arXiv:2602.01305v1 [cs.CV] 1 Feb 2026

STORYSTATE：基于智能体的状态控制，实现一致且可编辑的故事书

Ayushman Sarkar[1], *Zhenyu Yu*[2,*], *Wei Tang*[3], *Chu Chen*[3], *Kangning Cui*[4], *Mohd Yamani Idna Idris*[2]

[1] 比尔布姆工程与技术学院，[2] 马来亚大学，[3] 香港城市大学，[4] 维克森林大学

摘要

大型多模态模型使一键式故事书生成成为可能：用户提供一段简短描述，即可获得一本包含多页插图的故事书。然而，底层故事状态（如角色、世界设定和页面级对象）仍然是隐式的，因此编辑粒度较粗，且往往会破坏视觉一致性。我们提出了*StoryS-tate*，这是一种基于智能体的编排层，在免训练的文本到图像生成之上引入了显式且可编辑的故事状态。*StoryState* 将每个故事表示为一个结构化对象，由角色表、全局设定和逐页场景约束组成，并利用一组小型 LLM 智能体来维护此状态，并生成用于创作和编辑的 1Prompt1Story 风格提示词。*StoryState*完全通过提示词运行，因此与模型无关，并兼容各种生成后端。针对多页编辑任务开展的系统级实验表明，与 1Prompt1Story 相比，*StoryState* 能够实现局部页面编辑、提升跨页面一致性，并减少非预期更改、交互轮次和编辑时间，同时达到接近 Gemini Storybook 一次生成时的一致性。代码见 https://github.com/YuZhenyuLindy/StoryState

索引词— 基于智能体的系统、交互式编辑、多模态、故事书、文本到图像

1. 引言

大型多模态模型的最新进展，让消费级应用能够实现“一键式”故事创作。Gemini Storybook 等系统只需一次前向传播 [1–4], ，就能将简短的文字描述（可选择加入图像）转化为包含插图和旁白的多页故事，大幅减少了手动编排故事板、提示工程和版面设计的需求。尽管方便，这类一键式系统本质上仍是一次性且基于版本的。其底层故事结构，包括角色属性、世界设定和各页面的构成，仍然隐含而不可访问。因此，用户通常只能通过自然语言请求进行编辑，而这会触发完全重新生成，往往导致多个页面出现意料之外的视觉变化 [5]。鉴于当前图像生成器的视觉一致性并不完美，即使是细微的编辑也可能造成累积的跨页偏移，例如角色外观不一致或道具缺失 [3,6]。我们将这一挑战称为故事状态控制：在保持跨页面视觉一致性的同时，实现精确的页面级编辑。

提升视觉叙事一致性的既有工作大致可分为模型级方法和免训练方法。模型级个性化方法，例如文本反演、DreamBooth、SuTI、The Chosen One 和 StoryMaker，通过使用特定角色或风格对生成模型进行微调来实现主体一致性 [7–11]。免训练方法，包括 ConsiStory、StoryDiffusion 和 One-Prompt-One-Story (1Prompt1Story)，则通过精心设计的提示词和推理时控制来提升一致性 [12–17]。尽管这些方法能够根据精心编写的提示词生成一致的视觉叙事，但它们主要在提示词层面 运行，对迭代式局部编辑的支持有限，而这对于完善多页故事至关重要 [18–20]。

在本研究中，我们提出*StoryState*，这是一种基于智能体的编排层，通过显式且持久的故事状态抽象增强一键式故事生成。*StoryState*通过三个结构化组件表示故事：(i) 捕捉持久视觉属性的角色表，(ii) 全局世界与风格设定，以及 (iii) 定义特定视觉元素的逐页场景约束。少量大语言模型智能体负责维护和更新此状态、生成用于生成与编辑的结构化提示词，并验证输出与已存储状态之间的一致性。通过将故事结构从生成模型中外置，*StoryState*能够在保留全局视觉一致性的同时，精确地进行页面级编辑，且无需修改或重新训练任何底层模型。我们的主要**贡献**如下：

- 我们提出一**种明确的故事状态表示**，支持局部页面级编辑，同时保持跨页面视觉一致性；
- 我们提出一**种基于智能体的编排流程**，将全局身份与页面特定内容分离，以实现一致的生成与编辑；
- 我们提出了***StoryState***，这是一个模型无关、无需训练的框架，仅通过结构化提示词与多模态后端交互。

