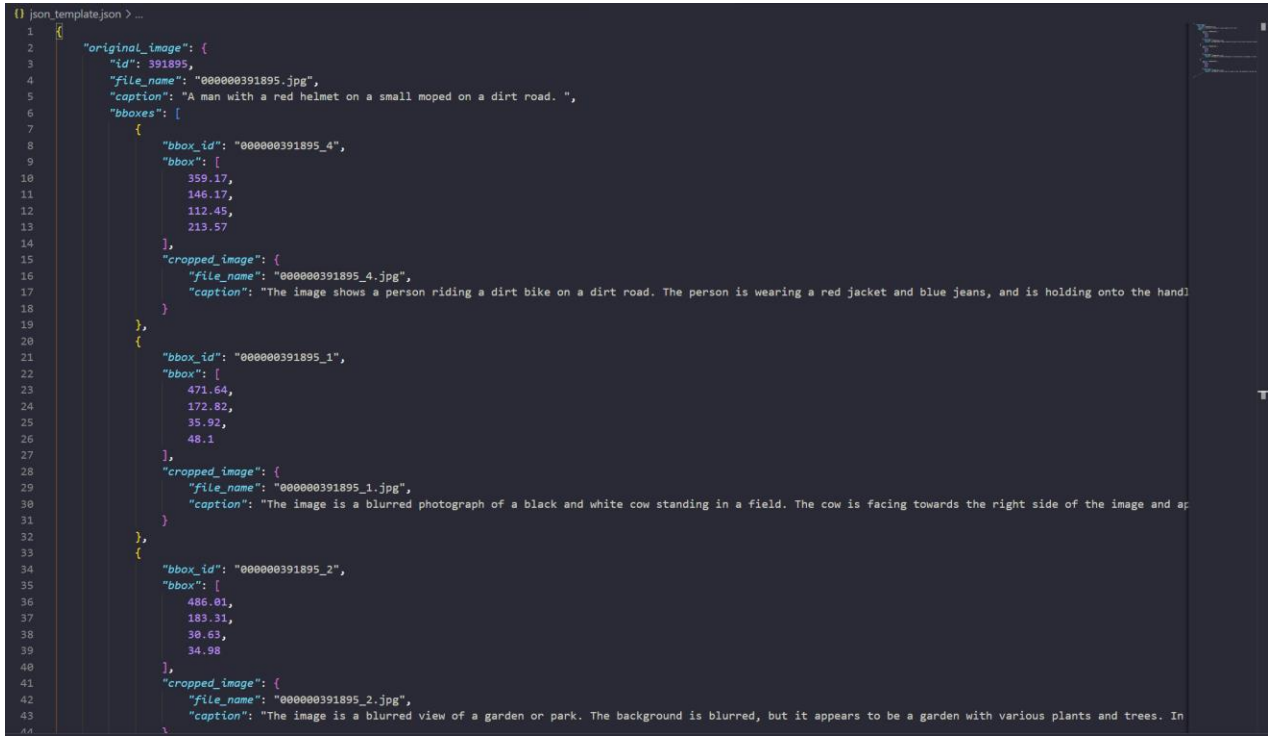
- ◆ 1. 找到一个能够准确检测出图像上所有 object 的模型/方法 - **直接使用cocoAPI**
- ◆ 2. 为一个检测出的bbox内的图像打上内容 caption 描述 - **Florence-2**
- ◆ 3. 用 prompt 提示要求 vlm 描述当前图像内容的整体 caption 描述（要求描述需要侧重于 object 之间的关系的视角）

