1 Diffusion 模型基础原理与推导

Diffusion Model是一类基于变分推断的生成模型，其核心思想是将数据分布逐步加噪，构建一个正向扩散过程 $q_\phi(x_{1\,T}|x_0)$，以及一个可训练的逆向去噪过程 $p_\theta(x_{0\,T})$。以最经典的 **DDPM**（Denoising Diffusion Probabilistic Models，Ho 等人 2020）为例，其正向过程将输入图像 x_0 按时间步 $t = 1, \ldots, T$ 逐步加高斯噪声：

$$q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\big), \quad q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}),$$

其中 $\beta_{t\,t=1}^{T}$ 是预先设定的噪声方差调度（通常逐渐增大），在 t 越大时噪声越强，最终 x_T 接近标准高斯。对应的逆向过程为：

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t),\ \Sigma_\theta(x_t, t)\big),$$

常用参数化形式为 $\Sigma_\theta(x_t, t) = \sigma_t^2 I$ 或学习噪声预测 $\epsilon_\theta(x_t, t)$ 的方式。在训练时，对数似然 $\log p_\theta(x_0)$ 难以直接优化，因此利用变分下界（Evidence Lower Bound, ELBO）构造可优化目标。对 $\log p(x_0)$ 应用 Jensen 不等式可得：

$$\log p_\theta(x_0) \geq \mathbb{E}_{q_\phi(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q_\phi(x_{1:T} \mid x_0)}\right] \equiv \mathrm{ELBO}_{\phi,\theta}(x_0).$$

推导可得（参见）ELBO 分解为多项 KL 散度之和：

$$\begin{aligned}\mathrm{ELBO}(x_0) = \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - \mathbb{E}_{q(x_{T-1}|x_0)}[D_{\mathrm{KL}}(q(x_T|x_{T-1})\|p(x_T))] \\ - \sum_{t=1}^{T-1} \mathbb{E}_{q(x_{t-1},x_{t+1}|x_0)}\left[D_{\mathrm{KL}}(q(x_t|x_{t-1})\|p_\theta(x_t|x_{t+1}))\right].\end{aligned}$$

其中通常取先验 $p(x_T) = \mathcal{N}(0, I)$。该 ELBO 可通过变分推断细致展开，但在实际训练中往往采用简化目标。Ho 等人证明，只训练网络预测噪声 ϵ_θ 等价于最大化 ELBO 的一个变形，简化为均方误差损失：

$$L_{\text{simple}} = \mathbb{E}_{t, x_0, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right],$$

其中 $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$（$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$）。训练过程中，为每个时刻 t 样本一个噪声向量 $\epsilon \sim \mathcal{N}(0, I)$，并监督网络预测该噪声。这样可以避免显式计算复杂的 KL 散度。

DDPM 与 DDIM 的区别：DDPM 在反向采样时必须按完整 Markov 链逐步采样，步骤较多。Song 等人提出的 DDIM（2021）将正向扩散过程推广到一类非马尔可夫过程，并给出与 DDPM 相同训练目标

但可以确定性采样的方法。具体而言，DDIM 中可以通过选择 $\sigma_t = 0$ 得到确定性更新：

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x^0(x_t, t) + \sqrt{1 - \bar{\alpha}_{t-1}}\, \epsilon_\theta(x_t, t),$$

其中 $x^0(x_t, t)$ 为从当前噪声估计出的原始图像预测。如此从相同初始噪声 x_T 开始，每次生成都一致，无需随机采样；同时可跳过部分步长（采用子序列采样）大幅加速生成。**总结**：DDPM 强调建立明确的概率链（有噪声）并优化ELBO，而 DDIM 则给出一种训练后采样的优化，可通过取消噪声项实现加速采样，其训练目标与 DDPM 一致。