2. 相关工作

视觉叙事系统。 近期的多模态系统支持通过单条提示词[1,2]一键生成故事书。这些系统提供端到端的生成流程，将创作复杂性抽象化。虽然使用方便，但这些系统通常采用一次性、基于版本的操作方式，对角色、布局或叙事流程等中间表示的控制有限 [5]。

一致性文本到图像生成。 通过个性化方法（例如 Textual Inversion、DreamBooth）在角色专属输入上微调模型 [7, 8], ，以及通过 ConsiStory 和 StoryDiffusion 等免训练方法在推理时调整提示词或生成约束 [12, 13]，研究者探索了如何在多图像生成中保持视觉一致性。这些技术虽能有效稳定静态叙事中的视觉属性，却假定提示词固定，且对结构化故事更新支持有限。

交互式生成与基于智能体的生成。 近期，交互式生成和规划与细化流水线的进展通过迭代推理或基于智能体的分解 [21, 22]，探索了用户控制与连贯性。然而，现有方法很少维护故事演变状态的显式持久表示，因此难以追踪编辑过程中的一致性，或随时间保持连贯性。

3. *StoryState*：基于智能体的状态控制

StoryState 在故事书风格的文本生成和免训练文本到图像（T2I）后端之上，引入由一小组基于大语言模型的智能体维护的显式、可编辑故事状态。*StoryState* 不再将故事结构隐式编码在模型激活或提示词中，而是将角色、世界设定和页面级场景表示为持久状态，作为文本和图像生成的唯一事实来源。智能体读取和写入此状态，以实现状态驱动生成和局部编辑。虽然我们的原型采用基于 Gemini 的文本智能体和 1Prompt1Story 风格的图像流水线，*StoryState* 与模型无关，并且无需重新训练即可兼容任何基于提示词的文本或图像生成后端。

3.1. *StoryState*构建

故事状态表示。 每个故事都由结构化状态表示

$$S = (C, W, \{S_i\}_{i=1}^{N}), \tag{1}$$

其中 N 为页面数量。角色表 C 是由角色条目组成的集合，每个条目包含：(i) 名称和叙事角色，(ii) 持久视觉属性（例如物种、年龄、外观、服装），以及 (iii) 可选的参考图像。世界设定 W 编码了页面间共享的全局元素，包括风格、基调以及反复出现的地点或道具。每个页面状态 S_i 编码页面专属信息，包括：(i) 简短的场景描述，(ii) 该页面中的角色（链接至 C），(iii) 明确的视觉约束（例如“与第 1 页相同的黄色雨衣”或“电视在左侧”），以及 (iv) 指向相关文本和图像素材的指针。

在内部， S 以一种可由智能体检查和修改的类 JSON 结构存储。重要的是，这种表示支持局部状态更新：编辑通过修改（C、W、$\{S_i\}$）中的最小子集来表达，而所有无关组件保持不变。不同于默认的 Gemini Storybook 工作流（其中此类信息隐含在模型内部，并在每次修订期间重新生成），*StoryState*会在整个编辑周期中显式且持久地维护这一表示 [1, 5]。

规划智能体。 给定初始用户提示（例如“一个害羞的男孩在城市里发现了一个迷路的机器人”），规划智能体会将其分解为页面级大纲。该大纲规定了由 N 个页面组成的序列，其中包含场景描述、高层叙事流程（例如开端、高潮、结局）以及初始的角色-页面分配。规划智能体输出一种机器可读的结构，用于初始化页面状态 $\{S_i\}$，为后续生成和编辑提供基础。

状态管理智能体。 状态管理器接收规划器的输出，并构建一致的角色和世界表征， C 和 W。当大纲中的多个指称（例如“一个男孩”“蒂姆”和“那个孩子”）指向同一实体时，状态管理器会将它们统一为一个角色条目，并为其分配稳定的视觉属性。它还会记录身份不变量，例如“除非明确更改，蒂姆始终穿黄色雨衣”，并在生成和编辑过程中将其视为持久约束。这种设计将存储在 C中的全局身份信息，与存储在 $\{S_i\}$中的页面特定场景语义解耦。通过将含糊或冗余的指称规范化为规范状态表征，状态管理器确保身份属性在各页面间保持一致，同时允许页面层面的语义独立变化。统一的故事状态 S 成为核心对象，所有智能体在整个生成和编辑过程中都从中读取信息并向其中写入信息。

