arXiv:2305.05464v1 [cs.CV] 9 May 2023

Style-A-Video：面向任意基于文本的视频风格迁移的敏捷扩散

Nisha Huang[1,2] Yuxin Zhang[1,2] Weiming Dong[1,21]中国科学院大学人工智能学院 [2]中国科学院自动化研究所 MAIS
https://github.com/haha-lisa/Style-A-Video

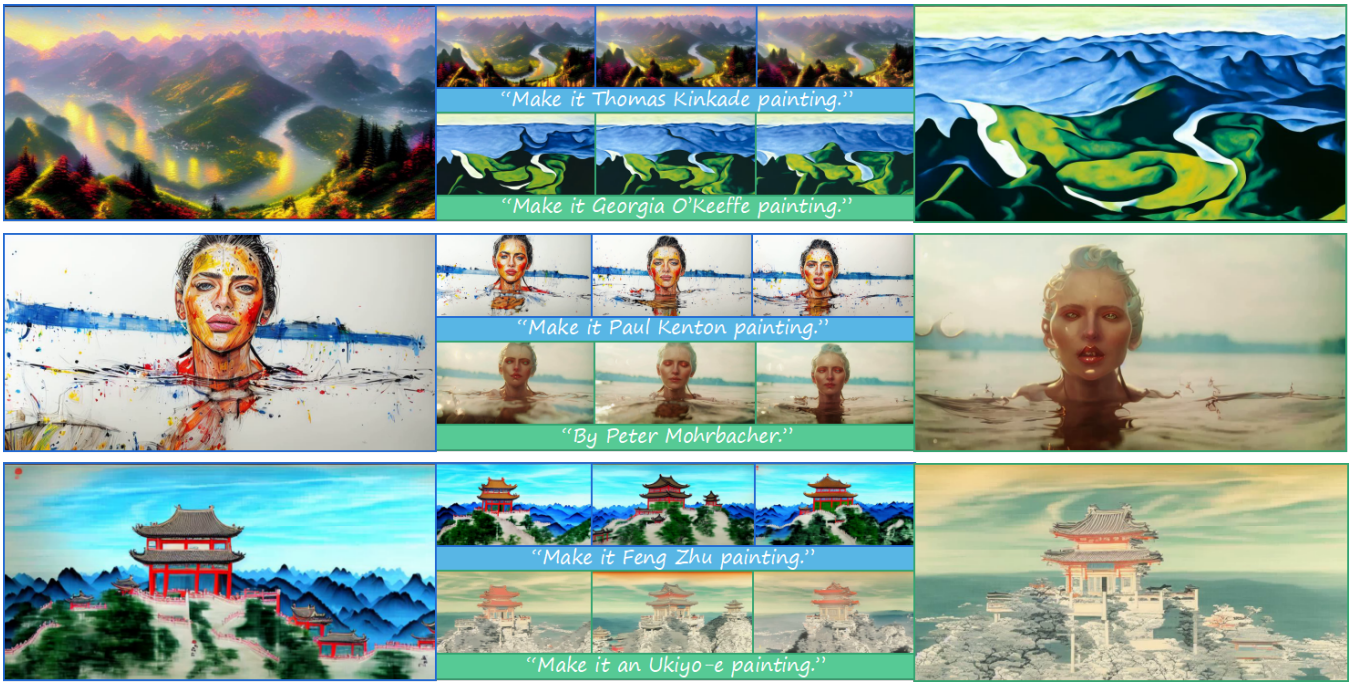

图 1：我们提出了一种基于扩散模型、由文本驱动的视频风格化方法 Style-A-Video。

摘要

大规模文本生成视频扩散模型展现出了合成多样化视频的卓越能力。然而，由于缺乏大规模文本生成视频数据集以及训练所需的计算资源，直接将这些模型用于视频风格化仍然十分困难。此外，输入内容的加噪过程具有随机性且会造成破坏，因此很难满足风格迁移对内容保持的要求。本文提出了一种名为 Style-A-Video 的零样本视频风格化方法，该方法利用生成式预训练 Transformer 与图像潜在扩散模型，实现简洁的文本控制视频风格化。我们改进了去噪过程中的引导条件，在艺术表现与结构保持之间取得平衡。此外，为了减少帧间闪烁并避免产生额外伪影，我们采用了采样优化和时间一致性模块。大量实验表明，与以往方法相比，我们的方法在消耗更少资源的同时，能够实现更出色的内容保持效果和风格表现。

1 引言

随着数亿规模的大型多模态数据集 [37, 44]向公众开放，文本驱动图像生成领域取得了突破。DALL-E 2 [38], 、Imagen [43], 和 Stable Diffusion [40], 等文本条件生成方法，既具备强大的图像生成能力，又提升了易用性，使新手也能生成高质量图像。这为利用当前生成式研究成果编辑真实世界中的视觉视频开辟了诸多可能性。然而，将扩散模型用于编辑真实世界的视频仍然困难重重。

由于难以收集大规模的文本与视频配对数据，视频领域的发展相比图像领域更为有限。从头开始训练文本引导的视频生成范式 [15, 45] 既耗时又耗费资源，而且难以获取和泛化。因此，利用现有的文本-图像模型生成视频更为实用。与此同时，研究人员已基于文本-图像模型开展了视频相关研究，并取得了可喜的成果。Text2LIVE [2] 通过计算显式对应关系来传播编辑器，

Nisha Huang[1,2] Yuxin Zhang[1,2] WeimingDong[1,2]中国科学院大学人工智能学院 [2]中国科学院自动化研究所 MAIS

https://github.com/haha-lisa/Style-A-Video

到预训练的神经分层图集（NLA） [25]模型。Tune-A- 视频 [51] 在现有图像生成模型的基础上，对每个输入视频进行微调。尽管这些努力有助于减少资源浪费，但训练 NLA 以及微调输入仍然会带来高昂的计算和时间成本。相比之下，我们希望对任意视频实现快速推理，以避免高昂的计算消耗。

此外，由于扩散模型中的噪声添加过程会破坏内容，以及去噪过程具有随机性，输入视频的内容往往难以保留。此外，文本-视频模型中的文本提示仅告诉模型用户希望改变什么，却没有传达用户希望保留什么。因此，文本引导的视频迁移会导致视频 [9, 51], 中的内容发生修改，这几乎无法满足风格化任务 [5, 6, 13] 的基本要求。