2 主流模型架构及实现

近年来出现了多种扩散模型架构，各有所长。**稳定扩散（Stable Diffusion）是目前最常用的文本到图像模型之一，由 CompVis 等提出。其采用潜在扩散模型**（Latent Diffusion Model, LDM）作为基础：首先训练一个自编码器（AutoencoderKL）将高分辨率图像 x 编码为低维潜在空间 z；然后在这个潜在空间上进行扩散过程，最后再解码回图像。Stable Diffusion 使用**固定的 CLIP-ViT-L/14 文本编码器**作为条件，将文本提示映射到潜在空间并通过交叉注意力注入到 U-Net 网络中。在架构上，其 U-Net 网络含约 8.6 亿参数，文本编码器约1.23亿参数。相比直接在像素空间训练高分辨率 DPM，LDM 显著降低了计算开销和显存需求，同时通过引入交叉注意力实现了强大的条件生成能力。如 HF 文档所示：“Stable Diffusion 是一种 *latent* 扩散模型……使用固定 CLIP 编码器来处理文本提示，模型相对轻量，可在消费级 GPU 上运行”。

Imagen（谷歌）通过使用**超大语言模型（T5）**来改进文本理解，从而提升图像质量。Saharia 等人的 Imagen 采用分级生成：先用一个小规模的扩散模型生成 64×64 图，然后逐级放大为更高分辨率。在训练时，他们发现：增加语言模型规模对生成效果提升比增大扩散模型更有效。Imagen 的关键在于使用预训练的大型变换器（如 T5-XXL）作为文本编码器，将高级语义信息很好地融合到图像扩散过程中，从而生成质量极高、语义一致性强的图像。研究表明，Imagen 在标准评测集（COCO）上达到很低的 FID 分数，甚至超越了同类方法。

PIXART-α（华为 Noah Lab）是一种最新的纯Transformer 架构的 T2I 扩散模型。该模型仅使用 Transformer 核心（不是 U-Net），并在其中嵌入交叉注意力模块，将文本条件直接注入 Diffusion Transformer（DiT）中。作者提出了分步训练策略：先分别优化像素依赖性、文图对齐以及美学质量，然后联合训练。在构建上，PIXART-α 采用了**高效的 T2I Transformer**，去除了传统扩散模型中的类别条件分支，通过交叉注意力实现文本控制。实验表明，其生成效果可与 Imagen、Stable Diffusion XL、Midjourney 等 SOTA 模型相媲美，但训练成本远低于传统大模型，大幅节省了时间和资源。

DALL·E 3（OpenAI）最新发布于 2023 年底，是基于 DALL·E 2 的改进版本。据 OpenAI 官方介绍，DALL·E 3 原生集成在 ChatGPT 中，使用户可以直接用自然语言与 DALL·E 交互。DALL·E 3 利用 ChatGPT 生成和润色提示词，将想法转化为详细提示，然后进行图像生成。虽然 OpenAI 并未公开全部技术细节，但据信其生成引擎依然基于扩散模型的思想，且优化了文本理解与对齐。指出：“DALL·E 3

原生构建在 ChatGPT 上……当用户提出创意时，ChatGPT 会自动为 DALL·E 3 生成量身定制的详细提示……如果希望调整，用户可继续用简单语言指导 ChatGPT 对生成图像做修改。” 这一设计强调了大语言模型在提示生成和交互中的作用。在实际应用中，DALL·E 3 显示了对细节的更好捕捉和与文本的更高一致性，相比 DALL·E 2 有显著提升。

3 扩散控制与个性化生成技术

为了在生成过程中引入更丰富的控制信号或实现个性化主题，一系列扩散控制方法相继提出。**ControlNet**（2023，由 Stability AI 与 CompVis 等提出）是其中典型代表。ControlNet 的核心是**“侧分支”网络结构**：在保持原有大型扩散模型（如 Stable Diffusion）冻结不变的情况下，引入一个与主网络结构相同但仅与特定控制信号（如边缘图、人体姿态图等）相连的副网络。该副网络的输出通过“零卷积”逐层融合到主网络中，以可控的方式影响最终生成。这样，ControlNet 能够利用预训练模型的强大表征能力，同时根据输入的任意结构化条件（边缘、分割图、深度图、草图等）实时调整生成结果。

