

ToonComposer: 利用 生成后关键帧简化卡通制作

Lingen Li^{1,2} Guangzhi Wang^{2*†} Zhaoyang Zhang^{2*†} Yaowei Li^{3,2} Xiaoyu Li²

Qi Dou¹ Jinwei Gu¹ Tianfan Xue^{1†} Ying Shan²

¹The Chinese University of Hong Kong ²ARC Lab, Tencent PCG ³Peking University

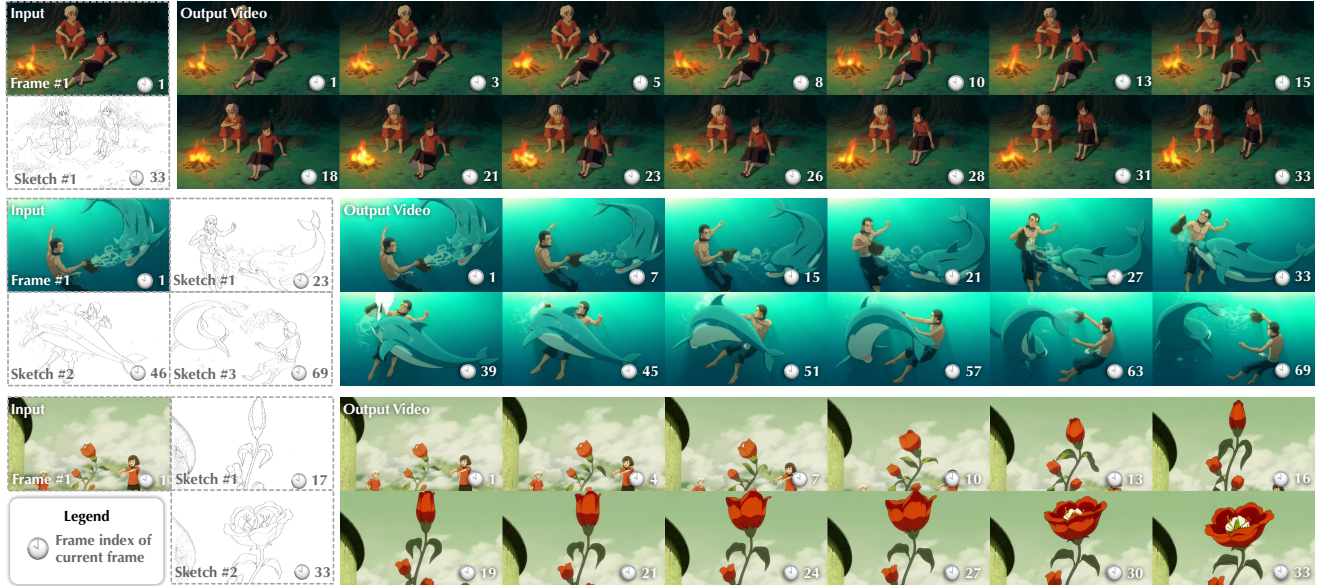


Figure 1. Video samples generated by ToonComposer using sparse keyframe sketches, featuring scenes from cartoon movies (Big Fish & Begonia, used with permission). ToonComposer supports precise keyframe control and flexible inference with varying numbers of input keyframe sketches and output video lengths (33 frames for the first and third samples, and 69 frames for the second sample). The frames are evenly sampled for illustration in this figure. Each input and output frame is annotated with its corresponding temporal index in the bottom right corner. These movies were excluded from the training data.

Abstract

传统的卡通和动画制作涉及关键帧、过渡帧和上色阶段，这些都需要大量的人工劳动。尽管在人工智能领域最近取得了一些进展，现有的方法通常将这些阶段分开处理，导致错误积累和瑕疵。例如，过渡帧方法在处理大幅度运动时存在困难，而上色方法则需要每帧的密集草图。为了解决这些问题，我们引入了ToonComposer，这是一种生成模型，将过渡帧和上色统一到单一的后关键帧阶段。ToonComposer使用一种稀疏草图注入机制，通过关键帧草图提供精确控制。此外，它采用了一种卡通适配方法，结合空间低秩适配器，在保持其时间先验完整的同时，将现代视频基础模型调整为卡通领域。ToonComposer 仅需一幅草图

和一个有色参考帧即可在稀疏输入下表现出色，同时也支持在任意时间位置使用多个草图以更精确地控制运动。这种双重能力减少了人工工作量并提高了灵活性，使得艺术家在现实场景中更具能力。为了评估我们的模型，我们进一步创建了 PKBench，这是一项包含人类绘制草图的基准，模拟现实使用案例。我们的评估表明，ToonComposer 在视觉质量、运动一致性和生产效率方面均优于现有方法，为人工智能辅助卡通制作提供了一种更优越且灵活的解决方案。项目页面：<https://lg-li.github.io/project/tooncomposer>。

1. 介绍

卡通和动漫以其充满活力的美学和复杂的叙述而著称，是全球娱乐的基石。传统的卡通制作涉及关键帧、过渡帧和上色阶段，每个阶段都需要艺术家制作大量的帧，

* 项目负责人。

† 通讯作者。

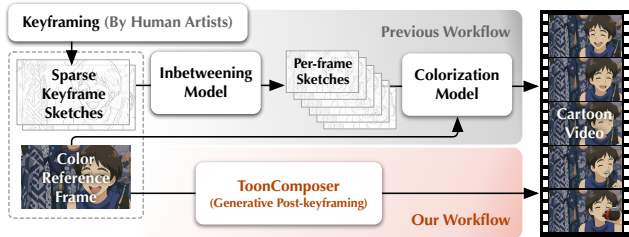


Figure 2. 我们与之前动画制作流程的比较。ToonComposer 实现了 后关键帧处理 阶段，将动画中间帧绘制和上色无缝集成到一个自动化流程中，与传统和现有的 AI 辅助流程相比，简化了动画制作。

以确保流畅的运动和风格的一致性 [21]。虽然关键帧阶段是体现人类艺术创造力的过程，但后续的过渡帧和上色阶段却是高度劳动密集且耗时的。具体而言，过渡帧和上色阶段需要较少的创意输入，却要求为短短几秒钟的动画制作数百张图稿，导致显著的时间和资源成本。

生成模型的最新进展促进了中间补间和上色阶段的发展，比如 ToonCrafter [10, 27]、AniDoc [15] 或 LVCD [8]。然而，这些方法存在关键限制：(1) 插值方法 [10, 27] 难以从稀疏的草图输入中插值出大的动作，往往需要多个关键帧来获得平滑的动作；(2) 上色方法 [8, 15] 需要每帧详细的草图，加大了艺术家的工作负担。(3) 此外，它们的顺序处理导致误差累积，其中插值草图中的不准确会影响上色阶段，导致产生瑕疵和质量下降 [21]。这些缺点凸显了要实现高效的生产流程中，高质量结果的显著差距。