• 训练数据合成结果

- 共生成了20w+的captions，舍弃掉了8w左右的blurred images
- 与原数据集的信息进行合并，生成json文件



图片过滤条目



json文件单元例

Model Training & Experiments

• Hunyuan-DiT 性能验证

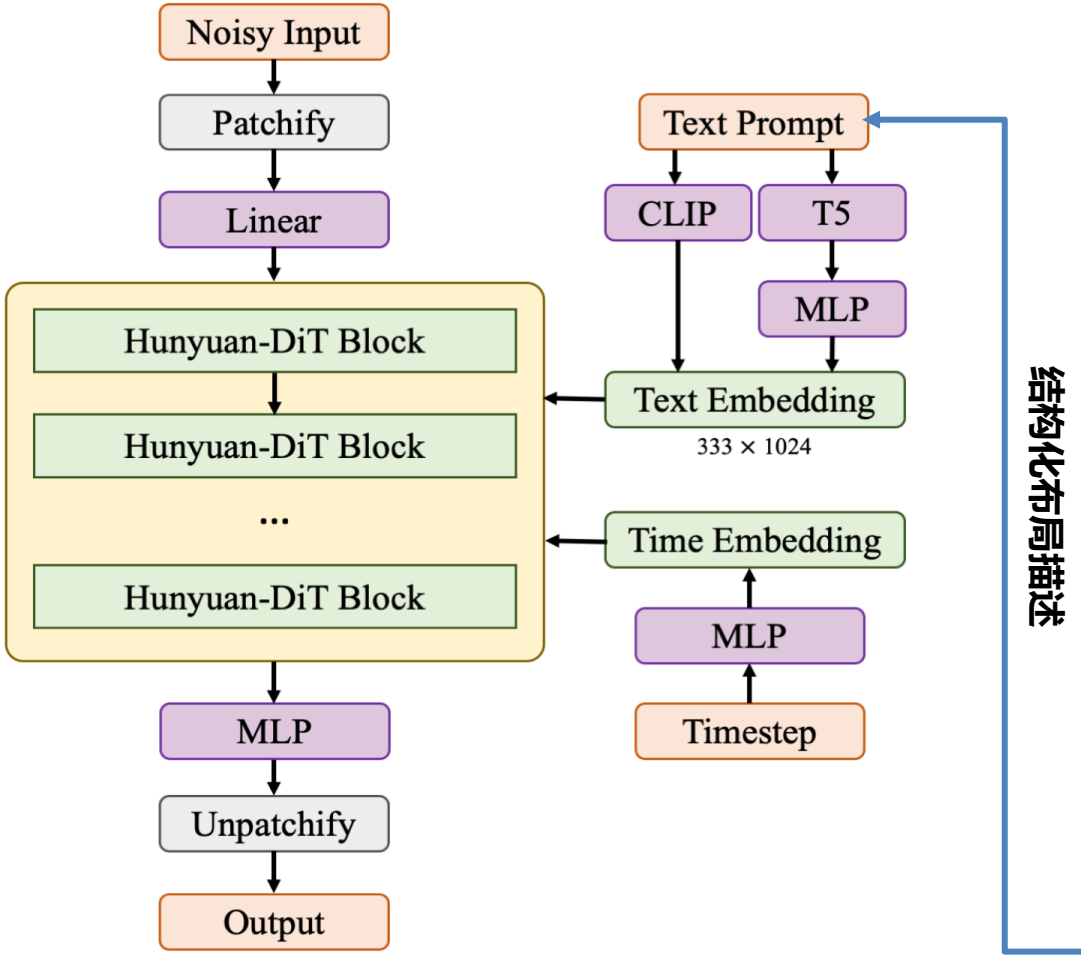
- 位置/关系理解
- 物理性质/大小理解
- 数量理解
- ……

Two young girl sitting on a couch with a box of pizza.

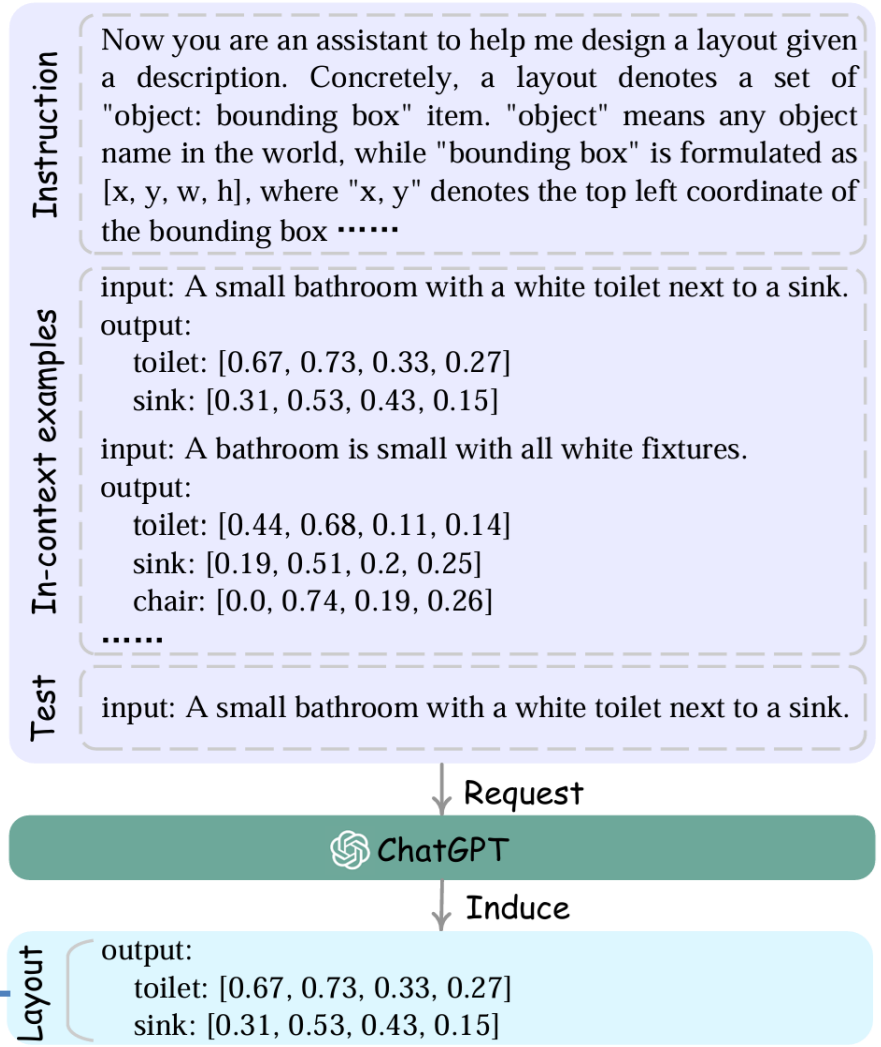
A banana on the left of a chair.