LoRA（Low-Rank Adaptation，低秩适配）最初用于大语言模型的高效微调，被引入到扩散模型的个性化训练中。其思想是：冻结原模型参数，只在特定层（如 Transformer 的注意力层或 MLP）中注入可训练的低秩矩阵。对于 Stable Diffusion，一般将 LoRA 插入到**交叉注意力层**，专门学习将文本和图像潜在表示之间的映射更新。HF 博客指出：“在 Stable Diffusion 的跨注意力层加入 LoRA，可以让微调时只训练少量的矩阵参数，从而以很低的成本使模型适应新风格或新主题。” LoRA 常用于快速训练自定义风格、角色等，并已广泛应用于开源社区。

T2I-Adapter（2023）提出了一种轻量级的适配器结构，用以引入外部控制信号而无需修改预训练模型。具体而言，T2I-Adapter 在 U-Net 的每个阶段插入一个小型的卷积网络或 Transformer 分支，以接收外部条件（如色彩、线稿、纹理等）并输出与主网络特征尺寸匹配的特征图。论文中给出的架构示意图表明：整个系统由（1）预训练冻结的 Stable Diffusion 主干，以及（2）多个针对不同控制模态训练的 T2I-Adapter 组成。不同的 Adapter 可以组合使用，用户可通过加权叠加它们的输出以同时控制多种条件。实验表明，T2I-Adapter 在丰富控制能力的同时，保持了多条件的可组合性。

组合引导（Compositional Guidance）技术主要针对复杂场景中多个对象的生成。以 Snap 公司（Parmar 等，2025）提出的“Visual Composer”为例，他们引入了对象级指导（object-level compositional guidance）机制。具体做法是在扩散的采样过程中加入额外的梯度或注意力机制，逐步强化每个目标对象的准确性和布局正确性。文中提到：“在推理阶段，我们提出了对象级组合引导，以提高身份一致性和布局正确性”。此外，他们还设计了混合键值机制，将不同分辨率的编码器键值拼接，实现粗略布局与细节生成的协同。总体而言，组合引导通过对多对象场景施加单独指令或损失来改善生成质量，适合需要多物体组合的任务。

Token-Binding（标记绑定）方法源于对难以获取大规模数据的新概念（如独特商标、Logo）生成的需求。在 LogoSticker（Fudan 等，2024）工作中，提出了“Logo Token Binding”（标志令牌绑定）策

略。其核心是在训练数据集中将目标 Logo 贴到简单背景，使用文本反演（Textual Inversion）优化一个特殊文本令牌 V，使得模型能够准确识别该 Logo。具体流程为：第一，**Logo Token Binding Set**：将 Logo 粘贴到随机纯色背景，使用 Textual Inversion 优化令牌 V，使模型在这些简单图像中集中关注该 Logo。得到令牌后，第二步构建复杂场景数据集，将 Logo 放到真实场景图像中，并微调 U-Net，使模型学习 Logo 的细节特征。这样，V 与目标对象绑定，生成时在提示中包含 V 即可复现该 Logo，避免了仅用普通词汇难以捕捉细节的问题。

IP-Adapter（Image Prompt Adapter，Ye 等，2023）专注于利用参考图像引导生成。其架构在 Stable Diffusion 中额外开辟了一条跨注意力分支，将参考图像编码并注入到扩散网络中。简言之，IP-Adapter 令模型可以同时处理文字提示和输入图像信息，从而让参考图像对最终生成产生显著影响。如论文所述："IP-Adapters 将整个输入图像编码后通过独立的跨注意力层注入模型，使模型能同时处理文本和视觉信息。"这种方法常用于实现风格迁移、参考场景约束等任务，可以在生成时保持对给定图像主题或风格的连续关注。