实际上，中间帧处理和上色阶段是高度相互依赖的。两者都需要在关键帧草图或颜色参考帧中寻找对应关系。因此，我们引入了 后关键帧处理 阶段，这是一种新的范式，它遵循关键帧创建阶段，并将中间帧处理和上色合并为单一的自动化过程。这种统一使模型能够在单一阶段中共同利用关键帧草图和参考帧中的基本和风格信息，避免跨阶段错误累积的风险。如 Figure 2 所示，后关键帧处理 阶段仅需少量的关键帧草图和一个上色的参考帧即可生成完整的高质量卡通视频。这种方法显著减少了人工工作量，让艺术家专注于创意关键帧设计，而 AI 负责管理重复性任务。

为此，我们采用现代扩散变压器 (DiT) 基础模型，[22]，显示出优越的视频生成性能。虽然它为卡通视频制作提供了新的可能性，但也提出了两个重要的挑战：1) 可控性：DiT 基础模型通常弱条件于文本提示或初始帧，缺乏将稀疏关键帧草图在特定时间位置用于运动引导的精确性。2) 适应性：由于它们是在自然视频数据集上训练的，适应卡通领域是生产高质量卡通视频所必需的。然而，之前针对卡通的适应方法 [27] 仅限于基于 UNet 的模型，只通过调整分离的空间层来更改模型的空间行为，同时保留时间先验。在我们的背景下，DiT 模型中的完整注意机制共同学习空间和时间行为，以前的卡通适应方法 [27] 不再有效。

为了解决上述挑战，我们提出了 ToonComposer，这是一种基于最先进的 DiT 视频基础模型 Wan 2.1 [22]

构建的生成性 后关键帧处理 模型。首先，为了解决稀疏草图精确可控性的问题，我们设计了稀疏草图注入机制，使得使用稀疏关键帧草图在卡通生成中实现精确控制。其次，为了解决 DiT 模型中的领域适应性问题，我们为 ToonComposer 量身定制了卡通适应策略。通过一种新颖的空间低秩适配器 (SLRA) 策略，将基础的 DiT 模型适应到卡通领域，在适应到卡通领域外观的同时，保持其强大的时间先验不变。此外，我们通过引入区域控制来提高 ToonComposer 的灵活性，使得在指定区域内无需绘制草图即可实现灵活的运动生成。这些贡献确保 ToonComposer 能够生成风格连贯的动画，所需输入最少，如 Figure 1 所示，有效地在一个模型内实现 后关键帧处理 阶段。

为了支持所提出模型的训练，我们策划了一个大型数据集 药代动力学数据，其中包含高质量的动漫和卡通视频片段。每个片段都附带有多种风格的关键帧素描，为训练提供了坚实的基础。除了在合成基准上评估我们的模型，我们还策划了 PKBench，这是一个新的基准，包含 30 个原创卡通场景，其中有人绘制的关键帧素描和参考彩图帧。在这两个基准上的大量实验表明，ToonComposer 在视觉质量、运动一致性和生产效率方面优于现有方法。

我们的贡献总结如下：

- 我们引入了 后关键帧处理 阶段，这是一种新的卡通制作范式，将中间补间和上色整合为由人工智能驱动的统一过程，显著减少了人工劳动。
- 我们提出了 ToonComposer，这是第一个基于 DiT 的用于 后关键帧处理 的卡通生成模型，结合稀疏草图注入和区域控制，从稀疏输入生成高质量卡通视频。
- 我们设计了一种使用 SLRA 的卡通适配机制，这是一种新颖的低秩适配策略，能够有效地调整 DiT 模型在卡通领域的空间行为，同时保留其时间先验。
- 我们整理了一个包含多样草图的卡通视频片段数据集用于训练，并开发了一个高质量的基准，PKBench，这是一套用于卡通 后关键帧处理 评估的真人手绘草图集。

人工智能越来越多地被应用于动画和动漫制作中自动化处理劳动密集型的任务，例如中间帧插入和上色。在中间帧插入方面，早期方法如 AnimeInterp [12] 和 AutoFI [18] 专注于线性和简单运动插值。最近，基于扩散的方法 [10, 27] 通过利用预训练模型的生成先验，能够处理具有更复杂运动的情况。在上色方面，早期基于 GAN 的方法 [9] 和最近基于扩散的方法 [8, 15, 34, 35] 已经实现了根据一个或一系列参考帧对线稿进行自动上色。

然而，虽然现有的 AI 辅助方法通过自动化中间帧和着色来推动卡通制作的发展，但它们往往需要密集的帧输入，作为独立的阶段操作，并在处理复杂动作和风格一致性方面面临挑战 [21]。ToonComposer 通过提供一个统一的稀疏输入解决方案来进行后关键帧处理，从而简化制作工作流程，克服了这些困难。

扩散模型已成为生成任务的基石，尤其是在图像和

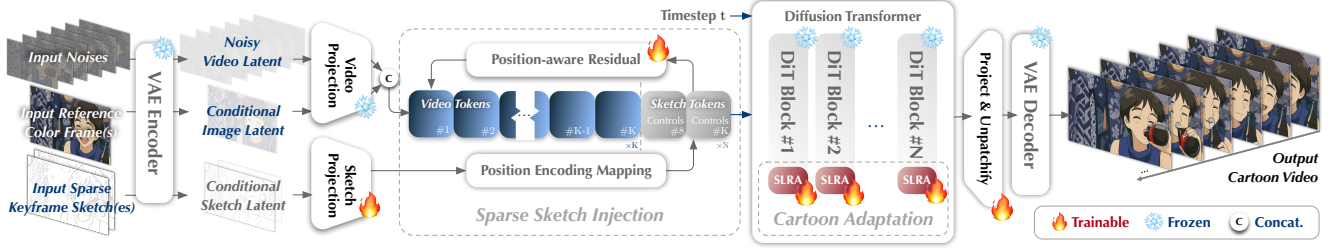


Figure 3. ToonComposer 的模型设计。稀疏草图注入机制通过关键帧草图实现精确控制，并且结合空间低秩适配器的卡通适应方法将基于 DiT 的视频基础模型调整至卡通领域，保留其时间先验。

视频合成方面，其通过迭代地对噪声分布中的样本进行去噪来产生高质量的输出。在视频生成中，这些模型必须有效地捕捉空间细节和时间动态，这一挑战促使出现了不同的架构方法。传统的基于 UNet 的扩散模型通过结合 3D 卷积和分离的空间与时间注意力层将二维 U-Net 扩展以处理视频。在这些模型中，空间注意力层处理帧内特征，通常在跨通道或空间位置上，而独立的时间注意力层则建模帧间依赖关系。相比之下，扩散变压器 (DiTs) 利用变压器架构，用完整的注意力机制替换 UNet 的卷积主干，以在空间和时间维度上建模长距离依赖。虽然与空间-时间解耦设计相比表现出更强性能，但这种完整的注意力机制消除了适合于特定领域如卡通的空间适应性。我们的工作建立在基于 DiT 的基础模型之上，以一种新的卡通适应机制来利用高质量视频先验，将 DiT 基础模型适应于卡通领域中的空间表现，同时保持其时间运动先验不变。