A cat in the middle of chair and dog.

• 结构化布局描述下的生成（text->picture）



结构化布局描述



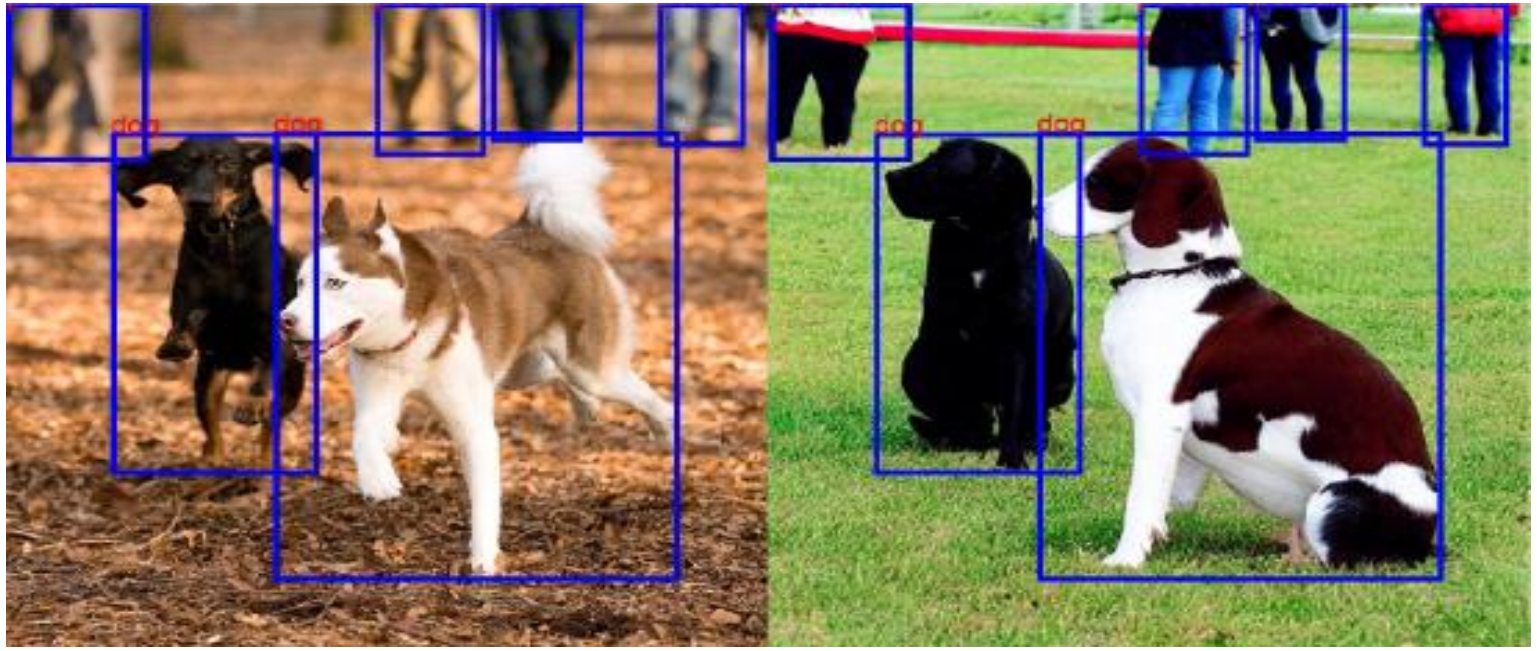
STEP2 :根据描述生成对应信息

STEP1 :生成结构化描述信息

Qu L, Wu S, Fei H, et al. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 643-654.