为了实现文本提示的风格表征、保留输入视频的内容、确保帧间一致性，并加快推理过程的优化，我们着重关注以下几个核心组成部分。首先，我们提出一种新的控制条件组合，包括文本、视频帧和注意力图；具体而言，文本用于风格引导，视频帧用于内容引导，注意力图用于细节引导。我们还采用自定义引导方法调整噪声预测过程，并将其与无分类器引导相结合。在实现文本引导风格化的同时，输入视频的整体内容得以良好保留。无需微调或额外训练，我们便能在推理时快速收敛，从而支持文本与视频输入的任意风格化任务。最后，我们还引入时间一致性，以消除闪烁并增强时间连贯性。

总而言之，我们提出了以下贡献 s:

- 在这项工作中，我们提出了 Style-A-Video，这是一种基于扩散模型、用于任意文本驱动视频风格化的新型框架。整个过程均在推理时完成，无需额外的逐视频训练或微调。
- 我们提出了新颖的噪声预测引导公式，以实现对风格、内容和结构的同时控制。此外，我们还实现了推理过程中对时间和内容一致性的控制。
- 大量实验和用户研究表明，与相应的基线方法相比，我们的方法具有更高的视觉质量和有效性。

2 相关工作

图像与视频风格迁移。近年来，图像风格迁移得到了广泛研究，使人们无需具备专业画家的技能，也能创作出艺术画作。Gatys 等人 [13] 发现，CNN中特征图的内积可用于表征风格，并提出了一种通过连续优化迭代实现的神经风格迁移（NST）方法。然而，这一优化过程耗时较长，难以广泛应用。

许多技术 [7, 24, 31, 55]通过对齐风格图像与内容图像的二阶统计量，实现任意风格迁移。采用自适应实例归一化（AdaIN），利用风格特征的均值和方差对内容特征进行归一化，Huang 等 [24] 提出了这种任意风格迁移方法。为了分别生成内容和风格的领域特定序列，Deng 等 [7] 提出了 StyTr2，它包含两个独立的 Transformer 编码器。通过比较和对比各种风格，并考虑风格分布，Zhang 等 [55]提出了对比任意风格迁移（CAST），使用户能够直接从图像属性中学习风格表示。

大多数图像风格迁移方法 [5, 12, 42]都被用于视频风格迁移。光流计算的准确性对基于序列的方法的效果有显著影响。使用不准确的光流会导致重影伪影。此外，由于预测光流的计算开销很大，处理高分辨率或长视频颇具挑战。我们旨在提出一种高效的零样本视频风格迁移方法，使其兼具良好的时间一致性和风格表现。

视频风格迁移将参考风格图像的风格通过统计方式应用到输入视频 [5, 6, 12, 27, 42] 上。相比之下，我们的方法在提取的结构数据约束下，将输入文本提示或图像中的风格与内容相结合。通过从数据中学习生成模型，我们的方法能够生成语义一致的结果，而非仅匹配特征统计信息。

文本到图像合成与编辑。自条件生成对抗网络被用于文本引导的图像生成以来，这一领域取得了显著进展 [30, 39]。ManiGAN [30]包含一种文本条件生成对抗网络，可在保留图像内容的同时编辑对象的外观。然而，这类多模态生成对抗网络方法局限于特定图像领域，且文本表达能力有限。DALL-E [39] 通过在海量数据集上学习联合图像-文本分布来解决这一问题。DALL-E 能够以出色的准确度将文本转换为图像，但其设计目的并非编辑已有图像。

另一方面，去噪扩散概率模型（DDPMs） [16] 已成功应用于文本生成图像。GLIDE [33] 进一步采用了这一方法，同时支持文本生成图像和图像修复。DALL-E 2 [38] 利用 CLIP [37] 潜在空间和先验模型。为提高效率，Stable Diffusion [40] 在潜在空间而非像素空间中从图像生成文本。

近期涌现的多种方法使用预训练生成器 [11, 35] 和预训练的 CLIP [37] ，在生成过程中提供文本引导。StyleCLIP [35], 、StyleGAN-NADA [11], 、CLIPstyler [28], 和 LDAST [10] 基于 CLIP 优化来修改内容图像。StyleCLIP 和 StyleGAN-NADA 受限于预训练生成器，只能处理特定数据领域。CLIPstyler 和 LDAST 则能够处理任意内容图像，但在风格化表示方面仍有不足。随着生成式研究的迅速发展，扩散方法进一步推动了风格化技术的发展。多种基于扩散的方法，例如 MGAD [22],、DiffStyler [23],和 InST [54] ，展现出令人瞩目的效果，并拓宽了应用场景。

文本到视频合成与编辑。尽管文本生成图像已取得显著进展，文本到视频生成仍是一个新兴研究领域，主要是因为高质量文本与视频配对数据并不充足。通过将文本 token 映射到视频

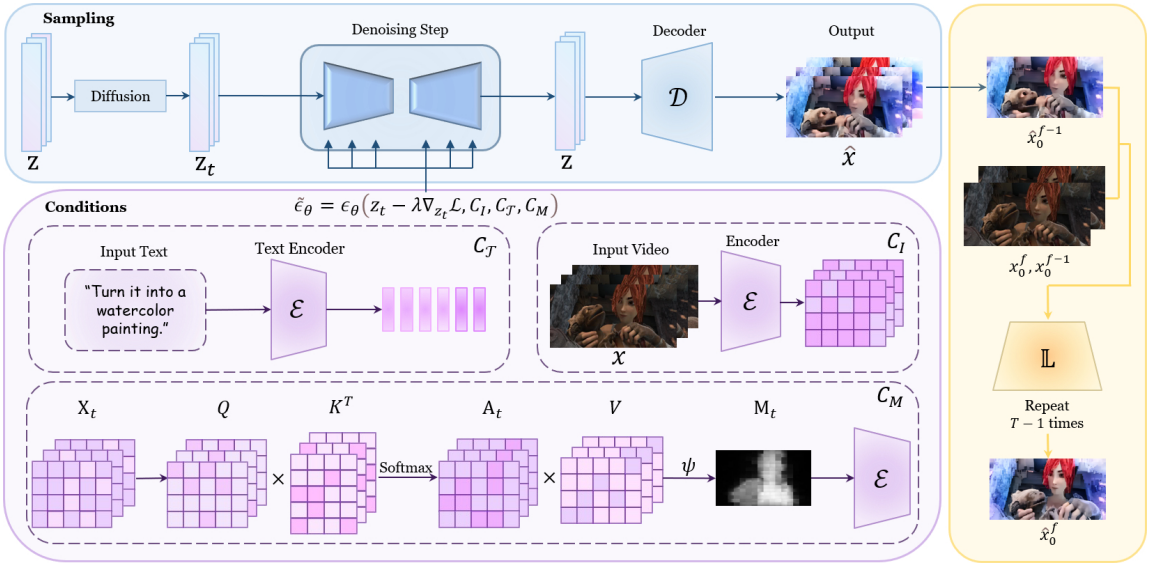

图 2：我们用于文本驱动的视频风格迁移的方法整体框架。紫色、蓝色和黄色部分分别表示条件表示、采样模块和时间一致性模块。该方法在保留输入视频内容的同时，通过对内容的文本描述迁移风格，从而生成风格化视频。