4 角色一致性与多角色生成

在漫画生成任务中，**角色一致性**是一个核心难题，即同一角色在不同场景中应保持相同外观和身份识别。已有多个方法用于这一目的：

- **Textual Inversion**（文本反演）：Gal 等人（2022）提出的技术，使用少量示例图片为新概念（如人物或风格）学习一个新的"词"嵌入（文本令牌）。具体做法是将若干个包含目标概念的图像作为训练数据，在冻结扩散模型和词表的情况下，仅优化一个随机初始化的文本向量。训练完成后，这个向量表示新概念，并可嵌入生成提示中，从而在生成的图像中复现该概念。文中指出，只需 3–5 张示例图像即可生成稳定且忠实的概念表达。这种方法适合快速增加新的专属角色或风格，而无需微调模型权重。
- **DreamBooth**（Ruiz 等，2022）：该方法通过微调方式将特定个体的形象注入到生成模型中。给定少量目标人物的照片，DreamBooth 在一个预训练好的文本-图像生成模型上微调，使模型学会"唯一标识符"（如一个特殊词）与该人物绑定。训练时引入先验保持损失（prior preservation），防止模型忘记原有常识。结果是，当提示中包含该唯一标识符时，模型能够在不同场景和姿势中生成该人物的形象。DreamBooth 强调"将唯一标识符与特定对象绑定"，并在多样场景中保持其特征，常用于角色一致的生成需求。
- **面部ID引导**：近期亦有针对面部身份一致性的工作，如 ID-Booth（2025，Tomašević 等）。ID-Booth 提出在扩散训练中引入三元组身份损失，使生成图像在保持高多样性的同时，更好地与目标身份对齐。作者指出，他们的新框架"利用三元组身份训练目标实现了身份一致的图像生成，同时保留了预训练模型的生成能力"。虽然此类方法侧重于真实人脸，但思路可扩展到二次元角色，通过加入人脸识别网络损失来强化身份连续性。

- **Reference Attention**：一些工作将参考图像特征融入注意力机制以促进一致性。以 GeoDiffuser（2025）为例，他们在编辑网络中加入了**参考注意力层**。该机制使用参考图像的注意力查询来指导编辑注意力，从而确保背景和风格的一致保持。虽然 GeoDiffuser 主要用于几何编辑，但其提出的“参考注意力”（Reference Attention）和“编辑注意力”分支思想同样适用于多角色生成：即为每个角色或参考图分配独立的注意力流，保证角色特征仅从对应参考吸收信息。
- **Textual Prompting 多角色**：另外，DreamStory（2024）等新颖工作通过 LLM 引导提示生成来保持角色一致性。DreamStory 框架首先用大语言模型自动为长故事生成场景提示和角色描述，然后生成**角色肖像**作为多模态锚（anchors），最后在生成每帧场景时将这些锚输入扩散过程。其“Masked Mutual Self-Attention/Cross-Attention” 模块确保每个角色从对应的多模态信息中提取属性，从而避免角色互相融合。总体而言，通过对每个角色建立专门的参考（图像或文本），并在注意力层分离其信息流，可以有效提升多角色生成的身份和外观一致性。

5 长文本处理与分镜策略

对于“**AI一键小说生成漫画**”任务，需要将长篇小说（如 2000 字以上）划分成若干场景提示，然后逐帧生成画面。一个可行的处理流程包括：首先使用大语言模型（LLM，例如 GPT-4、Llama 等）对原始小说进行分段和摘要，将其转化为一系列结构化的场景描述（Scene Descriptions）。每个场景描述通常包括场景背景、关键事件、所含人物、情感语气等信息，作为对应漫画帧的提示。具体方法可以参考 DreamStory 等工作：使用 LLM 生成带有**场景与角色标注**的详细提示，并生成每个角色的画像或特征描述。

接下来，对每个场景提示进行图像生成。为了保证连贯性，可以采用**多轮迭代生成**策略。典型的做法是：先利用提示生成初步图像，然后利用类似 Story-Adapter的框架对生成结果进行反复优化。例如，Story-Adapter 提出一个“全局参考跨注意力”模块（global reference cross-attention），在每次迭代中将**所有之前生成的帧**作为全局上下文，通过跨注意力融合进当前生成，强化语义一致性。同时，每轮迭代也重新融入原始文本约束，从而逐步细化人物互动细节和场景连贯性。