3.2. 生成流程

构建故事状态 S 后，文本和图像生成通过两个相互协调的分支进行，共享同一状态表示。两个分支都将生成视为以 S 为条件的确定性映射，以确保相同状态产生一致的输出。

文本智能体。 对于每个页面 i，文本智能体都会读取页面状态 S_i 以及全局上下文（C、W），并生成该页面的最终叙述文本。在初次创作故事时，此过程会应用于所有页面 $i = 1, \ldots\ldots, N$。在编辑过程中，如果页面状态 S_i 被修改（例如调整基调或长度），文本智能体只会重新生成受影响的页面，而



图 1：*StoryState* 工作流概览。系统从用户提示开始，构建显式故事状态，其中包含角色表、全局世界设定和逐页场景描述。由大语言模型智能体组成的团队会迭代构建并更新此状态，并通过结构化提示驱动文本和图像生成。通过持续维护此状态，*StoryState* 无需修改底层模型，即可实现局部编辑和跨页一致性。

其他页面的叙述文本保持不变。这种选择性重新生成确保文本编辑的影响局限于局部且可预测。

提示词编写智能体和文本到图像后端。 对于图像生成，*StoryState* 连接到支持 1Prompt1Story 式提示 [12–14]的免训练文本到图像后端。提示词编写智能体将完整状态 $S = (C,\ W,\ \{S_i\})$ 转换为一组结构化提示，其中包括全局身份提示 P_0 和页面专属提示 $\{P_i\}_{i=1}^{N}$。后端应用奇异值重加权和身份保持交叉注意力等技术，在保留逐帧语义 [14] 的同时强化身份特征。通过将共享身份信息集中于 P_0，并将特定场景内容隔离到每个 P_i中，*StoryState*减少了页面间的语义冗余，并缓解了全局视觉漂移。

3.3. 编辑与一致性检查

通过状态更新进行局部与全局编辑。 与依赖全故事重新生成来应对任何更改的系统不同， [5], *StoryState* 仅通过修改状态中的相关组件来执行编辑 S。对于页面j 上的局部视觉修订（例如，“在第 3 页，Lily 应穿着与第 1 页相同的黄色外套，且电视应位于左侧”），状态管理器会更新 S_j 以反映新的约束。随后，提示词编写器仅重新计算对应的页面提示词P_j，同时保持 P_0 及所有其他 P_i 不变，文本到图像后端则为页面 j重新生成图像。

对于全局身份编辑（例如，“Lily 在整个故事中都有绿色眼睛”），状态管理器会更新角色设定表 C。提示词编写器据此重写全局提示词 P_0 以及受影响的 $\{P_i\}$，后端则有选择地仅重新生成受影响的页面。文本智能体遵循相同原则，仅在相应的 S_i 发生变化时更新叙述文本。该设计确保编辑精确地且仅在预期之处传播。

一致性评审智能体。重新生成后，一致性评审智能体会检查更新后的内容是否符合当前故事状态 S，并在必要时核对相邻页面。评审智能体通过对文本和图像进行多模态推理，检查是否存在与 C不符的属性、相对于 W 和 S_i缺失或位置错误的元素，以及布局违规。检测到差异时，评审智能体会生成结构化反馈，建议对 S_j 进行最小幅度的修正。如果建议获采纳，系统便会更新 S_j，并重新执行该页面的生成流程。这一反馈循环使 *StoryState*能够检测并修复跨页不一致，而无需进行模型再训练。由于所有控制均通过结构化提示实现，该框架仍兼容多种文本到图像后端和大语言模型。

4. 实验

4.1. 实现细节

数据集。 我们构建了一个包含192本十页插图故事书的数据集，遵循 ConsiStory+ 基准协议 [14]。每个故事均由 GPT-5.2 根据提示词生成，提示词指定了主角、目标视觉风格和简洁的叙事大纲。对于每本基础故事书，我们还定义了多个页面级编辑请求，以模拟真实的修订场景。这使我们能够评估迭代编辑性能。附录A.1提供了代表性故事和视觉示例。

设置。 我们比较三种方法：(i) *GeminiStorybook*[1], 使用 Gemini 的官方一键生成工具，不支持迭代编辑 [5]; (ii)*1Prompt1Story* [14], 根据我们的提示词格式调整的免训练文本到图像流程；以及 (iii) *StoryState*。所有方法均使用相同的基础提示和目标页数。Gemini Storybook 仅评估一致性，而 1Prompt1Story 和 *StoryState*还会进一步评估交互式编辑。对于*StoryState*，我们初始化状态 S，生成故事，并通过修改 S 中的相关组件并重新生成受影响的页面来执行编辑。