token，GODIVA [49] 是首个将 VQ-VAE [48] 拓展到文本生成视频的项目。Nüwa [50] 提出了一种集成的自回归框架，用于处理将文本生成图像和视频的任务。CogVideo [20] 将 CogView-2 [8] 拓展到文本到视频生成，并利用预训练文本生成图像模型和时间注意力模块来提高视频生成质量。视频扩散模型（VDM） [18] 提出了一种因式分解时空 U-Net，可直接在像素上执行扩散过程。为了生成高分辨率视频，Imagen Video [15] 近期通过采用级联扩散模型和 v 预测参数化，进一步改进了 VDM。类似地，Make-A-Video [45] 旨在将文本生成图像领域的重大进展应用于文本到视频生成。该方法将无监督视频素材中的世界运动与文本图像数据中的外观文本信息相结合。

随后，文本引导的视频编辑 [2, 9, 34, 51] 也迅速发展起来。Text2LIVE [2] 通过将视频分解为神经层，实现了使用文本指令修改输入影片。分层视频表示一旦建立，便可在各帧之间实现持续传播。SinFusion [34] 通过在单个视频上微调扩散模型，能够生成视频的变体和外推结果。类似地，Tune-A-Video [51] 在单个视频上优化由图像模型生成的视频，以便进行修改。然而，由于逐视频训练成本高昂，这些方法在创意工具中的可行性受到限制。最近，Esser 等人提出了 Gen-1 [9]，，可以使用文本编辑任意视频，但编辑后的视频在一定程度上会丢失输入视频的内容信息。

3 方法

我们提出的方法旨在根据文本条件 T 对输入视频进行风格化编辑，同时保留输入视频的内容和结构。内容具体指输入视频的外观和语义信息。结构指输入视频的形状和几何特征。为此，我们将输入帧信息（记为 I）和自注意力信息（记为 M）加入生成模型条件中。我们推断编码文本 T 与去噪器ϵ_θ的中间特征之间的自注意力图 M。具体流程如图 2 所示。首先，将生成骨干网络实现为条件势视频扩散模型。接着，计算并编码引导条件。然后，将引导条件与无分类器引导相结合，以进行噪声预测。在采样过程中执行逐步优化。最后，对帧间的时间一致性进行优化。

3.1 扩散模型

扩散模型 [46] 学习逆转一个固定的前向扩散过程。每个样本 x_{t-1} 都会被逐渐添加服从高斯分布的噪声，以获得 x_t。前向过程对固定的马尔可夫链进行建模，噪声取决于方差调度 β_t，其中 $t \in \{1, \ldots, T\}$，T 是扩散链中的总步数，且 $x_0 := x$。

去噪扩散概率模型（DDPMs） [16] 是潜在生成模型，其训练目标是重建固定的前向马尔可夫链 $x_1, \ldots, x_T$。给定数据分布 $x_0 \sim q(x_0)$，马尔可夫转移 $q(x_t \mid x_{t-1})$被定义为具有方差调度 $\beta_t \in (0,1)$ 的高斯分布，即，

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbb{I}\right), \tag{1}$$

其中 $t = 1$、……、 T。根据贝叶斯规则和马尔可夫性质，可以明确表示条件概率 $q(x_t \mid x_0)$和

Nisha Huang[1,2] Yuxin Zhang[1,2] WeimingDong[1,2]中国科学院大学人工智能学院 [2]中国科学院自动化研究所 MAIS
https://github.com/haha-lisa/Style-A-Video

$q(x_{t-1} \mid x_t, x_0)$ 作为

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbb{I}\right), \tag{2}$$

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbb{I}\right), \tag{3}$$

$$\begin{gathered}\text{w.r.t. } t = 1, \ldots, T, \alpha_t = 1 - \beta_t, \\ \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \\ \tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_t}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t.\end{gathered} \tag{4}$$

为生成马尔可夫链 $x_1, , x_T$，去噪扩散概率模型利用反向过程，采用先验分布 $p(x_T) = \mathrm{N}(x_T; 0, \mathbb{I})$ 和高斯转移

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{5}$$

其中 $t \in \{T, \ldots, 1\}$。可学习参数 θ经过训练，以确保生成的反向过程接近前向过程。

为此，去噪扩散概率模型遵循变分推断原则，通过最大化负对数似然的变分下界进行训练。该下界可根据高斯分布之间的 KL 散度以封闭形式表示。从经验上看，这些模型可以解释为一系列权重共享的去噪自编码器 $\epsilon_\theta(x_t, t)$，它们经过训练，可预测输入 x_t 的去噪变体。目标函数可简化为

$$\mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \tag{6}$$

算法1条件引导扩散采样。

Input: The text prompt $\mathcal{T}$, video frame I,
frame number F, Diffusion model Model(z_t).
Output: Stylized frames $x_0^1, ..., x_0^F$.
$\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ a unit Gaussian random variable with specific seed S
  for f in $1, ..., F$ **do**
  $x_t^f \leftarrow noising(x_0^f)$
  $z_t = \mathcal{E}(x_t^f)$
    for t in $T, ..., 1$ **do**
      $\epsilon, \Sigma, M \leftarrow \mathrm{Model}(\mathbf{z}_t)$
      $\Delta z_t = \nabla_{z_t}\mathcal{L}_s$
      $C_I, C_\mathcal{T}, C_M \leftarrow$ CLIP embedding $(I, \mathcal{T}, M)$
      $\tilde{\epsilon}_\theta = \epsilon_\theta(z_t - \lambda\Delta z_t, C_I, C_\mathcal{T}, C_M)$
      $\mathbf{z}_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\tilde{\epsilon}\right), \Sigma\right)$
    end for
    return z_0
    $\hat{x}_0^f = \mathcal{D}(\mathbf{z}_0)$
    if $f = 1$
      $\hat{x}_0^{f-1} = \varnothing$
    else
      $\hat{x}_0^f \leftarrow x_0^f, \hat{x}_0^{f-1}$
  end for
  return x_0^f