另一种做法是**提示融合（Prompt Fusion）**：将分解出的多个提示（如场景提示、角色提示、动作提示）进行适当融合，形成复合提示引导生成。例如，可以将同一帧的多个人物关键词与环境描述拼接成一个综合 prompt，以保证图像中各元素合理共存。也可借鉴 compositional guidance 思想，对不同角色或物体施加单独权重或条件，再在生成过程中分阶段采样各部分内容。无论哪种方式，长文本到多帧图像的转换都强调**结构化提示**和**逐步迭代**：通过 LLM 自动分段摘要和多次生成-优化循环，可有效生成符合小说逻辑且风格统一的漫画分镜。

6 模块化漫画生成系统架构

针对任务要求，可设计一个模块化的漫画生成系统，包含以下主要组件（示意架构图见下方）：

- **文本解析模块**：接收原始小说文本，调用大语言模型进行场景切分和摘要，输出结构化的场景提示列表。
- **角色建模模块**：为小说中的重要角色生成或学习对应的概念向量或模型参数，包括使用 Textual Inversion、DreamBooth 等方法得到角色令牌，或者通过 LoRA 微调得到角色专属表示。这些角色向量与头像图像（若可获得）一起作为后续生成的输入。
- **画面控制模块**：处理场景布局和动作信息，可应用 ControlNet 等来加载场景草图或姿态图，以引导生成符合剧情的构图。例如，对话场景可用姿态骨架控制人物位置，对风景可用边缘图控制背景结构。
- **风格融合模块**：用于统一漫画风格或艺术效果，可能采用 LoRA 或 Style 模块，将目标风格注入扩散模型。也可使用 T2I-Adapter 以调节特定视觉特征（如线条粗细、着色风格）。
- **扩散生成核心**：以 Stable Diffusion / LDM 为基础的文本到图像生成引擎，结合上述模块输入的提示、角色向量和控制信号进行实际生成。模型应支持多条件融合，例如 CLIP 文本编码、跨注意力来自角色/风格向量、以及 ControlNet/T2I-Adapter 输出的特征。
- **分镜优化与一致性模块**：负责多帧间一致性，如 Story-Adapter 或 DreamStory 提出的迭代优化方法，实现角色外观和故事连贯的全局控制。它可能在每轮生成后评估并修正上一帧的细节，也可能在生成过程中融合全局语义嵌入。
- **系统交互与调度**：管理各模块数据流和资源，包括队列管理、硬件调度、以及用户交互界面（提示调整、选择分镜等）。

各模块通过清晰的接口交互：文本解析输出的场景提示传递给扩散引擎和控制模块；角色建模生成的向量提供给扩散核心；风格参数注入生成过程；分镜优化模块不断反馈以调整生成结果。整体架构遵循可插拔原则，允许在不同阶段加入新控制器（如新的 ControlNet 模型）或替换生成引擎（如使用 PixArt-α 进行特殊场景渲染）。通过这种模块化设计，系统既方便维护和扩展，也便于逐步优化各部分性能。

7 关键论文技术点评与适用性分析

- **DDPM（Ho 等，2020）**：提出了扩散模型的标准框架，包括前向噪声链和反向去噪网络。其详细的 ELBO 推导和噪声预测模型设计为后续研究奠定基础。该论文清晰给出正反过程形式和训练目标，是理解扩散模型数学原理的经典资源。
- **LDM（Rombach 等，2022）**：首次在潜在空间中训练扩散模型，以极大地降低了训练与推理成本。关键创新在于**图像-潜在空间**双阶段训练：先训练感知良好的自编码器，然后在其潜在向量上运行扩散模型。此方法保持了生成质量的同时，使高分辨率图像生成变得可行。对于本任务，LDM 代表了主流思路：我们也应在潜在空间进行扩散，以平衡质量与资源需求。