可控生成 可控生成旨在通过明确条件来引导图像和视频合成，例如参考图像 [30, 31]、深度图 [28]、人体姿态 [5, 33] 和语义标签。技术如 IP-Adapter [30] 和 ControlNet [31] 将这些视觉线索注入扩散模型中，与文本提示相结合，允许对内容和风格进行细粒度操作。生成的可控性价值在特定领域管道中尤为明显。对于电影摄制，摄像机感知生成器提供了 2D 场景布局和 3D 摄像机轨迹 [3, 13, 23, 25] 的操控手柄，使视频创作者能够以高精度构图和运动。在卡通制作中，草图引导模型支持插值、中间帧绘制和着色 [8, 15, 27]。我们的方法侧重于使用稀疏关键帧草图进行可控的卡通生成，以加速制作过程。

我们引入了 ToonComposer，这是一种新颖的生成后关键帧模型，能够在稀疏控制下生成高质量的卡通视频。为此，我们提出了一种精选的稀疏草图注入策略，该策略能够在任意时间点上实现精确的草图控制 (章节 1.1)。此外，为了充分利用视频生成模型中的时间先验，我们设计了一种新颖的低秩适应策略，使得在保持时间先验不变的同时，有效地将空间先验适应于卡通领域 (章节 1.2)。为了进一步减轻艺术家的工作负担并提高效率，我们的方法还实现了区域控制，允许艺术家只绘制部分草图，而由模型推理空白区域的运动应如何生成 (章节 1.3)。

近年来，卡通行业受益于生成式 AI 的发展，显著促

进了补间动画 [27] 和色彩化 [8, 15] 阶段的进步。虽然这些技术在卡通视频制作中非常有用，但现有的方法往往因高劳动力需求或视频质量不佳而遭遇瓶颈。例如，色彩化方法往往需要一个上色的参考帧和每帧的草图，而这些获得成本很高。尽管最近的补间方法 [27] 可以用于生成每帧的草图，但在处理大幅运动时仍面临挑战，导致在色彩化阶段出现误差累积。

事实上，这两个阶段是高度相互依赖的：两者都需要沿关键帧/草图中的元素之间的对应关系进行检索和插值，这表明它们的内部机制是相似的。受到此启发，我们提出了 后关键帧处理 阶段，一个新的阶段，自动化卡通制作，并将过渡和上色整合为一个统一的生成过程。给定一个上色的参考帧和一个草图帧，后关键帧处理 阶段旨在直接生成一个高质量的卡通视频，该视频依照这些输入提供的指导。这一过程显著减轻了每帧密集草图的需求，避免了跨阶段错误积累的风险。

形式上，给定一个带颜色的参考帧 f_1 和一个草图帧 s_j ，我们旨在获得一个模型 G_θ ，该模型可以直接生成具有 K 帧的高质量卡通视频：

$$\{\hat{f}_k\}_{k=1}^K = G_\theta(f_1, s_j, e_{text}) \quad (1)$$

其中 j 表示 s_j 的时间位置， e_{text} 表示描述场景的文本提示。

在这项工作中，我们采用了最近提出的强大视频生成模型 Wan [22] 作为基础。

1.1. 稀疏草图注入

高级视频生成模型，例如 Wan [22]，在生成高质量视频方面表现出色。虽然其图像到视频 (I2V) 变体支持通过初始帧指导的视频生成，但在任意时间位置使用稀疏草图进行精确控制仍未得到探索。

为此，我们引入了一种新颖的稀疏素描注入机制，该机制将素描无缝整合到 I2V DiT 模型的潜在令牌序列中，以实现精确的时间控制。在标准的 I2V DiT 模型 ϵ_θ 中，输入图像与沿通道维度的噪声潜在 z 连接：

$$\hat{\epsilon} = \epsilon_\theta \left(\left[\{z_k^{(t)}\}_{k=1}^K, \text{pad}(f_1^t) \right]_c, e_{text}, t \right), \quad (2)$$

其中 $\hat{\epsilon}$ 表示模型输出， $z_k^{(t)}$ 是时间步 t 的第 k 帧的噪声潜在令牌， $\text{pad}(\cdot)$ 表示在时间维度上的零填充， $[\cdot]_c$

表示在通道维度上的连接， f'_1 表示由 VAE 编码的 f_1 的潜在。我们的稀疏素描注入机制增强了模型 ϵ_θ ，以通过位置编码映射和位置感知残差机制支持使用稀疏素描和关键帧进行精确控制。

为了将草图帧 s_j 注入 DiT 模型 ϵ_θ 的潜在表示中，以实现生成卡通中时序位置 j 的精确控制，我们首先引入一个额外的投影头，该投影头将条件草图潜在向量嵌入到与模型潜在在维度兼容的草图令牌 s'_j 中。然后，我们应用位置嵌入映射过程，该过程在每个注意力操作之前从相应视频令牌的时间索引 j 中借用 RoPE [20] 编码。这些稀疏草图令牌与视频令牌沿序列维度串联，以便于注意力过程。

这种机制能够在生成过程中有效地将草图条件与时间意识整合到潜在空间中。此外，它还促进了多帧关键帧和草图作为控制输入的同时使用。鉴于某些动画场景中运动的复杂性，精确控制通常需要多个关键帧和草图。因此，我们扩展了公式以支持多个彩色参考帧和多个草图输入。因此，DiT 模型的前向步骤表示为：

$$\hat{\epsilon} = \epsilon_\theta \left(\left[\left[\{z_k^{(t)}\}_{k=1}^K, \text{pad}(\{f'_{ic}\}_{c=1}^C) \right]_c, \{s'_{in}\}_{n=1}^N \right]_s, e_{\text{text}}, t \right), \quad (3)$$

，其中 $\{s'_{in}\}_{n=1}^N$ 代表 N 草图帧， $\{f'_{ic}\}_{c=1}^C$ 表示 C 彩色参考帧。 $[\cdot, \cdot]_s$ 表示沿着令牌序列维度的连接。该公式允许对多个输入进行精确控制，同时也支持后关键帧处理阶段的最小输入要求（一个彩色和一个草图帧，在 Section 1 中描述）。

为了增强草图控制的灵活性，我们的 ToonComposer 还允许用户在推理过程中通过可调权重动态调整输入草图的控制力度。这是通过注入过程中的新位置感知残差模块实现的。对于在控制关键帧索引 $\{i_n\}_{n=1}^N$ 处的稀疏草图标记，我们通过线性层 W_{res} 转换这些标记，并将其与相同索引的视频标记相加，使用缩放权重 α ：

$$\{z_k^{(t)}\}_{k \in \{i_n\}_{n=1}^N} := \{z_k^{(t)}\}_{k \in \{i_n\}_{n=1}^N} + \alpha \{s'_{in}\}_{n=1}^N W_{res}, \quad (4)$$