潜在扩散模型（潜在扩散模型） [40] 是新近提出的去噪扩散概率模型变体，在自编码器的潜在空间中运行。潜在扩散模型能够提高扩散模型的效率和质量

这是通过在预训练变分自编码器 [26] 的潜在空间中运行实现的。潜在扩散模型包含两个关键组件。首先，使用大型图像集上的分块损失训练自编码器。对于图像 x，扩散过程会向编码后的潜在表示 $z = \mathrm{E}(x)$ 添加噪声，生成带噪潜在表示 z_t，其中噪声水平会随时间步 $t \in T$ 推移而增加。解码器 D 则学习将潜在表示重建回像素空间，使得 $\mathrm{D}(\mathrm{E}(x)) \approx x$。第二个组件是一个去噪扩散概率模型，它经过训练以去除添加到图像潜在表示中的噪声。该扩散模型可由类别标签的编码嵌入作为条件。我们训练一个网络 θ，在给定图像条件 c_I、文本提示条件 $c_\mathcal{T}$ 和注意力图条件 c_M 的情况下，预测添加到带噪潜在表示 z_t 中的噪声。我们最小化以下潜在扩散目标函数：

$$L = \mathbb{E}_{\mathcal{E}(x),c_I,c_\mathcal{T},c_M,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_I, c_\mathcal{T}, c_M))\|_2^2\right]. \tag{7}$$

3.2 条件表示

风格条件。 条件扩散模型非常适合对诸如 $p(x \mid c)$ 的条件分布进行建模。在这种情况下，前向过程 q 保持不变，而条件变量 c 则成为模型的额外输入。我们的目标是根据描述所需编辑视频的文本提示来编辑输入视频。因此，在采样过程中，我们用文本提示描述替代类别条件。目标风格可以从文本提示 $c_\mathcal{T}$ 中获得，并可由以下公式表示：

$$z \sim p_\theta(z \mid c_\mathcal{T}), \quad x = \mathcal{D}(z). \tag{8}$$

内容条件。 对于风格化任务而言，除了风格表示之外，内容保留也是另一个关键问题。由于扩散模型的加噪过程对内容图本身具有破坏性，以及去噪过程具有随机性，因此很难在内容保留方面取得更好的效果。以往使用 CLIP [37] 图像嵌入来表示内容条件的工作 [1, 9],难以满足传统的风格化要求。相比之下，我们在第一个卷积层中添加额外的输入通道，将 z_t 和 c_I进行拼接，从而使最终生成的结果相对于输入视频在语义内容上具有更高的一致性。我们使用预训练检查点初始化扩散模型的所有可用权重，并将作用于新添加输入通道的权重设为零。

自特征条件。近期的大规模扩散模型 [14,22, 33] 通过在去噪 U-Net ϵ_θ 中加入注意力层 [14, 19]来实现条件控制。自注意力机制是注意力机制的一种变体，它对外部信息的依赖较少，更擅长捕捉自特征之间的内部相关性。具体而言，给定时间步 t下任意特征图 $X_t \in \mathbb{R}^{(H)w\times d}$，其高度为 H、宽度为 W， N头自注意力定义如下：

$$Q_t^{(h)} = X_t W_Q^{(h)}, \quad K_t^{(h)} = X_t W_K^{(h)}, \tag{9}$$

其中 $W_{(h)Q}, W_{(h)K} \in \mathbb{R}^{C\times d}$对于 $h = 0, 1, \ldots, N-1$。

$$A_t^{(h)} = \mathrm{softmax}\left(Q_t^{(h)}\left(K_t^{(h)}\right)^T/\sqrt{d}\right). \tag{10}$$



图 3：Style-A-Video 的更多视频风格迁移结果。我们的方法支持简单、任意的风格描述，无需像 [36, 51] require 那样详细描述视觉内容。每个结果的输入内容以插图形式展示。

随后，将每个

$$A_t \qquad V_t = X_t W_V \qquad W_V \quad \in (h) \text{右乘} (h)$$

$\mathbb{R}^{C\times d}$ (h)，其中(h)。给定掩码阈值 ψ，其实际取值设为 A_t 的均值，根据自注意力图对 x_t 中的块进行掩码，具体形式如下：

$$M_t = \mathbb{1}\left(A_t > \psi\right) \tag{11}$$

3.3 条件引导

无分类器引导 [17] 是一种在扩散模型生成的样本质量与多样性之间进行权衡的方法。它常用于类别条件和文本条件图像生成，以提升生成图像的视觉质量，并使采样图像更符合其条件。无分类器引导有效地将概率质量转移至

隐式分类器 $p_\theta(z_t \mid c)$ 会为条件 c赋予较高似然度的数据。无分类器引导的实现方式是联合训练扩散模型进行条件去噪和无条件去噪，并在推理时结合两种分数估计。训练无条件去噪时，只需在训练过程中以一定频率将条件设为固定的空值 $c = \varnothing$。在推理时，使用引导尺度 $s \geq 1$，将修改后的分数估计 $\tilde{\epsilon_\theta}(z_t, c)$ 沿着条件分数 $\epsilon_\theta(z_t, c)$ 的方向、远离无条件分数 $\epsilon_\theta(z_t, \varnothing)$ 进行外推：

$$\tilde{\epsilon_\theta}\left(z_t, c\right) = \epsilon_\theta\left(z_t, \varnothing\right) + s \cdot \left(\epsilon_\theta\left(z_t, c\right) - \epsilon_\theta\left(z_t, \varnothing\right)\right). \tag{12}$$

条件引导。 对于我们的任务，评分网络$\tilde{\epsilon_\theta}(z_t, c_I, c_\mathcal{T})$有三个条件：输入图像 c_I、文本提示 $c_\mathcal{T}$和自注意力图 c_M。我们发现，利用此方法很有益。

Nisha Huang[1,2] Yuxin Zhang[1,2] WeimingDong[1,2]1中国科学院大学人工智能学院 2中国科学院自动化研究所多媒体智能技术实验室