- **Stable Diffusion（CompVis 等，2022）**：在 LDM 基础上整合 CLIP 文本编码器，专门针对文本到图像生成进行了优化。其贡献在于提供了完整的开源模型和工具，使得大规模的预训练模型可在普通硬件上运行。实践中，我们可以直接使用 Stable Diffusion 作为生成核心，并利用其丰富的社区资源微调风格或人物。
- **Imagen（Saharia 等，2022）**：强调利用大型预训练语言模型提高文本条件的表达能力。虽然实际训练成本高，但揭示出“文本编码器规模对图像质量的关键作用”。对于漫画生成，我们可借鉴其经验：在可能的情况下采用更强大的文本模型或更好的文本提示，以改善长文本场景的理解和对齐。
- **PIXART-α（Li 等，2024）**：作为最近提出的高效 T2I Transformer，其强调**分步训练策略**和**轻量级架构**。其经验表明，通过巧妙的网络设计和训练分解，即使在资源有限的情况下也可接近 SOTA 性能。若资源受限，可考虑类似策略（如单独优化文本对齐和视觉细节），或者在不拥有大规模算力时借鉴其高效插入方法。
- **ControlNet（Wu 等，2023）**：其可插拔的侧链架构提供了一种强大的扩散条件控制方案。对于需要用户交互或严格布局控制的漫画创作任务，ControlNet 可通过边缘、姿态等提示精确指导构图。其设计也为后续研究提供思路：我们可以训练专用的 ControlNet 模型（如针对漫画线稿、角色姿势），以满足特定风格需求。
- **LoRA（Hu 等，2021；在 SD 应用见）**：低秩适配在微调上表现出极高效率。介绍了其原理：仅训练低秩矩阵极少参数即可调整大模型。用于漫画生成时，LoRA 可用于“轻量级角色/风格微调”，例如快速学习一个新角色的服装或背景风格，而无需大量数据或长时间训练。
- **T2I-Adapter（Zhang 等，2023）**：提出了一种灵活的控制模块，通过在 U-Net 不同阶段添加小型网络接收条件信号。此论文说明可以在不触碰主干网络的前提下实现多种控制。实践中，我们可借此设计自定义 Adapter（如颜色适配器、光照适配器等），满足漫画生成的多样化需求。
- **Visual Composer（Parmar 等，2025）**：其对象级组合生成方法引入了**KV-mixed**机制和组合引导。该工作对复杂多物体场景生成具有借鉴意义：我们可以采用类似策略，通过分对象控制与指导，提高漫画场景中多角色/多物体的一致性和准确性。
- **LogoSticker（Wang 等，2024）**：Token-Binding 概念在此得到体现。对于需要引入特殊元素（如标志、特殊纹样）的场景，此思路非常有用。在漫画中若出现特定符号或 Logo，可使用类似的标记绑定技术，确保生成时能精确再现细节。
- **Story-Adapter（Mao 等，2024）**：该工作针对长故事生成提出训练-free 的迭代方案。其“参考交叉注意力”模块可为我们的多帧一致性提供思路，例如在漫画生成中使用全球上下文来指导每帧。尽管 Story-Adapter 需要额外计算，其原理（使用全局先前生成信息）对提高连贯性非常有启发。
- **DreamStory（He 等，2024）**：结合 LLM 和多主体扩散的框架提供了一种端到端的故事可视化思路。其强调 LLM 在场景与角色规划中的作用，以及通过多模态锚点确保角色一致。该工作凸显了将大型语言模型用于提示生成的可行性，我们也可以借鉴其分阶段生成与参考图像引导方法来提升漫画叙事的整体连贯性。