• 结构化布局描述下的生成结果

在给定文本提示的情况下：two dogs brown white and black and some people



真实结果

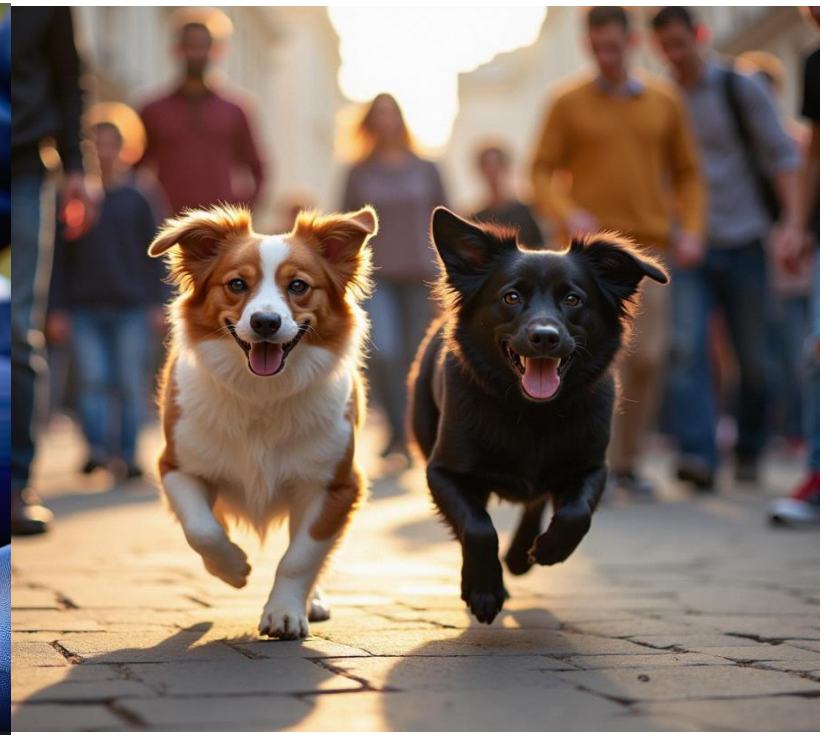
生成布局描述提示："bbox": [[0.6208046875000001, 0.5444453507340946, 0.531109375, 0.6918597063621533], [0.277109375, 0.462903752039151 7, 0.27228125, 0.5242088091353996], [0.5596171875, 0.1179771615008 1565, 0.146359375, 0.235954323001 6313], [0.914765625, 0.1101141924959 2169, 0.1119375, 0.2202283849918 4338], [0.096421875, 0.1215497553017 9446, 0.18428125, 0.2430995106035 8892], [0.6988749999999999, 0.1069738988580 7504, 0.12021875, 0.2139477977161 5008]], "label": ["dog", "dog", "person", "person", "person", "person"], "