图 2：不同故事编辑框架之间的定性比较。每一行展示了由 Gemini Storybook、1Prompt1Story 或 *StoryState* 生成的 10 个页面。Gemini 在编辑时需要完全重新生成；1Prompt1Story 提升了一致性，但限制了动作与姿态的多样性；*StoryState* 支持局部编辑，同时保留跨页面的身份一致性与视觉变化。

指标。 我们报告：(i)视觉一致性，以相邻页面之间 CLIP 嵌入的平均余弦相似度衡量（越高越好）；(ii)更改页数，即单次编辑后被修改的图像页面平均数量（越低越好）；以及 (iii) 用户投入，通过用户交互轮次和每次编辑的挂钟时间来衡量（越低越好）。在用户研究中，我们还会收集关于感知一致性和控制的 5 点李克特评分。

4.2. 与基线比较

定量比较。 如表 1 所示，Gemini 故事书可作为一键式参考，并取得了最高的一致性得分（0.89），但不支持迭代编辑，因此未对交互性指标进行评估。在两种交互式方法中，*StoryState* 的一致性优于 1Prompt1Story（0.83 对 0.78），同时更改的页面更少（1.6 对 4.5）、用户轮次更少（3.1 对 4.3），每次编辑耗时也更短（74 秒对 96 秒）。这些结果表明，显式状态表示和基于智能体的提示控制有助于实现局部且高效的编辑。**定量比较**。 图 2 对比了三个模型生成的两个故事示例（提示词见附录 A.2）。

Gemini Storybook 通过重新生成整本书来应用编辑，往往会导致其他页面发生意外更改。1Prompt1Story 提升了角色身份的一致性，但动作与姿态的多样性有限。相比之下，*StoryState* 在保持角色身份一致的同时，也能在不同页面呈现多样的姿态和场景构图。其基于状态的控制可针对特定页面状态进行定向编辑，而不会意外修改其余页面。

4.3. 用户研究

为评估感知到的视觉质量和编辑可控性，我们邀请了 $N = 100$ 位参与者（熟悉绘本的非专业人士）开展用户研究。针对每个故事提示词，我们分别使用 Gemini Storybook、1Prompt1Story 和 *StoryState* 生成三个版本。参与者会看到全部三个版本，并回答：（i）哪个故事在各页面之间看起来更加一致；（ii）哪个系统在编辑过程中提供了更大的控制权。参与者还会使用 5 点李克特量表对每个故事的整体质量进行评分。本节重点关注关于一致性和控制的成对偏好。

如图3所示，在一致性方面，*StoryState* 在36%的比较中更受青睐，略高于 Gemini

表 1：在一致性和编辑效率方面的系统级比较。*StoryState*在保持高一致性的同时，以更少的更改页数、更少的轮次和更短的每次编辑时间实现了较高的编辑效率。Gemini Storybook在一致性方面得分最高，但缺乏交互式编辑能力。

方法	一致性↑	更改页数↓	轮次↓	时间（秒）↓
Gemini 故事书	**0.89**	-	-	-
1Prompt1Story	0.78	4.5	4.3	96
StoryState（我们的方法）	0.83	**1.6**	**3.1**	**74**

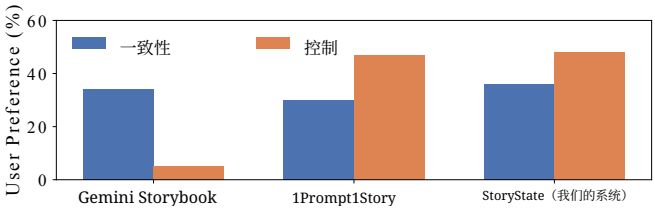


图3：对三种故事编辑系统的一致性和可控性进行的用户研究。*StoryState* 在可控性方面最受青睐（48%），一致性偏好度也最高（36%）。详细结果见表A.4。

Storybook（34%）和 1Prompt1Story（30%）（见表A.4）。在感知控制方面，*StoryState* 再次以48%领先，紧随其后的是1Prompt1Story（47%），而 Gemini Storybook 仅在5%的情况下被选中。这些结果表明，参与者认为*StoryState* 是最可控的系统，同时又不损害一致性。尤其是*StoryState*和1Prompt1Story 在控制方面的出色表现，凸显了智能体驱动的编辑和结构化提示词相较于 Gemini 单次生成界面的优势。

4.4. 消融研究

我们进行了消融研究，以分别考察*StoryState*中三个关键组件的影响：(i) 移除显式故事状态表示，并仅通过提示词重新生成来应用编辑；(ii) 禁用页面级重新生成，同时保留状态；以及 (iii) 删除一致性批评器。