其中， $W_{res} \in \mathbb{R}^{D \times D}$ 是可训练权重， D 是标记的特征维度。在训练期间，缩放权重 α 设置为 1。在推理期间，用户可以调整权重 α 以放松或加强关键帧草图的控制。

1.2. 卡通适应

先前的工作 [27] 已证明成功将视频生成模型适应于卡通领域。通过仅调整时空 U-Net 的空间层，保留了原始模型中的时间运动先验，同时将外观部分调整为卡通。然而，随着视频生成模型的发展，现代视频生成模型中广泛使用了 3D 全注意力 [11, 22, 29]，其中空间和时间表现在潜在空间中交织。因此，无法像以前的工作那样直接进行空间适应。

为了解决这个问题，我们引入了空间低秩适配器 (SLRA)，这是一种创新的低秩适配机制，旨在有效地将视频生成模型适配到卡通领域。与传统的 LoRA [4] 类似，我们的 SLRA 也包含两个可训练的矩阵 $\mathbf{W}^{\text{down}}$

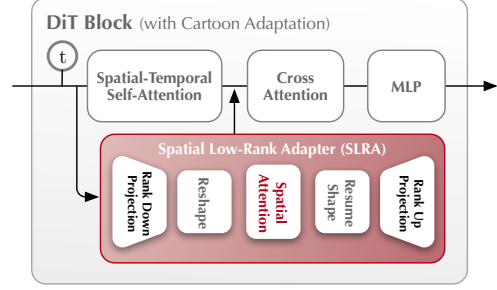


Figure 4. 用于 ToonComposer 中卡通适应的空间低秩适配器 (SLRA) 的结构。SLRA 将空间-时间自注意模块之前的隐藏状态作为输入，并输出一个在自注意操作后添加的残差。

和 $\mathbf{W}^{\text{up}}$ 。不同于传统的 LoRA，我们的 SLRA 仅改变注意力模块的空间行为，以保留基础模型中的时间先验。

设 H 和 W 为 DiT 的潜符号的空间大小， K 和 N 为视频符号和草图符号的时间长度。在 ϵ_θ 的每个自注意模块中给定一个符号序列 $h \in \mathbb{R}^{L \times D}$ ，其中 $L = (K + N) \times H \times W$ 为完整符号序列的长度，SLRA 首先通过使用线性层对输入隐藏符号的特征维度进行下采样操作：

$$h^{\text{low}} = h \mathbf{W}^{\text{down}}, \quad (5)$$

，其中 $\mathbf{W}^{\text{down}} \in \mathbb{R}^{D \times D'}$ $D' \ll D$ ，得到 $h^{\text{low}} \in \mathbb{R}^{L \times D'}$ 。

然后，SLRA 将 h^{low} 重塑为 $\tilde{h}^{\text{low}} \in \mathbb{R}^{[K+N] \times [H \times W] \times D'}$ ，恢复其原始的空间-时间排列。在重塑后，我们仅在空间维度上进行自注意力操作。这通过在每个帧的空间维度上独立执行注意力机制来实现：

$$Q = \tilde{h}_l^{\text{low}} \mathbf{W}_Q, \quad K = \tilde{h}_l^{\text{low}} \mathbf{W}_K, \quad V = \tilde{h}_l^{\text{low}} \mathbf{W}_V \quad (6)$$

$$O = \text{softmax} \left(\frac{QK^T}{\sqrt{D'}} \right) V, \quad (7)$$

其中下标 l 表示沿 l -th 时间维度的索引。因此，注意力计算是在每个帧内执行的。 $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D' \times D'}$ 是可训练的矩阵以及 $O \in \mathbb{R}^{H \times W \times D'}$ 。在此注意力操作期间，将对视频和草图令牌应用与主模型相同的位置嵌入。因此，在 SLRA 模块中，信息传播仅在空间维度内进行，而时间维度保持不变。

接下来，我们将 O 重塑为 $\hat{h}^{\text{low}} \in \mathbb{R}^{L \times D'}$ 作为一个序列，然后上采样至原始维度：

$$h_{\text{res}} = \hat{h}^{\text{low}} \mathbf{W}^{\text{up}}, \quad (8)$$

，其中 $\mathbf{W}^{\text{up}} \in \mathbb{R}^{D' \times D}$ 。最后，自适应的自注意力输出将被修改为：

$$h' = \text{SelfAttention}(h) + h_{\text{res}}. \quad (9)$$

。SLRA 的操作过程在 Figure 4 中进行了说明。

SLRA 确保卡通特定的空间特征在不破坏时间连贯性的情况下被学习，从而有效地将基于 DiT 的视频扩散模型适应于卡通领域。

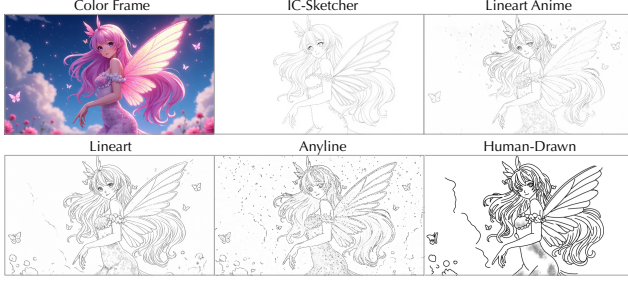


Figure 5. 训练和评估中使用的不同草图类型的示例。除了人工绘制的草图外，所有变体都包含在训练集中。训练草图的多样性提高了 ToonComposer 在实际使用案例中对不同草图风格的鲁棒性。人工绘制的草图则保留用于真实基准中的评估，如 ?? 所述。

1.3. 区域控制

有时卡通创作者可能只想绘制前景草图，并让生成器为他们创建背景。如果他们简单地将背景留空，这可能会产生不良的伪影，如 Figure 9 第二行所示。

为此，我们提出了一种新颖的区域控制机制，允许艺术家指定素描中的空白区域，以便模型根据上下文或文本提示生成合理的内容。在训练过程中，随机掩码 $m_{i_n} \in \{0, 1\}^{H \times W}$ 应用于素描帧 s_{i_n} ，其中 $m_{i_n}(i, j) = 0$ 表示未提供素描的区域。一个额外的通道与 s_{i_n} 连接，其编码为：

$$\tilde{s}'_{i_n} = [\mathcal{E}(s_{i_n}), m_{i_n}]_c, \quad (10)$$

，其中 $\tilde{s}'_{i_n}$ 用于替换训练期间在 Equation (3) 中描述的 s'_{i_n} 。该模型学习在掩蔽区域重建完整帧，从而实现灵活的推断，艺术家可以指定 m_{i_n} 的值，并将掩蔽区域留空以进行基于上下文的生成。