• 结构化布局描述下的生成结果

在给定文本提示的情况下：two dogs brown white and black and some people

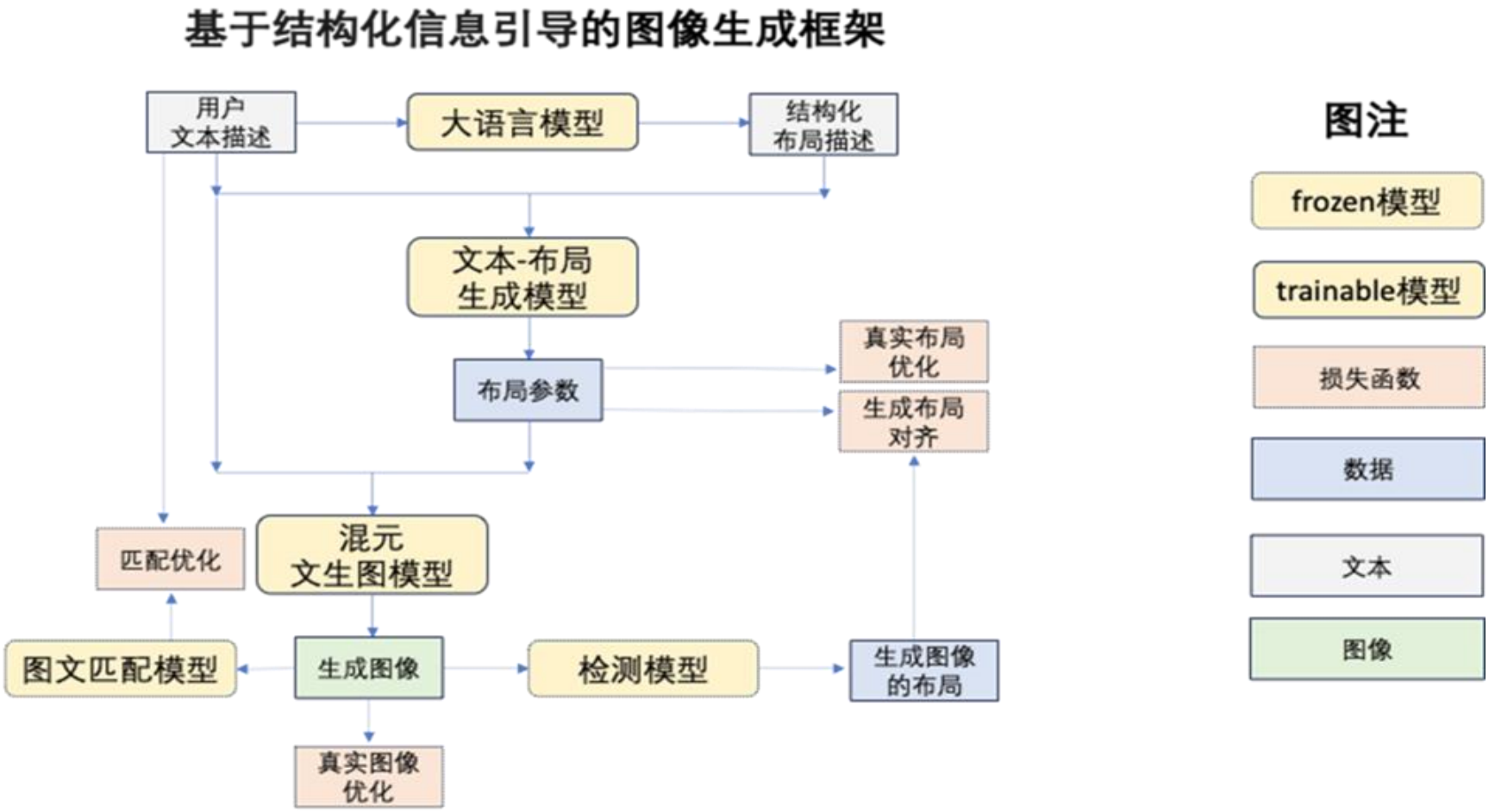
生成布局描述提示："bbox": [[0.6208046875000001, 0.5444453507340946, 0.531109375, 0.6918597063621533], [0.277109375, 0.462903752039 1517, 0.27228125, 0.524208809135 3996], [0.5596171875, 0.1179771615008 1565, 0.146359375, 0.235954323001 6313], [0.914765625, 0.1101141924959 2169, 0.1119375, 0.220228384991 84338], [0.096421875, 0.1215497553017 9446, 0.18428125, 0.2430995106035 8892], [0.6988749999999999, 0.1069738988580 7504, 0.12021875, 0.2139477977161 5008]], "label": ["dog", "dog", "person", "person", "person", "person"], "