如表2所示，移除显式状态会导致更改的页面大幅增加，用户投入也随之增加，这表明持久化的结构化表示对于局部编辑至关重要。禁用页面级重新生成虽然能保持一致性，但每次编辑都会导致故事几乎完全重新生成，从而显著增加更改页数和编辑时间。排除一致性批评器后，量化得分相近，但在定性检查中，身份和布局违规出现得更频繁，尤其是在连续进行多次编辑时。

这些发现证实，*StoryState* 所带来的优势源于持久状态跟踪、针对性的页面级更新和结构化一致性检查的结合，而不仅仅是提示工程。

表2：*StoryState* 组件的消融研究。显式状态表示与页面级重新生成可显著减少更改页数和用户投入。移除一致性评判器的影响较小，但会降低一致性。

变体	一致性↑	更改页数↓	轮次↓	时间（秒）↓
完整 *StoryState*	0.83	**1.6**	**3.1**	**74**
无显式状态	0.78	4.4	4.2	95
无页面级重新生成	0.83	9.8	3.0	132
无一致性批评器	0.81	1.7	3.1	75

表 3：故事编辑框架结构对比。*StoryState* 引入了显式、持久的状态，并支持逐页编辑，不同于基于版本或仅提示词的系统。

框架	显式故事状态	编辑单元
Gemini Storybook	✗	版本
ConsiStory / StoryDiffusion	✗	-
1Prompt1Story	✗	提示词
StoryState（我们的）	✓	Page

5. 讨论

*StoryState*通过从仅依赖提示词或基于版本的工作流转向结构化的状态驱动范式，重新思考了故事书的生成与编辑方式。不同于现有系统依赖不透明提示词运行，或为进行细微编辑而重新生成整个故事，*StoryState* 将叙事骨架（角色、全局设定和逐页细节）外化为一种持久且可编辑的表示形式（表 3）。

这种设计让身份保持、细粒度编辑和一致性检查等多项核心操作不仅切实可行，而且透明、可控。为此，*StoryState*将故事创作从一次性生成重新构想为渐进式、以状态为中心的创作。它在完全依赖手工制作的流程与黑盒端到端生成之间提供了折中方案，在长篇多模态内容创作中平衡自动化与用户控制。

6. 结论

本文介绍了*StoryState*，这是一种基于智能体的状态控制器，旨在实现故事书的一致性与可编辑性。通过将角色、设置和页面级场景显式建模为结构化、可编辑的状态，并将其与提示驱动生成相结合，*StoryState*支持精确的页面级编辑，同时保持跨页面视觉一致性。实验表明，*StoryState* 能够减少非预期的变化和用户投入，同时保持模型无关。除了故事书之外，基于状态的控制范式还为管理长时程多模态生成并进行局部编辑提供了一种通用方法。未来的工作包括扩展状态以涵盖更丰富的叙事结构、支持更细粒度的对象级或区域级编辑，并将 StoryState 式编排策略应用于视频生成和交互式媒体。

7. 参考文献

[1] 谷歌，“Gemini 故事书——只属于你的奇思妙想故事”，https://gemini.google/ overview/storybook/，2025，访问日期：2025- 11-15。

[2]谷歌，“在 Gemini 应用中创建个性化插画故事书”，https://blog.google/ products/gemini/storybooks/，2025，访问日期：2025-11-15。

[3]Zhedong Lin、Zhongsheng Wang、Qian Liu、Xinyu Zhang 和 Jiamou Liu，“叙事学与文本生成图像：人工智能生成故事书插画一致性的综述”，《人工智能评论》，2026。

[4] Zhenyu Yu、Mohd Yamani Idna Idris、Pei Wang 和 Rizwan Qureshi，“Cotextor：通过分层解耦和深度感知融合实现免训练的模块化多语言文本编辑”，载于第三十九届神经信息处理系统年会创意人工智能专题：人类，2025。

[5]谷歌，“在 Gemini 应用中创建插画故事书”，https://support.google.com/ gemini/answer/16434396，2025，帮助中心文章，访问日期：2025-11-15。

[6] Jay Peters，《谷歌 Gemini 现在可以创作 AI 生成的睡前故事》，*TheVerge*，2025，网络文章，访问日期：2025-11-15。

[7] Rinon Gal、Yuval Alaluf、Yuval Atzmon、Or Patashnik、Amit H. Bermano、Gal Chechik 和 Daniel Cohen-Or，《一张图像胜过一个词：利用文本反演实现个性化文本到图像生成》，载于*ICLR*，2023。