除了支持时间上稀疏的关键帧和草图外，我们的区域控制还允许输入草图在空间上稀疏，从而进一步减轻了卡通创作者的要求和负担。

1.4. 训练目标

ToonComposer 被训练为一个条件扩散模型，遵循 Rectified Flow [2]，它在从 logit-normal 分布中采样的时间步 t 预测速度 v_t 。为简单起见，我们将 Equation (3) 中的输入部分写为 x_{in} ：

$$x_{in} = \left[\left[\{z_k^{(t)}\}_{k=1}^K, \text{pad}(\{f'_{i_c}\}_{c=1}^C) \right]_c, \{\tilde{s}'_{i_n}\}_{n=1}^N \right]_s, \quad (11)$$

并设 $z_0 = \{z_k^{(0)}\}_{k=1}^K$ 为一个干净的卡通视频潜变量，训练目标是 최소화 期望的速度预测误差：

$$\mathcal{L} = \mathbb{E}_{z_0, \eta, t} \left[\|v_t - \epsilon_\theta(x_{in}, e_{\text{text}}, t)\|_2^2 \right], \quad (12)$$

其中 η 是随机高斯噪声， v_t 是从 $\{z_k^{(t)}\}_{k=1}^K - \eta$ 派生出的速度，而 ϵ_θ 是待训练的 ToonComposer 模型。

2. 实验

2.1. 实验设置

根据我们的内部视频资源，我们构建了 药代动力学数据，这是一个高质量的卡通数据集，包含 37K 个多样化的卡通视频片段。每个片段都有 CogVLM [24] 生成的描述性标题和一组草图帧。鉴于不同艺术家的偏好或创作工具，草图风格存在多样性，我们通过多样化的草图类型扩充了我们的数据集。具体而言，我们使用四个基于 CNN 的开源草图模型合成了每帧四个版本的草图，包括在 ControlNet [31] 中使用的两个基本线稿模型、Anime2Sketch [26] 和 Anyline [19]。此外，我们在多个艺术家的一个小型真实草图数据集上，使用 in-context LoRA [6] 调整了一个基于 FLUX 的图像到图像生成模型。该模型名为 IC-Sketcher，随后被用于生成另一个版本的草图。Figure 5 展示了一个具有多样化草图的示例帧。

基准测试 我们首先在从卡通电影获得的合成基准上评估我们的方法（经许可使用，仅用于评估），其中每个视频帧的草图是使用草图模型生成的。由于有地面实况，我们在该基准上采用基于参考的度量。此外，我们开发了 PKBench，一个以人为绘制草图特色的新基准，用于在真实场景中对卡通 后关键帧处理 进行更全面的评估。PKBench 包含 30 个样本，每个样本包括：1) 一个彩色的参考帧，2) 描述场景的文本提示，3) 两幅由专业艺术家绘制的真实草图，这些草图描绘了场景的开始和结束帧。

对于评价指标，我们采用以下方法：1) 基于参考的感知指标用于合成基准，包括 LPIPS [32]、DISTS [1] 和 CLIP [16] 图像相似性，2) 来自 VBench [7] 的无参考视频质量指标用于合成和真实基准，包括主体一致性 (S. C.)、运动一致性 (M. C.)、背景一致性 (B. C.) 和美学质量 (A. Q.)。3) 对真实基准人类感知质量的用户研究。

训练细节 我们使用数据集训练我们的模型 10 个周期，等效的批量大小为 16，使用 AdamW [14] 优化器和学习率为 10^{-5} 。SLRA 的秩 D_{low} 被设置为 144。我们使用零冗余优化器 [17] 第二阶段来减少训练期间的内存成本。

2.2. 合成基准测试评估

我们首先在一个合成卡通基准上评估我们的 ToonComposer，并与之前的方法进行比较，包括 AniDoc [15]、LVCD [8] 和 ToonCrafter [27]。在这个合成评估中，草图是从卡通视频帧中使用相同的草图模型 [26] 获得的。为了确保评估的公平性，我们在空间和时间维度上对齐了真实值，以适应每个模型的预定义设置进行指标计算。

尽管我们的模型只需一次推理即可获得最终的卡通视频，但之前的方法需要如 Figure 2 所示的两阶段过程。对于 ToonCrafter [27]，我们首先通过插值第一

Table 1. 在合成基准上的定量评估结果, 将 ToonComposer 与之前的 AI 辅助卡通生成方法进行比较: AniDoc [15]、LVCD [8] 和 ToonCrafter [27]。

Method	LPIPS↓	DISTS↓	CLIP↑	Subject Con.↑	Motion Smo.↑	Background Con.↑	Aesthetic Qua.↑
AniDoc [15]	0.3734	0.5461	0.8665	0.9067	0.9798	0.9408	0.4962
LVCD [8]	0.3910	0.5505	0.8428	0.8316	0.9810	0.9183	0.4984
ToonCrafter [27]	0.3830	0.5571	0.8463	0.8075	0.9550	0.8920	0.5035
ToonComposer (Ours)	0.1785	0.0926	0.9449	0.9451	0.9886	0.9547	0.5999

Table 2. 在真实素描基准 PKBench 上的定量评估结果, 将 ToonComposer 与之前的 AI 辅助卡通生成模型进行比较: AniDoc [15]、LVCD [8] 和 ToonCrafter [27]。

Method	S. C.↑	M. S.↑	B. C.↑	A. Q.↑
AniDoc [15]	0.9456	0.9842	0.9664	0.6611
LVCD [8]	0.8653	0.9724	0.9394	0.6479
ToonCrafter [27]	0.8567	0.9674	0.9343	0.6822
ToonComposer (Ours)	0.9509	0.9910	0.9681	0.7345

Table 3. 用户对 ToonCrafter [27]、AniDoc [15]、LVCD [8] 和 ToonComposer 生成的动画的美学质量和运动质量的偏好评分。

Method	Aesthetic Q.↑	Motion Q.↑
AniDoc [15]	4.45 %	5.34 %
LVCD [8]	7.54 %	7.91 %
ToonCrafter [27]	17.02 %	18.19 %
ToonComposer (Ours)	70.99 %	68.58 %

个和最后一个草图帧生成密集的草图序列, 然后使用其草图引导模式 (该模式需要第一个和最后一个彩色帧作为输入) 生成最终的卡通视频。对于 LVCD [8] 和 AniDoc [15], 我们首先生成由 ToonCrafter 插值的密集草图序列, 然后分别使用这两个模型将草图着色为最终的卡通视频。

结果 Table 1 显示了合成评价的数值结果。我们的方法在基于参考的指标和无需参考的指标中都优于以往的方法。例如, 我们的模型报告了明显更低的 DISTS 分数, 表明其感知质量明显优于同类方法。

Figure 6 展示了这些方法之间的定性比较, 并提供了作为参考的真实视频。在两个样本中, 我们的方法生成了平滑且自然的卡通视频帧, 而其他方法在处理这种具有稀疏草图的困难案例时显得无能为力。例如, 在第一个样本的放大图中, AniDoc 和 ToonCrafter 会产生扭曲的面部。LVCD 生成了合理的脸部, 但在后续帧中丢失了所有细节。相比之下, 我们的方法生成了一个清晰的面部, 保留了第一个参考帧的特征。这些观察结果与我们的方法在 Table 1 中的数值性能优势一致。更多结果可以在补充视频中查看。