https://github.com/haha-lisa/Style-A-Video

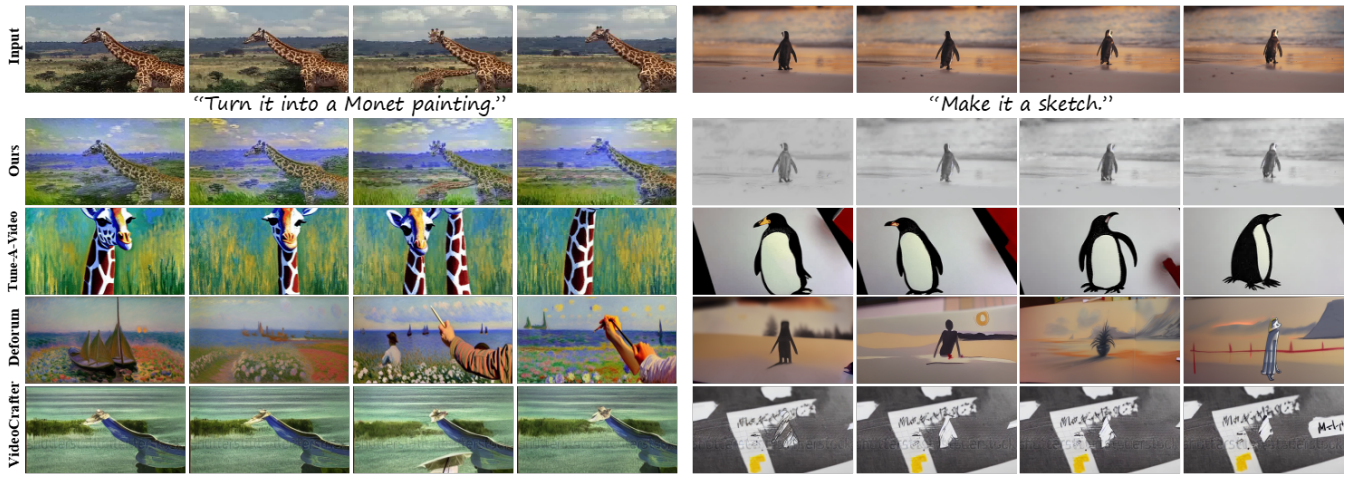

图 4：Style-A-Video 与其他基线方法的定性比较。我们的结果在时间一致性、内容保留和风格呈现方面更为出色。

无分类器引导:

$$\epsilon'_\theta = \epsilon_\theta\left(z_t, \varnothing, \varnothing\right) \tag{13}$$

关于每个条件。Liu等人 [32] 证明，有条件扩散模型可以组合来自多个不同条件值的分数估计。我们引入三个引导尺度，s_I、$s_\mathcal{T}$ 和 s_M，可以通过调整它们来权衡生成样本与条件的对应程度。修改后的分数估计如下:

$$\begin{aligned}\tilde{\epsilon_\theta}\left(z_t, c_I, c_\mathcal{T}, c_M\right) =& (1 - s_I - s_\mathcal{T} - s_M) \cdot \epsilon'_\theta \\ &+ s_I \cdot \epsilon_\theta\left(z_t, c_I, \varnothing\right) \\ &+ s_\mathcal{T} \cdot \epsilon_\theta\left(z_t, \varnothing, c_\mathcal{T}\right) \\ &+ s_M \cdot \epsilon_\theta\left(z_t, c_M, \varnothing\right).\end{aligned} \tag{14}$$

3.4 采样优化

损失。 我们希望在保留物体原始空间布局、形状和感知语义的同时，允许纹理和外观发生显著变化。尽管在风格迁移领域已有多种感知内容损失被提出，但其中大多数都使用从预训练的 VGG模型中提取的特征。我们则在 CLIP特征空间中定义损失。这样便能对生成的图像 Io 的内部 CLIP表示施加额外约束。受经典及近期研究 [47], 的启发，我们采用自相似性度量。具体而言，我们将图像输入 CLIP的 ViT编码器，并从最深层提取其空间标记。结构损失记为 L_s。通过最小化输入帧 x_0^f 与预测帧 x_t^f 之间的相似度来优化网络:

$$\mathcal{L}_s = 1 - \mathcal{D}_{\cos}\left(x_0^f, x_t^f\right), \tag{15}$$

其中 $f \in \{1、\ldots\ldots、 F\}$， F为输入视频的帧数。然后，我们根据 z 梯度 ∇_{z_t}执行一个梯度步:

$$\Delta z_t = \nabla_{z_t} \mathcal{L}_s, \tag{16}$$

表 1：基线方法的定量评估和用户研究 I。最佳结果以粗体突出显示，次佳结果以下划线标记。

	我们的方法 Tune-A-视频	Deforum	Text2LIVE	VideoCrafter	
Fra-Con↑	**0.987**	0.882	0.908	0.969	<u>0.973</u>
Pro-Con↑	**0.304**	0.235	0.263	<u>0.272</u>	0.266
Fra-Acc↑	<u>0.983</u>	0.75	0.872	**0.987**	0.945
偏好↑	-	0.157	0.086	0.229	0.286

如下方程所示，该操作会同时优化去噪网络和引导条件:

$$\tilde{\epsilon}_\theta = \epsilon_\theta(z_t - \lambda \Delta z_t, C_I, C_\mathcal{T}, C_M). \tag{17}$$

3.5 时间一致性

视频帧 $\left\{x_0^f\right\}_{f=1}^F$ 在短期和长期内整体上彼此一致。然而，由于输入帧与基于图谱的帧未对齐，可能会出现局部闪烁。因此，我们额外使用局部去闪烁网络进一步细化结果。以往研究表明，基于光流的正则化能够有效处理局部闪烁。因此，我们选择了经过修改的轻量级流程 [27] 。如图2所示，我们将连续两帧x_0^f，x_0^{f-1} 以及前一输出帧 $\hat{x}_0^{f-1}$ 输入局部细化网络 $\mathbb{L}$，以预测输出帧 $\hat{x}_0^f$。首先，连续两帧经过若干卷积层处理，然后与 $\hat{x}_0^{f-1}$ 融合。局部闪烁网络通过时间一致性损失进行训练，以去除局部闪烁伪影 [29]。



图 5：更多定性比较。我们使用 Text2LIVE 提供的预训练模型进行了实验 [2]。

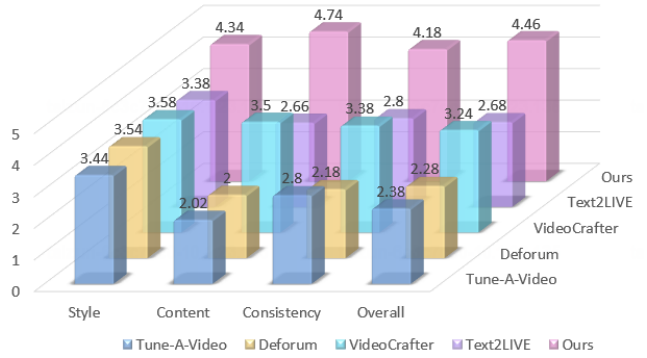

图 6：用户研究 II。四项指标下的用户平均评分结果。

4 实验

4.1 实现细节

在实践中，我们通过潜在扩散模型（潜在扩散模型） [40]实现了我们的方法。我们的开发基于两个大型预训练模型的能力，它们分别处理 [3] 不同模态：大型语言模型 GPT-3 [4] 和图像生成模型 Stable Diffusion [40]。Style-A-Video 在单张 NVIDIA GeForce RTX 3090 上生成一帧 512×256 需要 1 秒，并占用 3925 MiB。我们采用的 U-Net [41]架构基于 Wide-ResNet [52]。为确保结果质量并保持参数一致性，我们将本文实验所用的扩散步数和时间步数均设为 30。