仅文本描述提示



布局描述提示

- **结构化布局描述下的生成（text + layout->picture）**

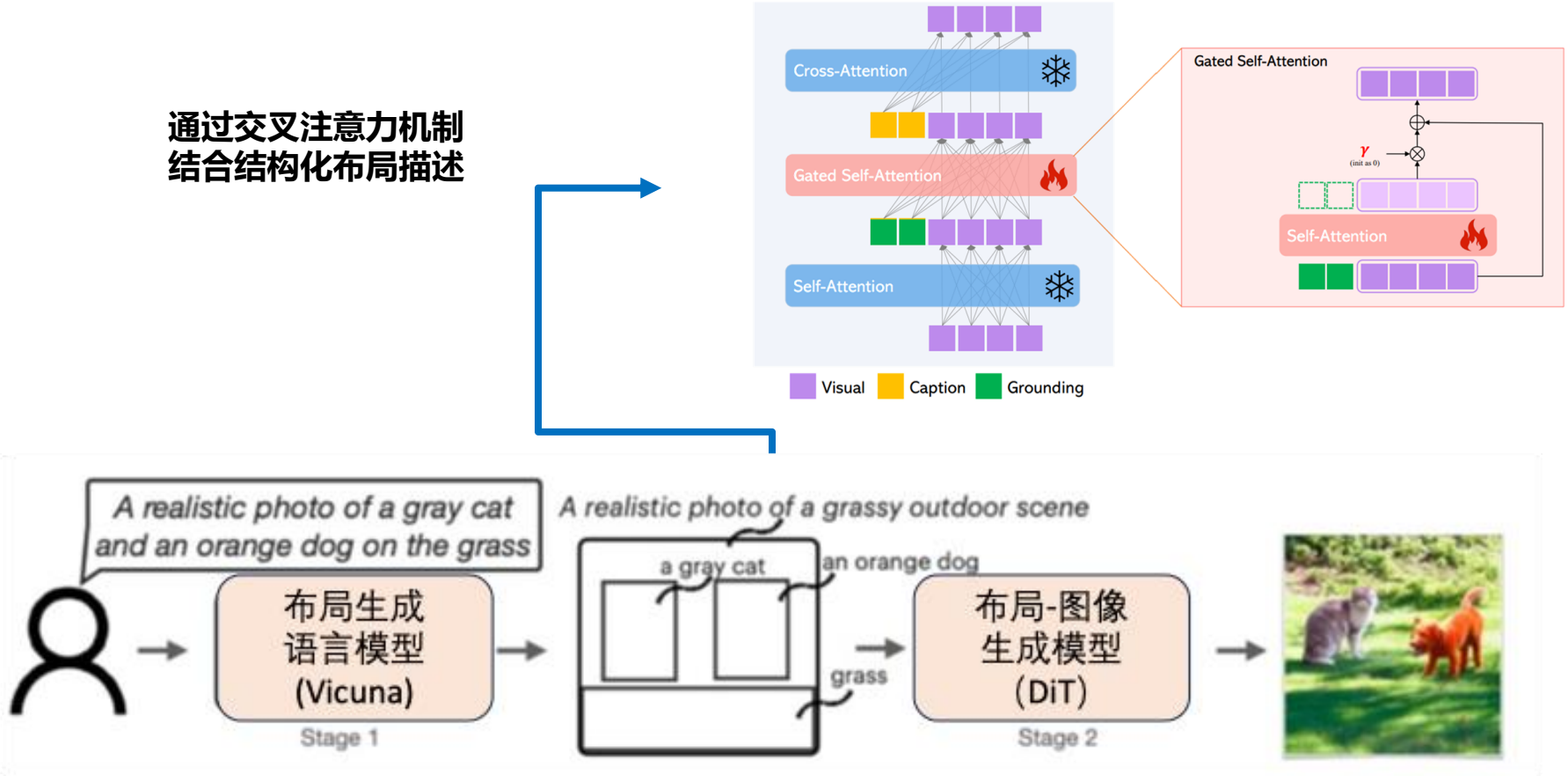


• 结构化布局描述下的生成（text+layout->picture）

1.要求预训练语言模型生成物体及其数量，然后生成物体的布局和描述

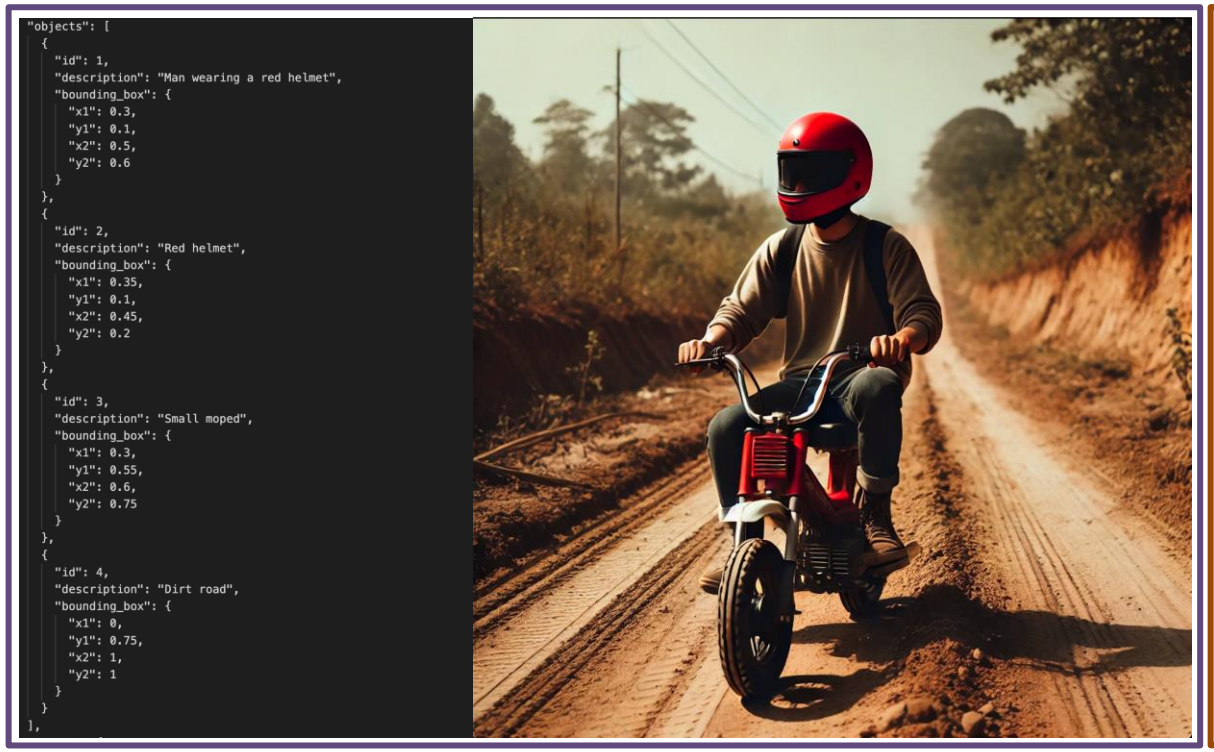
2.借鉴布局到图像生成模型GLIGEN 的思路并将布局描述结合进Hunyuan-DiT