[8]Nataniel Ruiz、Yuanzhen Li、Varun Jampani、Yael Pritch、Michael Rubinstein 和 Kfir Aberman，《DreamBooth：微调文本生成图像扩散模型以实现主体驱动生成》，载于*CVPR*，2023，第 22500–22510 页。

[9] Wenhu Chen、Hexiang Hu、Yandong Li、Nataniel Ruiz、Xuhui Jia、Ming-Wei Chang 和 William W. Cohen，《通过学徒式学习实现主体驱动的文本到图像生成》，载于 *NeurIPS*，2023。

[10] Omri Avrahami、Amir Hertz、Yael Vinker、Moab Arar、Shlomi Fruchter、Ohad Fried、Daniel Cohen-Or 和 Dani Lischinski，“The chosen one：文本生成图像扩散模型中的一致性角色”，载于*ACMSIG-GRAPH*，2024。

[11] Zhengguang Zhou、Jing Li、Huaxia Li、Nemo Chen与Xu Tang，“StoryMaker：迈向整体一致性文本到图像生成中的角色，”*arXiv*预印本 *arXiv:2409.12576*，2024。[12] Yoad Tewel、Omri Kaduri、Rinon Gal、Yoni Kasten、Lior Wolf、Gal Chechik与Yuval Atzmon，“免训练一致性文生图，”*ACM*图形学汇刊，2024。[13] Yupeng Zhou、Daquan Zhou、程明明、冯佳时与侯淇彬，“StoryDiffusion：面向长程图像与视频生成的一致性自注意力，”*NeurIPS*，2024。

[14] Tao Liu、Kai Wang、Senmao Li、Joost van de Weijer、Fahad Shahbaz Khan、Shiqi Yang、Yaxing Wang、Jian Yang 和 程明明，“One-Prompt-One-Story：使用单个提示词实现零成本的一致性文本到图像生成”，载于 *ICLR*，2025。

[15] Mingxiao Li、Mang Ning 和 Marie-Francine Moens，“采用非对称之字形采样的一致性故事生成”，*arXiv*预印本 *arXiv:2506.09612*，2025。

[16] Sibo Dong、Ismail Shaheen、Maggie Shen、Rupayan Mallick 和 Sarah Adel Bargal，“ViSTA：使用多模态适配器进行文本生成图像扩散模型的视觉故事讲述”，*arXiv*预印本 *arXiv:2506.12198*，2025。

[17] Guanying Wang、Huiguo He、Jiawei Mao、Zhicheng Zhang、NamLing 和 JunkaiYu，《Storyanchors：通过一致性身份与布局锚定实现长篇故事可视化》，*arXiv* 预印本 *arXiv:2501.03456*，2025。

[18] Huiguo He、Guanying Wang、Zhicheng Zhang、Cheng Zhang、Han Li、Shiqi Wang、Nam Ling 和 Junkai Yu，《Dreamstory：通过 LLM 引导的多主体一致性扩散实现开放域故事可视化》，*IEEE* 模式分析与机器智能汇刊，2025。

[19] Jiawei Mao、Huiguo He、Guanying Wang、Junkai Yu 和 Nam Ling，《Story-adapter：用于长篇故事可视化的免训练迭代框架》，*arXiv* 预印本 *arXiv:2410.06244*，2024。

[20] Patrick Kwon 和 Chen Chen，《Dreamingcomics：通过使用视频模型进行主体和布局定制生成的故事可视化流水线》，*arXiv* 预印本 *arXiv:2512.01686*，2025。

[21] Jiaqi Gong、Shiwei Zhang、Hangjie Yuan 等人，《基于潜扩散模型的交互式视频生成》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集（*CVPR*），2023。

[22]Xiaoxin Lu、Haoran Zhang、Yusen Zhang 和 Rui Zhang，《增强文本图像计划生成的多模态一致性与连贯性》，载于计算语言学协会论文集（*ACL*），2025。

A.1. 故事书数据集

清单 1 展示了故事书数据集中的一个具体示例。每个数据实例代表一个由十帧组成的故事，其中包括一个全局身份提示（身份提示）和一系列帧级提示（帧提示列表）。身份提示指定主体的核心视觉属性，并在所有帧中保持固定；而每个帧提示则描述一个连贯的场景，其中动作、环境或叙事状态各不相同。

该数据集旨在通过明确确保身份一致性和结构化的时间推进，支持长时程视觉叙事。如清单 1 所示，帧提示按照自然的叙事流程进行组织，包括场景引入、环境发展和结局。这种结构化设计有助于受控的多帧生成，并能够系统地评估基于故事的图像生成任务中的身份保持、时间连贯性和提示-图像对齐。