除了在合成测试集上的评估外, 我们还在我们提出

Table 4. 对空间低秩适配器 (SLRA) 在卡通适配中的消融研究, 通过修改其内部注意模块: 仅时间 (Temp. Adapt.), 时空 (S. T. Adapt.), 以及去除注意模块 (Linear Adapt.)。我们还与使用 LoRA [4] 进行适配的基线模型进行比较。

Adaptation Method	LPIPS↓	DISTS↓	CLIP↑
Temp. Adapt.	0.1956	0.1109	0.9581
S. T. Adapt.	0.1977	0.1068	0.9587
Linear Adapt.	0.2030	0.1091	0.9589
LoRA	0.1922	0.1082	0.9628
SLRA (Ours)	0.1874	0.0955	0.9634

的基准 PKBench 上与真实的人类素描进一步比较了所有方法。由于对于每个样本没有可用的真实数据, 我们使用 VBench [7] 中的不依赖参考的指标评估生成的视频。定量比较如 Table 2 所示, 我们的模型在所有指标上都优于之前的方法, 实现了更优的外观和运动质量。

Figure 7 可视化了所有方法的比较, 最右侧列中提供了来自随机选择区域的放大图块。可以观察到, 先前的方法偏离了第一参考帧的整体风格。具体来说, ToonCrafter 生成的中间帧具有突出的粗线条, 可能是受到人工绘制草图中粗笔触的影响, 显示出其对多样草图风格的有限鲁棒性。相比之下, 我们的 ToonComposer 生成的视频帧具有卓越的视觉质量、运动连贯性和风格一致性, 这与定量结果一致。

为了进一步研究用户对生成结果的偏好, 我们进行了人工评估, 比较了我们的方法和其他基准方法生成的结果。我们从基准测试中随机选择 30 个样本, 并使用上述流程为每种方法生成卡通视频。我们的评估过程包括 47 名参与者, 每位参与者被要求选择出在美学质量和运动质量上最好的视频。结果如 Table 3 所示, 我们的方法在这两个指标上均获得了最高的胜率, 显著超过了第二名竞争者。

2.3. 讨论与分析

为了评估空间自适应在 ToonComposer 中的重要性, 我们对 SLRA 进行了消融研究, 结果详见 Figure 4。我们修改了 SLRA 的内部注意机制, 以探索替代的自适应行为: a) 时间自适应 (Temp. Adapt.), 它关注时间动态; b) 时空自适应 (S.T. Adapt.), 同时调整两者; c) 退化的线性适配器 (Linear Adapt.), 完全移除注意块。d) 使用 LoRA [4] 的基线, 该基线修改

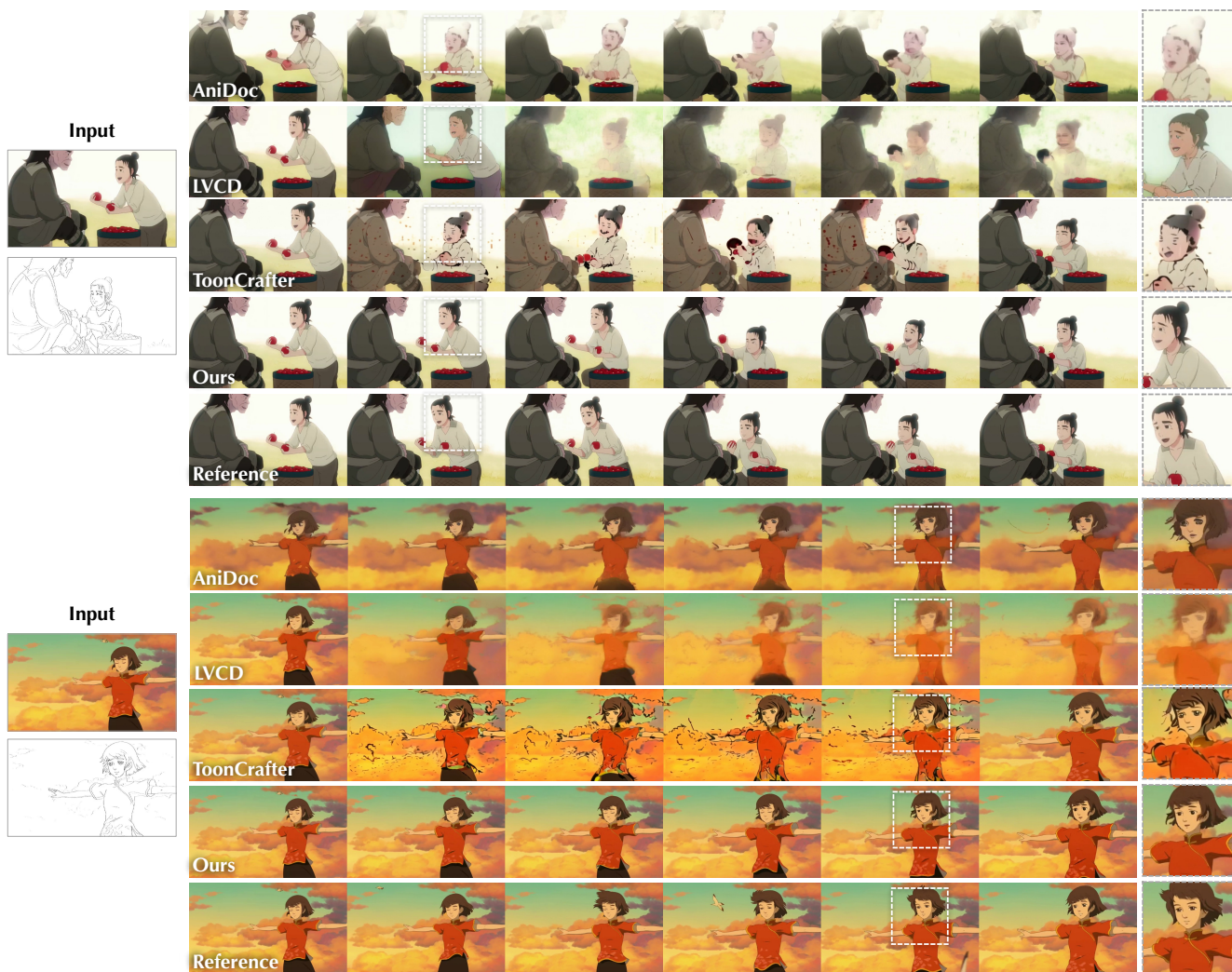


Figure 6. 在 AniDoc、LVCD、ToonCrafter 和我们的 ToonComposer 之间的综合基准比较。右侧列展示了随机选择区域的放大补丁。我们的方法显示出卓越的视觉质量、更流畅的运动和与输入图像更好的风格一致性。评估场景经过许可来自电影《妙先生》和《风语咒》。更多结果请参阅补充视频。