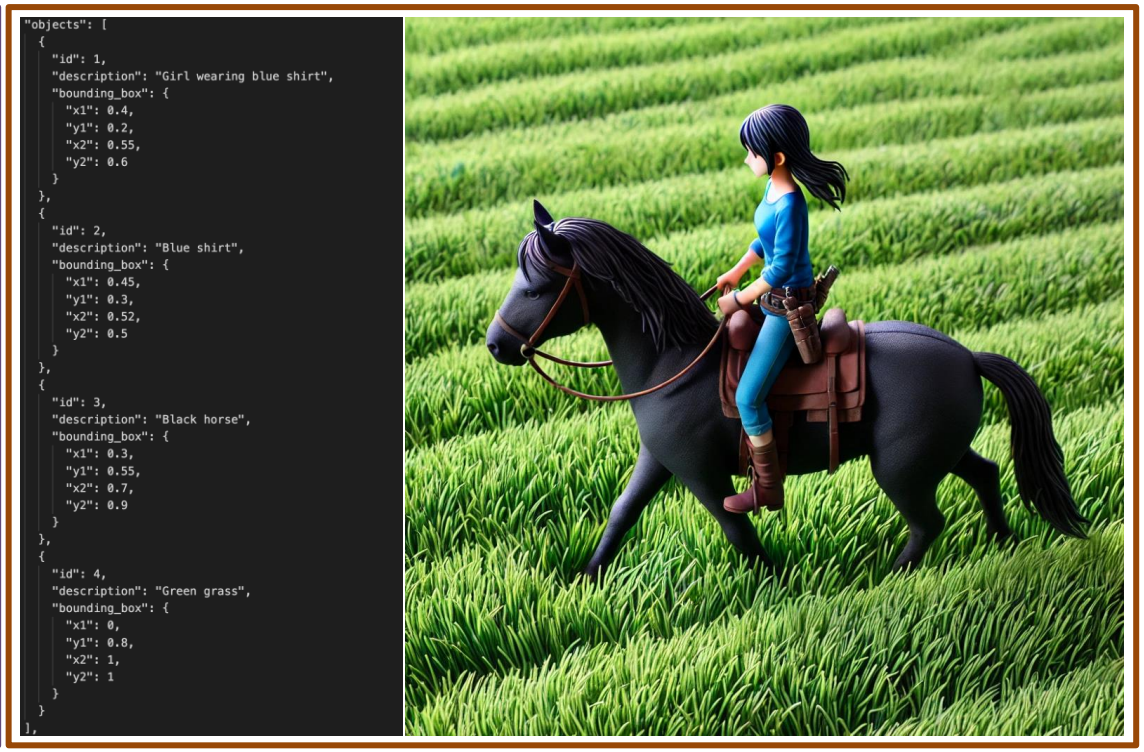
通过交叉注意力机制
结合结构化布局描述



• 结构化布局描述下的生成结果

输入描述："a man with a red helmet on a small moped on a dirt road."





输入描述： “a girl wearing blue shirt rides a black horse on the green grass.”

Conclusion

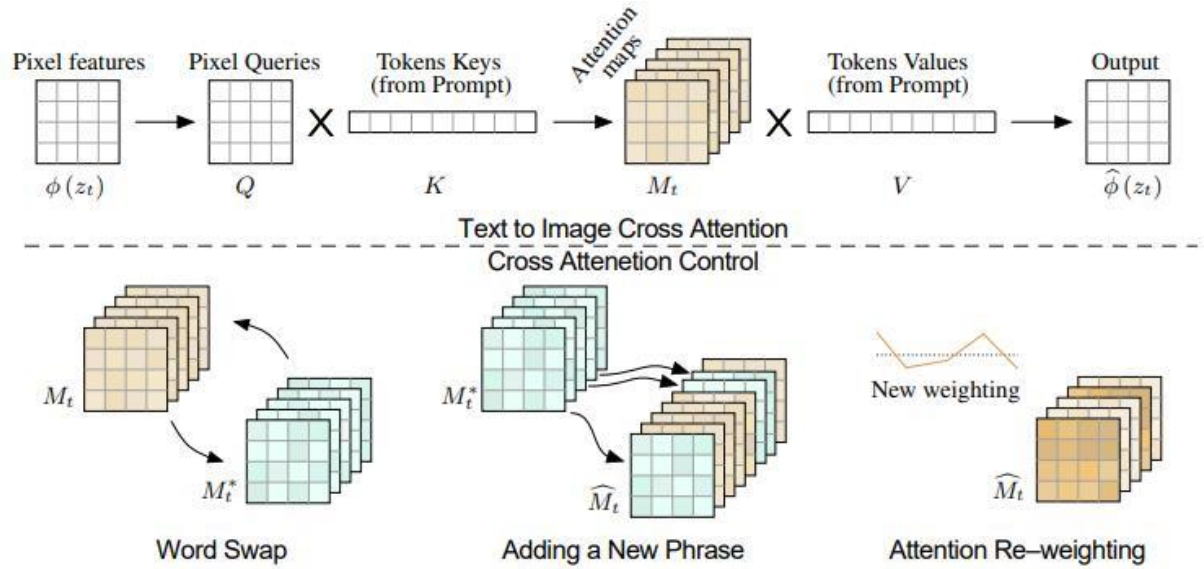
总结

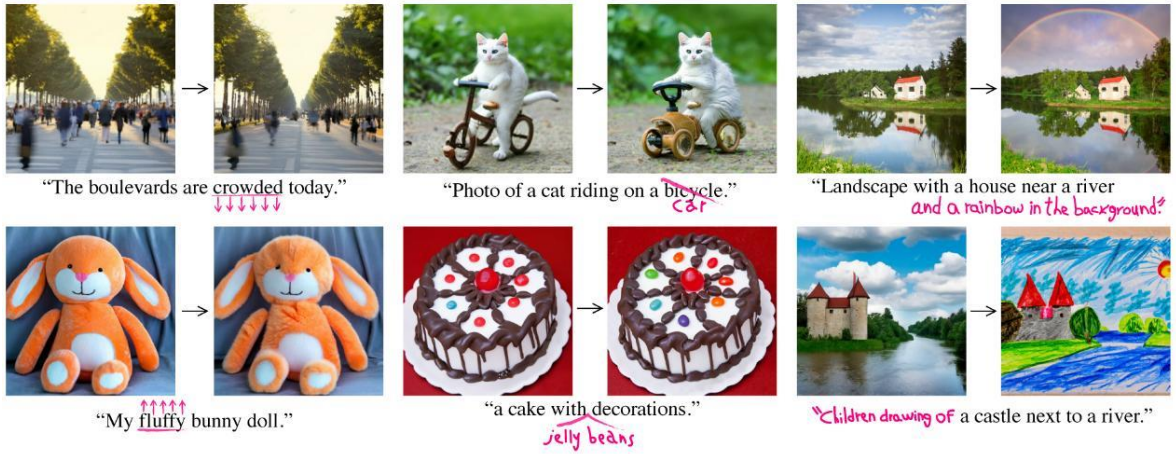
本课题组主要研究了**基于结构化信息引导的图像生成**方法，利用大型语言模型（LM）在视觉和语言任务中的控制能力，提升混元DiT模型在对象数量、空间关系和尺度方面的控制精度。