清单 1：数据集示例。

--id_prompt "一幅火凤凰的壮丽插画，它有着明亮的橙色羽毛。" --frame_prompt_list "一只长着明亮橙色羽毛的凤凰，从熊熊燃烧的灰烬中升起，引入场景；画面生动而连贯，采用电影感光照，空间背景清晰

且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，翱翔于发光的天空中，引入场景；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，栖息在山巅，展现场景中的环境发展；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，唱着一曲令人心悸的旋律，展现场景中的环境发展；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，用双翼点燃火焰，展现场景中的环境发展；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，从熊熊燃烧的灰烬中升起，展现场景中的环境发展；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，翱翔于发光的天空中，为场景带来平静的结局；画面生动而连贯，采用电影感光照，空间背景清晰，且具有很强的视觉连续性" "一只长着明亮橙色羽毛的凤凰，栖息在山巅，为场景带

平静的结局，呈现为一个生动而连贯的瞬间，配以电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着鲜艳橙色羽毛的凤凰，唱着一首令人难忘的旋律，将场景带向

平静的结局，呈现为一个生动而连贯的瞬间，配以电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着鲜艳橙色羽毛的凤凰，用翅膀点燃火焰，将场景带向平静的结局，呈现为一个生动而连贯的瞬间，配以电影感光照、清晰的空间背景和强烈的视觉连续性"

--id_提示词 "一幅鲜明夺目的肖像，描绘一只带有黑白条纹的斑马。" --frame_提示词_列表

"一只身披黑白条纹的斑马，在河流旁吃草，作为场景的引入；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在兽群中奔跑，穿越平原，作为场景的引入；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在金合欢树的树荫下休息，展现环境；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性"

"一只身披黑白条纹的斑马，在野外穿过一条尘土路，展现环境；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在热带草原上保护幼崽，展现环境；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在河流旁吃草，展现环境；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在兽群中奔跑，穿越平原，为场景带来平静的结局；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在金合欢树的树荫下休息，为场景带来平静的结局；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在野外穿过一条尘土路，为场景带来平静的结局；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马，在热带草原上保护幼崽，为场景带来平静的结局；画面生动而连贯，具有电影感光照、清晰的空间背景和强烈的视觉连续性"

--id_prompt "一只目光锐利、身形矫健敏捷的猎豹。" --frame_prompt_list "一只目光锐利的猎豹，正在热带草原上飞奔，场景由此展开；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在草地上悄然追踪一只瞪羚，场景由此展开；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在树下的阴凉处休憩，环境逐渐展现；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在用气味标记领地，环境逐渐展现；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，成功狩猎后正在打盹，环境逐渐展现；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在热带草原上飞奔，环境逐渐展现；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在草地上悄然追踪一只瞪羚，为场景带来平静的结局；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在树下的阴凉处休憩，为场景带来平静的结局；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，正在用气味标记领地，为场景带来平静的结局；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强" "一只目光锐利的猎豹，成功狩猎后正在打盹，为场景带来平静的结局；画面生动连贯，采用电影感光照，空间背景清晰，视觉连续性强"

--id_提示词 "一幅气势磅礴、庄严宏伟的画面，描绘一头长着长獠牙的海象。" --frame_提示词_列表 "一头长着长獠牙的海象悠闲地躺在一块浮冰上，引出场景；画面捕捉了一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的

视觉连续性" "一头长着长獠牙的海象正在海中潜水寻找蛤蜊，引出场景，画面捕捉了一个

s

一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在海岸上咆哮，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在积雪海岸线附近休息，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在海洋中优雅地游泳，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象悠闲地躺在一块浮冰上，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象潜入海中寻找蛤蜊，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在海岸上咆哮，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在积雪海岸线附近休息，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只长着巨大獠牙的海象在海洋中优雅地游泳，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" --id_prompt "一幅大胆醒目的插画，描绘一只身披黑白条纹的斑马。" --frame_prompt_list "一只身披黑白条纹的斑马在热带草原的阳光下吃草，开启场景，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马与一群兽群一起奔跑在草原上，开启场景，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马站在水坑附近，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马在荒野中穿过河流，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马在一头角马旁边小跑，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马在热带草原的阳光下吃草，逐步展现环境，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马与一群兽群一起奔跑在草原上，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马站在水坑附近，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马在荒野中穿过河流，让场景以平静的结局收尾，呈现为一个生动连贯的瞬间，具有电影感光照、清晰的空间背景和强烈的视觉连续性" "一只身披黑白条纹的斑马在一头角马旁边小跑，让场景以