了 DiT 注意模块中所有线性层 (query, key, value, 和 output)。这种设计隐含地改变了空间和时间行为。为确保公平，LoRA 的秩被设定为 24，以匹配 SLRA 的可训练参数数量。所有模型均在相同设置下训练。使用 LPIPS、DISTS 和 CLIP 图像相似度进行评估。

结果如 Table 4 和 Figure 8 所示，其中 SLRA 在数值结果和视觉质量方面均优于所有变体。具体来说，a) Temp. Adapt. 和 b) S.T. Adapt. 由于空间调整不足或冲突而产生更高的误差，而 c) Linear Adapt. 缺乏卡通美学所需的细致适应。尽管 d) LoRA 的适用范围更广，但由于其适应性不够针对性，破坏了对平滑过渡至关重要的时间先验，因此表现不如 SLRA。这些发现强调了 SLRA 在调整 DiT 的空间行为以适应卡通特征方面的有效性，同时保持时间先验不变。

我们形象化展示了区域控制如何影响生成的视频。

在没有区域控制的情况下，在关键帧草图中留一个空白区域会使模型将其解释为一个无纹理的区域，导致生成的帧中出现平坦的区域，如 Figure 9 的第二行所示。相反，在启用区域控制的情况下，用户可以简单地用刷子工具绘制一个区域，以指示需要根据上下文生成适当运动的区域。如 Figure 9 的最后一行所示，我们的模型能够从输入的关键帧、草图和给定的遮罩中推断，并自动生成遮罩区域内火车的合理运动。这一机制显著提高了灵活性，进一步减轻了实际场景中的手动工作量。

通过增加关键帧草图实现可控性 ToonComposer 的稀疏草图注入机制通过支持输入关键帧草图的可变数量，实现了灵活的控制，增加了其在动画制作流程中的实用性。这种适应性允许艺术家根据期望动作的复杂

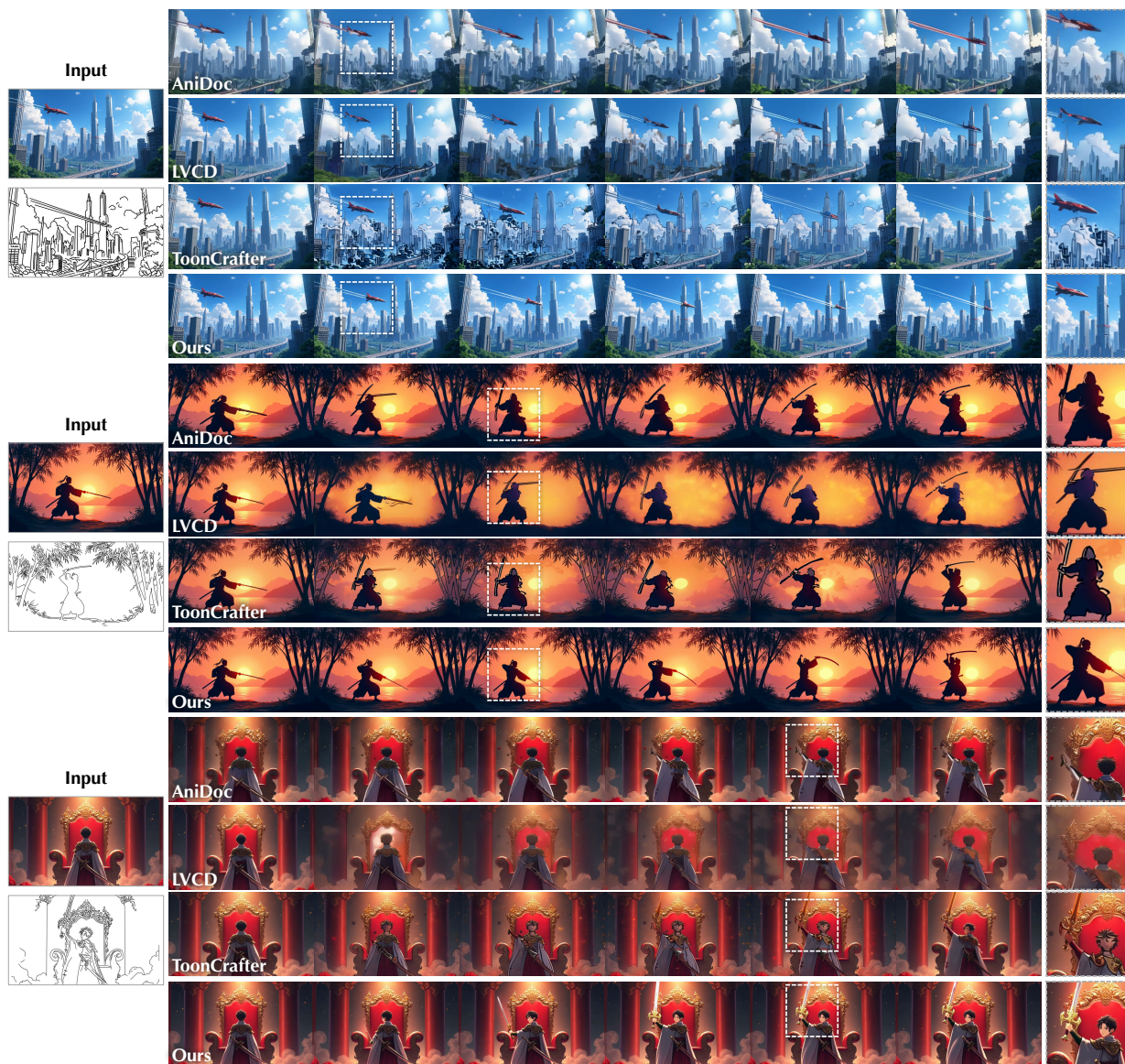


Figure 7. 在 PKBench 基准测试上的比较，使用由人类艺术家绘制的关键帧草图。右侧最末列显示了放大的细节。我们的方法能够从真实的草图输入中生成高质量的结果，而其他方法则难以保持视觉的一致性。请参考补充视频以获得更多示例。

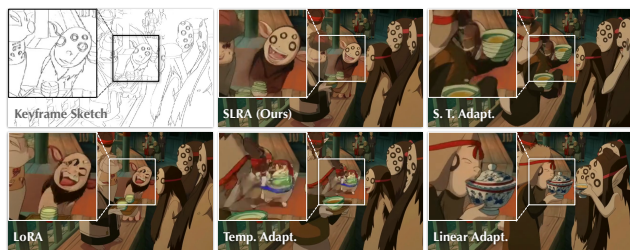


Figure 8. ToonComposer 中的空间低秩适配器 (SLRA) 消融研究，比较使用不同适配方法所生成的卡通输出帧。与其他方法相比，SLRA 提供了更高的视觉质量和更好的一致性，与输入关键帧草图更为一致。