4.2 定性评估

结果分析。我们在各种视频和风格化文本上测试了我们的方法。这些视频来自网络，包含各种对象类别，例如人物、动物、风景等。图 1 和图 3 展示了使用不同输入视频和文本提示进行引导所获得的结果。我们的方法能够处理任意镜头，例如包含人物特写的大型广角风景镜头。此外，无需额外的主体跟踪，我们的方法就能为静止镜头和运动镜头生成稳定的结果。

与基线方法比较。我们将 Style-A-Video 与最先进水平的文本驱动视频编辑方法进行比较，包括 Text2LIVE [2], Tune-A-Video [51], Defroum [21], 和 VideoCrafter [53]。Tune-A-Video 针对每个输入视频和文本都需要额外训练。图 4 的结果表明，Tune-A-Video 难以重现输入视频的内容、形状、运动和位置；例如，企鹅和长颈鹿的大小与姿势都与输入视频不符。此外，它还生成了一些与输入内容不符的内容和伪影；从结果来看，生成的背景也与输入内容不符。Deforum 生成的图像质量很高，但帧与帧之间缺乏连贯性。在长颈鹿示例中，长颈鹿变成了一艘船，还有一个正在画画的人的手臂。VideoCrafter 的呈现风格严重破坏了内容。与上述基线方法不同，我们的方法在编辑过程中能够很好地保持输入视频的内容和运动，同时实现风格化表达并获得高质量的视频风格化效果。此外，我们还将模型与 Text2LIVE 公开提供的基于"car"训练的预训练模型进行了比较；该模型能够较好地保持输入视频的内容，但难以满足文本提出的编辑要求。如图 5 所示，与我们的结果相比，Text2LIVE 的风格表现力更为有限。

4.3 定量评估

我们使用训练好的 CLIP [37] 模型进行定量评估，与先前的方法 [2, 21, 51, 53]相同。我们随机选取了由 Style-A-Video 和各最先进水平方法生成的 25 个视频。我们分别使用以下三项指标，量化时间一致性、提示一致性和帧准确性之间的权衡。

时间一致性。我们首先计算输出视频所有帧的 CLIP 图像嵌入，然后计算所有相邻帧对之间的余弦相似度，以此衡量帧的时间一致性。

提示一致性。我们首先计算输出视频中输出帧的 CLIP 图像嵌入，以及风格化文本提示的 CLIP 文本嵌入。此外，我们计算风格化帧嵌入与文本嵌入之间的平均余弦相似度，以此衡量文本与风格化效果的一致性。

帧准确度。我们计算输入视频帧及其对应风格化帧的 CLIP 图像嵌入，并计算两者之间的平均余弦相似度，以此衡量输入视频的内容保持程度。

如表 1 所示，与基线方法相比，Style-A-Video 方法在时间一致性和提示词一致性方面取得了更优结果，并且其内容保持准确度与 TextLIVE [2] 方法相当。

Nisha Huang[1,2] Yuxin Zhang[1,2] WeimingDong[1,2]中国科学院大学人工智能学院 [2]中国科学院自动化研究所 MAIS
https://github.com/haha-lisa/Style-A-Video

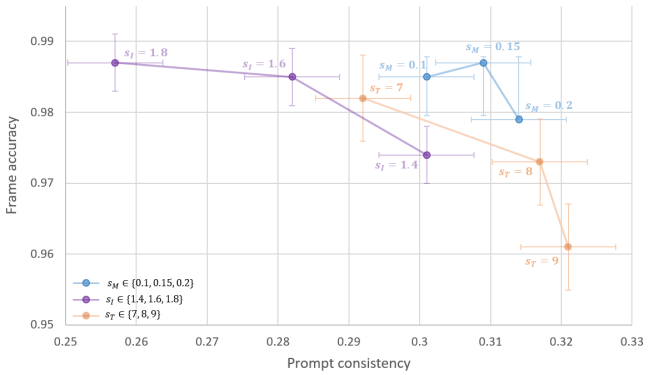

图 7：消融研究。该图展示了不同条件引导参数尺度对帧准确度和提示一致性的影响。放大查看效果更佳。

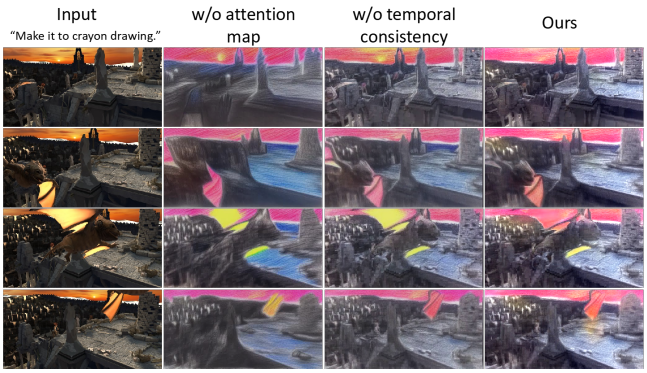

图 8：消融研究。该图展示了注意力图和时间一致性对 Style-A-Video 的影响。

4.4 用户研究

用户研究 I。 我们将本方法与几种开源的最先进水平文本引导的视频编辑方法进行了比较，包括 Text2LIVE [2], VideoCrafter [53], Deforum [21], 和 Tune-A-Video [51]。所有基线方法均使用公开可用的实现和默认配置进行训练与推理。我们随机展示了 20 组内容-风格对，每组包含 Style-A-Video 与另一种方法的结果。我们邀请 70 名参与者选择更受偏好的结果，共收集了 1400 票。表 1 的最后一行报告了各方法相较于 Style-A-Video 所获得的票数百分比。

用户研究 II. 我们设计了一项新颖而详细的用户研究，以更好地评估生成结果在各个方面的表现。我们随机向每位参与者提供 40 组文本-视频对，并请志愿者对以下标准进行评分："结果与文本描述风格的契合程度""结果与输入视频内容的契合程度""结果的时间一致性"以及"整体效果"。最终，我们收集了 70 位参与者的 11200 个评分。图 6 展示了各方法的平均分。结果表明，我们的方法在所有指标上都表现出色。

4.5 消融研究

图 7 展示了我们提出的模型中各引导条件 C_I、C_T 和 C_M 的作用。更大的引导系数 s_I、s_T 和 s_M 意味着引导权重增加。我们观察到，在基线模型中，提高强度参数会带来轻微的权衡。更大的 s_I 意味着与输入视频的一致性更强，但与风格引导的一致性会降低。同样，更大的 s_T 意味着风格化效果更强，因此与输入视频内容的偏离也更大。