一个平静的结局，以生动而连贯的瞬间呈现，具有电影感光照、清晰的空间背景和强烈的视觉连续性”

A.2. 对比提示词 Y

本节列出了我们实验中对比方法所使用的提示词（见列表 2）。为确保评估公平且可复现，所有基线方法均以相同的故事提示词为条件，这些提示词定义了主角身份和叙事场景序列。这些提示词直接用作相应对比流程的输入，不进行任何人工调优或针对特定方法的修改。

对比提示词采用与我们的方法相同的结构化格式，由固定的身份提示和一系列帧级提示组成。通过在所有方法中统一提示词内容，我们将性能差异归因于建模和生成机制，而非提示工程。本节列出了完整的提示词示例，以促进透明度、可复现性和未来的基准测试。

列表 2：数据集示例。

--id_prompt "Fira，一只好奇的狐狸，长着橙色皮毛和明亮的绿眼睛，脖子上围着一条小红围巾。" --frame_prompt_list "冬季的第一片雪花飘落在窗外时，她在舒适的树屋中醒来" "她走进一片白雪覆盖的空地，红围巾随风飘动，探索着宁静的森林" "她堆起一只小雪狐狸，看着神奇的雪精灵在空中翩翩起舞" "她沿着溪流漫步，积雪开始融化，娇小的春花也开始绽放" "她遇见一只发光的蝴蝶，它带领她来到一片隐秘的草甸，那里鲜花盛开" "她在樱花树下与其他动物嬉戏，春风轻柔地吹拂" "夏日里，她跳进一处清凉的池塘，池塘四周环绕着发光的芦苇和友善的青蛙" "温暖的夏夜里，她看着萤火虫在神奇的森林上空盘旋" "她沿着金色的小径漫步，秋叶飘落，红围巾与树木的颜色融为一体" "冬季再次来临，她蜷缩在树屋中，望着飘落的雪花，脸上带着平和的微笑"

--id_提示词 "一幅超写实的数字绘画，描绘一名留着短而竖直的黑发、身材略显单薄、拥有深棕色眼睛的少年。"

--frame_prompt_list "戴着绿色针织帽，静静站立，柔和的光线洒过他的脸庞" "在被混凝土坡道和涂鸦墙环绕的城市滑板公园中消磨时光" "穿过一片有着高大树木、阳光透过枝叶洒落的茂密森林" "身穿宇航员服装，仿佛已准备好踏上超越地球的想象之旅" "跪在地上，用一把小手铲仔细挖掘" "爬上一棵树，紧抓树枝，树叶在他周围沙沙作响" "参加一场热闹的科学展览会，现场摆满海报和模型，还有好奇的参观者" "手捧一杯热饮，蒸汽轻柔升起，享受片刻宁静" "在画布上作画，全神贯注于创作过程" "平静地坐着读书，沉浸在思考与想象之中"

A.3. 示例：*StoryState*

本节展示由 *StoryState* 生成的定性示例，以说明其在长时程视觉叙事中的表现（见图 A.4 和 A.5）。每个示例都展示了 *StoryState*如何在逐步更新底层状态，以反映跨帧的动作、环境和叙事背景变化的同时，保持对主要角色的一致表征。与将多帧生成视为彼此独立或弱耦合过程的基线方法不同，*StoryState*明确将生成建模为一个状态依赖过程。如示例所示，状态转移能够实现连贯的时间演进，避免身份漂移或视觉退化，即使是在较长的故事序列中也是如此。这些定性结果补充了正文中的定量分析，并直观地展示了状态感知推理如何促进稳定且可控的故事生成。



图 A.4：*StoryState*生成结果示例。



图 A.5：*StoryState*生成结果示例。

A.4. 附加表格

表A.4报告了关于感知视觉一致性和编辑可控性的用户研究详细结果。参与者需要比较由不同系统生成的故事书，并选择他们认为跨页面一致性更高且在编辑过程中更易于控制的故事书。百分比表示在两两比较中各系统被偏好的次数比例。如表中所示，*StoryState* 在一致性和控制两方面均获得了最高的偏好率，这表明明确的故事状态表示和基于智能体的编排能够提升用户感知的编辑体验。

表A.4：用户研究结果。**粗体**数值表示最佳性能。

系统	**一致性**↑	**控制**↑
Gemini Storybook	34%	5%
1Prompt1Story	30%	47%
StoryState（我们的）	**36%**	**48%**