性来平衡创作控制和自动化。如 Figure 10 所示，我们展示了 ToonComposer 在相同文本提示下，从不同数

量的输入草图生成不同动画序列的能力。更多示例可在补充视频中查看，该视频展示了我们方法在多种场景中的多样性。

尽管制作管道有所不同，ToonComposer 通过将初始参考帧适配为 3D 渲染图像，扩大了其在 3D 渲染动画中的适用性。我们在一个精简的 3D 动画片段数据集上微调了模型，使其能够以后制关键帧方式生成高质量的 3D 风格序列。这种适应性突出了 ToonComposer 的多功能性及其在更广泛动画应用中的潜力。这些 3D 动画样本在补充视频中提供。

在本文中，我们提出了一种新的模型 ToonComposer，通过一种称为“后关键帧处理”的统一生成过程，自动化中间帧和上色的繁琐任务，从而简化卡通制作。基于 DiT 架构，ToonComposer 利用稀疏关键

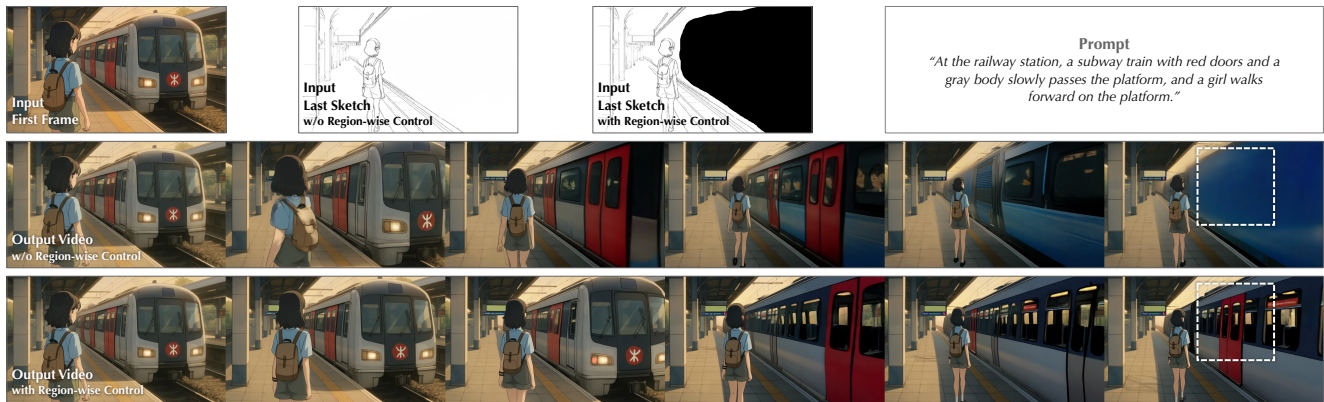


Figure 9. 在 ToonComposer 中逐区域控制的示例。如果没有逐区域控制，关键帧草图中的空白区域会被误解为无纹理区域，导致生成一个平面的蓝色火车（第二行，用虚线方框标出）。通过逐区域控制，用户可以指定利用上下文进行生成的区域，而无需明确的草图，从而使模型能够创建合理且详细的内容，例如动态的火车运动（第三行，用虚线方框标出）。



Figure 10. ToonComposer 的灵活可控性与不同的关键帧草图。仅使用草图 # 1 作为最终关键帧和提示词“老年人转身”，ToonComposer 生成一段老年人直接转身的序列（第一行）。在保持提示词不变的情况下，添加草图 # 2 来控制中间关键帧，会生成一段老年人先捡起水果然后再转身的序列（第二行）。

帧草图和单一带色参考帧来生成高质量、风格一致的卡通视频序列。实验表明，ToonComposer 在视觉保真度、运动一致性和制作效率方面超越了现有方法。稀疏草图注入和分区域控制等功能赋予艺术家精度和灵活性，使 ToonComposer 成为卡通创作的多功能系统。尽管存在如计算成本等限制，ToonComposer 通过生成模型为简化卡通制作流程提供了一个有前景的解决方案。

我们对 B & T 工作室和古东动画工作室表示诚挚的感谢，他们慷慨地允许我们在工作中使用他们的动画内容。

References

- [1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.
- [2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [3] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl:

Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101, 2024.

- [4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [5] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.
- [6] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775, 2024.
- [7] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- [8] Zhitong Huang, Mohan Zhang, and Jing Liao. Lvcd: reference-based lineart video colorization with diffusion models. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024.

- [9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004, 2017.
- [10] Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu, Bingwen Zhu, Jixuan Xu, et al. Exploring the frontiers of animation video generation in the sora era: Method, dataset and benchmark. arXiv preprint arXiv:2412.10255, 2024.
- [11] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.
- [12] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V. Sander. Deep sketch-guided cartoon video inbetweening. IEEE Transactions on Visualization and Computer Graphics, 28(8):2938–2952, 2021.
- [13] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. arXiv preprint arXiv:2406.15339, 2024.
- [14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- [15] Yihao Meng, Hao Ouyang, Hanlin Wang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Zhiheng Liu, Yujun Shen, and Huamin Qu. Anidoc: Animation creation made easier. arXiv preprint arXiv:2412.14173, 2024.
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.
- [17] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE, 2020.
- [18] Wang Shen, Cheng Ming, Wenbo Bao, Guangtao Zhai, Li Chenn, and Zhiyong Gao. Enhanced deep animation video interpolation. In 2022 IEEE International Conference on Image Processing (ICIP), pages 31–35. IEEE, 2022.
- [19] Xavier Soria, Yachuan Li, Mohammad Rouhani, and Angel D Sappa. Tiny and efficient model for the edge detection generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1364–1373, 2023.
- [20] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.
- [21] Yunlong Tang, Junjia Guo, Pinxin Liu, Zhiyuan Wang, Hang Hua, Jia-Xing Zhong, Yunzhong Xiao, Chao Huang, Luchuan Song, Susan Liang, Yizhi Song, Liu He, Jing Bi, Mingqian Feng, Xinyang Li, Zeliang Zhang, and Chenliang Xu. Generative ai for cel-animation: A survey. arXiv preprint arXiv:2503.00000, 2025.
- [22] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.
- [23] Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. arXiv preprint arXiv:2502.08639, 2025.
- [24] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems, 37:121475–121499, 2024.
- [25] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024.
- [26] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1434–1444, 2022.
- [27] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Toon-crafter: Generative cartoon interpolation. arXiv preprint arXiv:2405.12345, 2024.
- [28] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. IEEE Transactions on Visualization and Computer Graphics, 2024.
- [29] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072, 2024.
- [30] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.
- [31] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023.
- [32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

- [33] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In European Conference on Computer Vision, pages 145–162. Springer, 2024.
- [34] Junhao Zhuang, Xuan Ju, Zhaoyang Zhang, Yong Liu, Shiyi Zhang, Chun Yuan, and Ying Shan. Color-flow: Retrieval-augmented image sequence colorization. arXiv preprint arXiv:2412.11815, 2024.
- [35] Junhao Zhuang, Lingen Li, Xuan Ju, Zhaoyang Zhang, Chun Yuan, and Ying Shan. Cobra: Efficient line art colorization with broader references. arXiv preprint arXiv:2504.12240, 2025.