- 我们引入了文生图模型中LLM的prompt engineering的工作，在用户描述的基础上进行文本扩写，生成更好的描述，根据结构化描述的提示直接生成图像结果。
- 我们引入VPGen的生成流程，模型分为两部分：**(1) 布局生成和(2) 图像生成。**与传统的文本到图像（T2I）生成方法不同，我们通过文本描述对象及其数量和边界框，利用LM生成对象和布局，便于预训练语言模型处理未见过的对象。布局表示采用高效的边界框格式，并使用LoRA**微调Vicuna-13B模型**，结合**Flickr30K、MS COCO和PaintSkills数据集。**最后，我们使用基于**HunyuanDIT的GLIGEN模型**实现布局到图像的生成并进行微调。

后续愿景：prompt-to-prompt编辑

- 基于文本prompt的交互式图像编辑。
- 通过 Cross Attention maps 控制特定对象或图像区域的编辑。
- 无需重新生成图像，实现细粒度微调（如颜色、姿态、表情、物体位置）。





作用：

- 赋予模型局部、全局编辑控制能力。
- 文本驱动，无需用户提供mask。

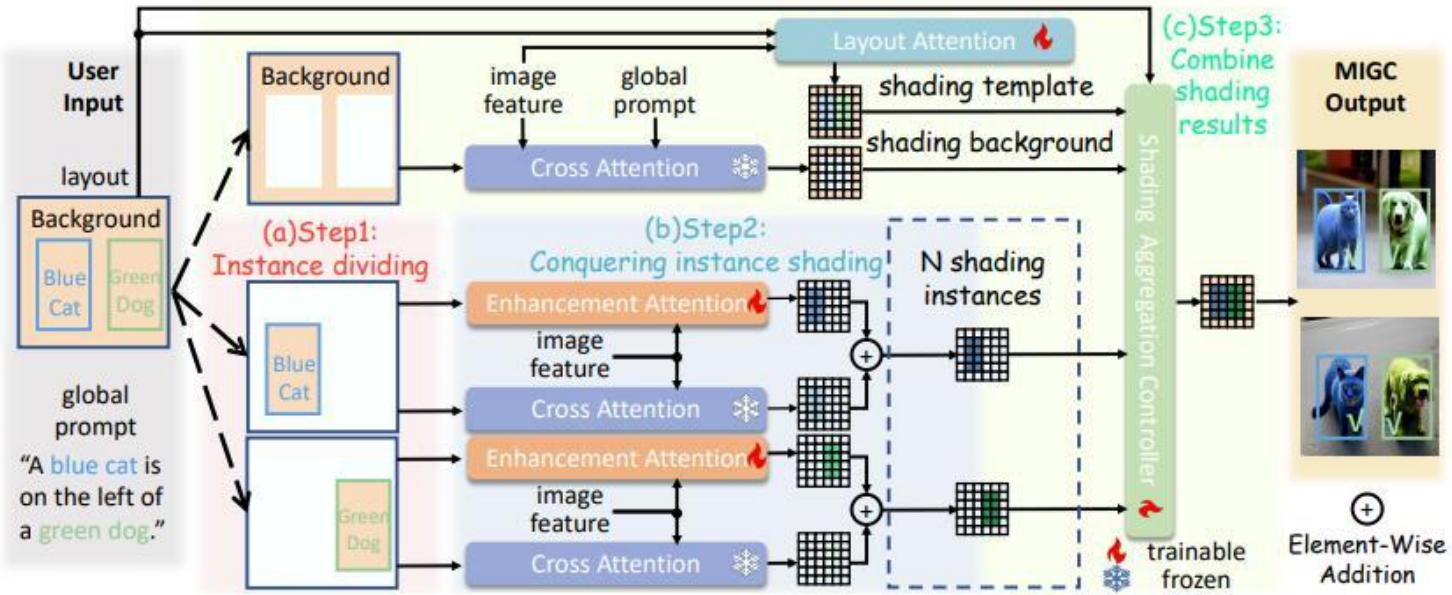
[Prompt-to-Prompt Image Editing with Cross-Attention Control]

后续愿景：MIGC

现状：生成多实例物体时缺乏精确控制

思路：

- **独立控制每个实例的生成过程，避免对象间的冲突。**
- **为每个实例生成独立的嵌入，以引导模型区分和生成多个对象。**
- **设定生成区域边界，保证多个实例在空间中的合理布局。**



挑战：提高生成图像质量的同时保持高效推理速度?

MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis

Thanks!

成员及贡献

- 杨德杰：文献调研、模型设计与实验
- 蒋官语：文献调研、数据集pipeline搭建及数据生产，模型设计及实验
- 郭栩菲：文献调研，验证混元DiT性能，模型设计及实验
- 陈钰豪：文献调研、数据生成方式调研测试
- 韦晗潇：文献调研、数据生成方式调研测试
- 吴可凡：文献调研、生成方式调研测试