8 技术路径建议与实现方案

针对“AI一键小说生成漫画”任务，可沿如下技术路径构建系统，并说明各步骤原理与可扩展性：

1. **文本预处理与分段**：使用大型预训练语言模型（如 GPT-4、LLaVA 等）对原始小说进行语义分析，将长文本自动拆分为逻辑连贯的场景。可以先进行段落级别的摘要，再细分关键事件。具体可采用链式调用：先请求模型给出故事大纲和章节划分（便于整体把控），再在每个章节内部生成分镜提示。该阶段主要依赖语言模型的理解和概括能力，具有高度可扩展性（更强模型可提高质量）。
2. **角色与元素建模**：分析文本中的主要角色、物品或风格，然后利用个性化生成技术创建这些概念的模型。例如，对每个主角可使用**Textual Inversion**或**DreamBooth**在 Stable Diffusion 上训练一个专属令牌，以便在后续每帧生成时复现该人物特征；同时，对漫画风格（如拟漫风格、配色方案）可使用 **LoRA** 做微调。角色建模需要准备少量示例图（3–10 张）并训练，过后生成过程只需引用对应令牌即可，这样确保跨帧角色外观一致且风格统一。
3. **提示工程与引导生成**：为每个场景构建综合提示（组合文本）。提示可以包含场景描述、角色令牌、情感语气以及控制信号。例如：**“夜晚森林中，两位主角（令牌A, 令牌B）并肩战斗，氛围阴森而紧张，蓝绿色调”**。如果需要更强的结构控制，可以利用 **ControlNet** 引入边缘图或姿态图：绘制每帧的草图轮廓（可由简单算法或手动绘制），并将其作为条件图输入，以确保角色位置和动作一致。同时，可用 **IP-Adapter** 注入参考图像或角色照片，以进一步保持人物细节。提示融合方面，可尝试将 LLM 输出的多个版本提示混合，或在迭代中更新提示以纠正生成偏差。
4. **生成模型与微调**：以 Stable Diffusion (或 Latent Diffusion) 作为基础生成模型。根据需求，可进行以下设置：
   - 决定是否在潜在空间或像素空间上生成：通常使用 LDM 潜在生成以效率和质量均衡。
   - 设定采样步骤和指导尺度：根据时效要求，可选用 DDIM 等快速采样方案；同时利用 Classifier-Free Guidance 或其他策略平衡创意性与文本一致性。
   - 模型微调：若已有特定漫画风格（如某漫画家风格），可用相似风格数据对模型进行 LoRA 微调或少量增量训练，以保证风格统一。
5. **迭代与一致性优化**：每帧生成后，可利用 **Story-Adapter** 或类似的策略对图像进行迭代优化：将已生成的前一帧或前几帧通过全局参考注意力模块反馈给当前帧的生成过程，以维持角色和场景的连贯。此外，可使用曝光融合、风格迁移等后处理技术统一整集风格。生成过程中可多次调用 LLM，要求其评价当前结果并提出修改建议，形成人机交互式迭代（类似 ChatGPT 辅助）。
6. **系统集成与可复现模块**：利用现有开源框架（如 Hugging Face Diffusers）组合上述模块。例如，Diffusers 可加载 Stable Diffusion、ControlNet、T2I-Adapter 等模型；LLM 可通过 API 调用。实现时应将各个自定义模块（角色令牌映射、场景提示生成、控制网络）封装为独立插件，以便替换和升级。重要的训练流程包括：角色令牌训练（Textual Inversion）、LoRA 微调、ControlNet 模型训练等，应设计良好的数据流水线和验证机制，以防过拟合或概念漂移。
7. **技术选型逻辑**：选择 Stable Diffusion 作为生成骨干是因为其高质量和开放性；引入大语言模型是因其在长文本理解上的优势；使用 ControlNet/IP-Adapter 等控制方法则能提供用户所需的可控性；采用 LoRA/TI/DreamBooth 等技术则兼顾了个性化与效率。每项技术的理论依据都是在已有工

作中验证的有效性：如大 LLM 可显著提升提示对齐性，多模态参考注意力可增强一致性。系统设计考虑可扩展性：如果将来需要加入视频生成或更细粒度风格控制，只需在相应阶段插入新的模型或算法即可。

8. **可扩展性与伦理考量**：系统架构应预留更新接口，比如未来可替换为更强的 T2I Transformer（如 PixArt-α）或新型注意力机制；对新角色的添加也可以通过增量训练快速实现。此外，应关注内容安全和版权问题：如对于具体小说内容或角色，需保证数据来源合法，并避免生成版权受保护的角色图像。

   $L(\theta) = E[\min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)]$