为评估自注意力掩码条件和时间一致性模块对 Style-A-Video 结果的影响，我们开展了消融研究。如图 8 所示，我们分别对每项设计进行消融，以分析其影响。第二列探究了自注意力掩码条件的有效性。移除自注意力图条件会导致细节丢失（例如城镇、寺庙、翅膀）。第三列探究了时间一致性模块。如果缺少该模块，帧与帧之间就会产生颜色和光照差异，从而导致视频出现闪烁。这些结果表明，上述所有关键设计都有助于我们的方法取得成功。

5 结论与未来工作

本文提出了一种新的文本驱动视频风格化框架 Style-A-Video。该框架基于潜在扩散模型，能够根据给定的风格和内容信息对视频进行风格化。我们在每个去噪过程中引入时间一致性引导、风格引导和内容保留引导。我们进一步提出将交叉注意力模块作为结构信息的引导条件，以提升框架在结构保持和整体质量方面的表现。我们尝试同时将文本、视频和交叉注意力图用作视频风格化应用的条件。我们的框架能够以更快的推理速度生成任意视频和任意风格。我们相信，这将为视频合成与编辑的广泛应用铺平道路。未来，我们计划研究其他条件（如三个参数和姿态估计）对视频稳定性的影响。

参考文献

[1] Yogesh Balaji、Seungjun Nah、Xun Huang、Arash Vahdat、Jiaming Song、Karsten Kreis、Miika Aittala、Timo Aila、Samuli Laine、Bryan Catanzaro 等。2022。ediffi：采用专家去噪器集成的文本到图像扩散模型。*arXiv* 预印本 *arXiv:2211.01324*（2022）。[2] Omer Bar-Tal、Dolev Ofri-Amar、Rafail Fridman、Yoni Kasten 和 Tali Dekel。2022。Text2live：文本驱动的分层图像与视频编辑。载于计算机视觉——*ECCV 2022*：第 *17* 届欧洲会议，以色列特拉维夫，*2022* 年 *10* 月 *23–27* 日，论文集，第 *XV* 部分。Springer，707–723。[3] Tim Brooks、Aleksander Holynski 和 Alexei A Efros。2022。Instructpix2pix：学习遵循图像编辑指令。*arXiv:2211.09800*（2022）。[4] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是小样本学习器。神经信息处理系统进展33（2020），1877–1901。[5] Dongdong Chen、Jing Liao、Lu Yuan、Nenghai Yu 和 Gang Hua。2017。连贯在线视频风格迁移。载于*IEEE*国际计算机视觉会议论文集。1105–1114。[6] Yingying Deng、Fan Tang、Weiming Dong、Haibin Huang、Chongyang Ma 和 Changsheng Xu。2021。通过多通道相关性实现任意视频风格迁移。载于第 *35* 届 *AAAI* 人工智能会议论文集（*AAAI*），第 35 卷。1210–1217。

[7] Yingying Deng、Fan Tang、Weiming Dong、Chongyang Ma、Xingjia Pan、Lei Wang 和 Changsheng Xu。2022。Stytr2：使用 Transformer 进行图像风格迁移。载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集。11326–11336。[8] Ming Ding、Wendi Zheng、Wenyi Hong 和 Jie Tang。2022。Cogview2：通过分层 Transformer 实现更快、更优的文本生成图像。arXiv 预印本 arXiv:2204.14217（2022）。[9] Patrick Esser、Johnathan Chiu、Parmida Atighehchian、Jonathan Granskog 和 Anastasis Germanidis。2023。使用扩散模型进行结构与内容引导的视频合成。arXiv:2302.03011（2023）。[10] Tsu-Jui Fu、Xin Eric Wang 和 William Yang Wang。2022。语言驱动的艺术风格迁移。载于欧洲计算机视觉会议（ECCV）。[11] Rinon Gal、Or Patashnik、Haggai Maron、Amit H Bermano、Gal Chechik 和 Daniel Cohen-Or。2022。StyleGAN-NADA：CLIP 引导的图像生成器域适配。《ACM 图形学汇刊》（TOG）（2022），1–13。[12] Wei Gao、Yijun Li、Yihang Yin 和 Ming-Hsuan Yang。2020。快速视频多风格迁移。载于 IEEE/CVF 冬季计算机视觉应用会议论文集。3222–3230。[13] Leon A Gatys、Alexander S Ecker 和 Matthias Bethge。2016。使用卷积神经网络进行图像风格迁移。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）。2414–2423。[14] Amir Hertz、Ron Mokady、Jay Tenenbaum、Kfir Aberman、Yael Pritch 和 Daniel Cohen-Or。2022。通过交叉注意力控制实现提示到提示的图像编辑。arXiv 预印本 arXiv:2208.01626（2022）。[15] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、Ruiqi Gao、Alexey Gritsenko、Diederik P Kingma、Ben Poole、Mohammad Norouzi、David J Fleet 等。2022。Imagen Video：使用扩散模型生成高清视频。arXiv 预印本 arXiv:2210.02303（2022）。[16] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。2020。去噪扩散概率模型。《神经信息处理系统进展》（NeurIPS）（2020），6840–6851。[17] Jonathan Ho 和 Tim Salimans。2022。无分类器扩散引导。arXiv:2207.12598（2022）。[18] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet。2022。视频扩散模型。arXiv:2204.03458（2022）。[19] Susung Hong、Gyuseong Lee、Wooseok Jang 和 Seungryong Kim。2022。利用自注意力引导提升扩散模型的样本质量。arXiv 预印本 arXiv:2210.00939（2022）。[20] Wenyi Hong、Ming Ding、Wendi Zheng、Xinghan Liu 和 Jie Tang。2022。Cogvideo：通过 Transformer 进行文本到视频生成的大规模预训练。arXiv 预印本 arXiv:2205.15868（2022）。[21] Eddy Hu。2023。Deforum Stable Diffusion V0.7。https://github.com/HelixNGC7293/ DeforumStableDiffusionLocal[22] Nisha Huang、Fan Tang、Weiming Dong 和 Changsheng Xu。2022。绘制你的艺术之梦：多模态引导的多样化数字艺术合成。载于 ACM 多媒体国际会议论文集（MM）。1085–1094。[23] Nisha Huang、Yuxin Zhang、Fan Tang、Chongyang Ma、Haibin Huang、Yong Zhang、Weiming Dong 和 Changsheng Xu。2022。DiffStyler：用于文本驱动图像风格化的可控双扩散。arXiv 预印本 arXiv:2211.10682（2022）。[24] Xun Huang 和 Serge Belongie。2017。利用自适应实例归一化实现实时任意风格迁移。载于 IEEE国际计算机视觉会议（ICCV）。1501–1510。[25] Yoni Kasten、Dolev Ofri、Oliver Wang 和 Tali Dekel。2021。用于一致视频编辑的分层神经图集。《ACM 图形学汇刊》（TOG）40，6（2021），1–12。[26] Diederik P Kingma 和 Max Welling。2013。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114（2013）。[27] Xiaoyu Kong、Yingying Deng、Fan Tang、Weiming Dong、Chongyang Ma、Yongy-ong Chen、Zhenyu He 和 Changsheng Xu。2023。探索任意风格迁移的时间一致性：通道视角。《IEEE 神经网络与学习系统汇刊》（2023）。[28] Gihyun Kwon 和 Jong Chul Ye。2022。CLIPstyler：使用单一文本条件进行图像风格迁移。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）。[29] Chenyang Lei、Xuanchi Ren、Zhaoxiang Zhang 和 Qifeng Chen。2023。通过存在缺陷的图集进行神经滤波，实现盲视频去闪烁。arXiv 预印本 arXiv:2303.08120（2023）。[30] Bowen Li、Xiaojuan Qi、Thomas Lukasiewicz 和 Philip HS Torr。2020。ManiGAN：文本引导的图像编辑。载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）。7880–7889。[31] Yijun Li、Chen Fang、Jimei Yang、Zhaowen Wang、Xin Lu 和 Ming-Hsuan Yang。2017。通过特征变换实现通用风格迁移。《神经信息处理系统进展》30（2017）。[32] Nan Liu、Shuang Li、Yilun Du、Antonio Torralba 和 Joshua B Tenenbaum。2022。使用可组合扩散模型进行组合式视觉生成。载于《计算机视觉——ECCV 2022：第 17 届欧洲会议，以色列特拉维夫，2022 年 10 月 23–27 日，论文集，第十七部分。Springer，423–439。[33] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2021。GLIDE：利用文本引导扩散模型实现趋于照片级真实感的图像生成与编辑。arXiv预印本 arXiv:2112.10741（2021）。[34] Yaniv Nikankin、Niv Haim 和 Michal Irani。2022。SinFusion：在单幅图像或视频上训练扩散模型。arXiv预印本 arXiv:2211.11743（2022）。[35] Or Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or 和 Dani Lischinski。2021。StyleCLIP：文本驱动的 StyleGAN 图像操控。载于 IEEE/CVF国际计算机视觉会议（ICCV）。2085–2094。[36] Chenyang Qi、Xiaodong Cun、Yong Zhang、Chenyang Lei、Xintao Wang、Ying Shan 和 Qifeng Chen。2023。FateZero：融合注意力，实现零样本文本视频编辑。arXiv:2303.09535（2023）。[37] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于 ICML论文集。PMLR，8748–8763。[38] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022。利用 CLIP潜变量进行分层文本条件图像生成。arXiv:2204.06125（2022）。[39] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。2021。零样本文本到图像生成。载于国际机器学习会议。PMLR，8821–8831。[40] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。利用潜在扩散模型进行高分辨率图像合成。载于 IEEE计算机视觉与模式识别会议（CVPR）论文集。10684–10695。[41] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。2015。U-Net：用于生物医学图像分割的卷积网络。载于《医学图像计算与计算机辅助介入——MICCAI 2015：第18届国际会议，德国慕尼黑，2015年10月5日至9日，论文集》第三部分第18卷。Springer，234–241。[42] Manuel Ruder、Alexey Dosovitskiy 和 Thomas Brox。2016。视频艺术风格迁移。载于《模式识别：第38届德国会议，GCPR 2016，德国汉诺威，2016年9月12日至15日，论文集》第38卷。Springer，26–36。[43] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。2022。利用深层语言理解实现照片级真实感的文本到图像扩散模型。arXiv:2205.11487（2022）。[44] Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman 等。2022。LAION-5B：用于训练新一代图像文本模型的开放大规模数据集。arXiv:2210.08402（2022）。[45] Uriel Singer、Adam Polyak、Thomas Hayes、Xi Yin、Jie An、Songyang Zhang、Qiyuan Hu、Harry Yang、Oron Ashual、Oran Gafni 等。2022。Make-A-Video：无需文本-视频数据的文本到视频生成。arXiv:2209.14792（2022）。[46] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。2015。利用非平衡热力学进行深度无监督学习。载于国际机器学习会议（ICML）。2256–2265。[47] Narek Tumanyan、Omer Bar-Tal、Shai Bagon 和 Tali Dekel。2022。拼接 ViT特征以实现语义外观迁移。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10748–10757。[48] Aaron Van Den Oord、Oriol Vinyals 等。2017。神经离散表示学习。《神经信息处理系统进展》第30卷（2017）。[49] Chenfei Wu、Lun Huang、Qianxi Zhang、Binyang Li、Lei Ji、Fan Yang、Guillermo Sapiro 和 Nan Duan。2021。GODIVA：根据自然语言描述生成开放域视频。arXiv预印本 arXiv:2104.14806（2021）。[50] Chenfei Wu、Jian Liang、Lei Ji、Fan Yang、Yuejian Fang、Daxin Jiang 和 Nan Duan。2022。Nüwa：用于神经视觉世界创建的视觉合成预训练。载于《计算机视觉——ECCV 2022：第17届欧洲会议，以色列特拉维夫，2022年10月23日至27日，论文集》第十六部分。Springer，720–736。[51] Jay Zhangjie Wu、Yixiao Ge、Xintao Wang、Weixian Lei、Yuchao Gu、Wynne Hsu、Ying Shan、Xiaohu Qie 和 Mike Zheng Shou。2022。Tune-A-Video：单次调优图像扩散模型，实现文本到视频生成。arXiv预印本 arXiv:2212.11565（2022）。[52] Sergey Zagoruyko 和 Nikos Komodakis。2016。宽残差网络。arXiv:1605.07146（2016）。[53] Yong Zhang、Yingqing He、Haoxin Chen 和 Menghan Xia。2023。VideoCrafter：文本到视频生成与编辑工具包。https://github.com/VideoCrafter/VideoCrafter[54] Yuxin Zhang、Nisha Huang、Fan Tang、Haibin Huang、Chongyang Ma、Weiming Dong 和 Changsheng Xu。2022。利用扩散模型进行基于反演的创造力迁移。arXiv预印本 arXiv:2211.13203（2022）。[55] Yuxin Zhang、Fan Tang、Weiming Dong、Haibin Huang、Chongyang Ma、Tong-Yee Lee 和 Changsheng Xu。2022。通过对比学习实现域增强任意图像风格迁移。载于 SIGGRAPH论文集。ACM，12:1–12:8。[56] Daquan Zhou、Weimin Wang、Hanshu Yan、Weiwei Lv、Yizhe Zhu 和 Jiashi Feng。2022。MagicVideo：利用潜在扩散模型实现高效视频生成。arXiv预印本 arXiv:2211.11018（2022）。