UniVid：用于高质量视频生成的金字塔扩散模型

Xinyu Xiao、Binbin Yang、Tingtian Li、Yipeng Yu、Sen Lei

arXiv:2603.13739v1 [cs.CV] 14 Mar 2026

摘要——**基于扩散模型的文本到视频生成（T2V）或图像到视频生成（I2V）已成为一个重要的研究方向。然而，如何将这两种生成范式整合到统一模型中仍然面临挑战。本文提出了一种统一视频生成模型（*Univid*），它以文本提示和参考图像作为混合条件。利用这两种可用的控制方式，我们的模型能够从文本提示中提取物体外观及其运动描述，同时从图像线索中获取纹理细节和结构信息，以引导视频生成过程。具体而言，我们通过引入时间金字塔跨帧时空注意力模块和卷积，对预训练文本到图像扩散模型进行扩展，从而生成时间连贯帧。为支持双模态控制，我们引入了一种双流交叉注意力机制，其注意力分数可在推理过程中自由重新加权，以实现单模态与双模态控制之间的插值。大量实验表明，我们的*Univid* 在文本到视频、图像到视频以及（T+I）2V任务上均实现了出色的时间连贯性。**

索引词—**扩散模型、视频生成、时空注意力**

I. 引言

近期研究 [1]–[4],致力于通过训练文本到视频（T2V）扩散模型，在视频生成领域复现这一成功。文本提示是图像合成的友好接口，而对于视频，使用参考图像作为提示对用户来说可能更加方便。在本研究中，我们旨在开发一种统一视频生成模型，充分利用文本和图像提示的优势，让用户能够灵活地使用图像、文本，甚至两种模态的组合来控制视频生成，即，T2V、I2V 和 (T+I)2V。

近期大多数研究 [2], [4] 通过在每个空间模块后插入 $1D$ 时间层，或将空间注意力改为全注意力，将 T2I 扩散模型转变为视频生成器。基于 Stable Diffusion 的 U-Net 主干网络 [5], ，在本研究中，我们引入了一种金字塔时空注意力模块，通过以动态视觉节奏探索帧间区域关系，改进动态建模并捕捉运动线索。如图 1 所示，

$3D$时空注意力可以通过学习熊猫在不同帧之间的相关性，充分建模移动熊猫的运动线索；而 $1D$ 时间注意力则有所不足，因为它仅考虑逐像素的时间依赖关系。

为此，我们提出 *Univid*，这是一种统一的视频扩散模型，通过灵活组合描述性句子和参考图像作为混合控制来生成高质量视频。对于时间维度的扩展，我们从视频理解中多层次视觉节奏建模在运动动态方面取得的成功中汲取灵感

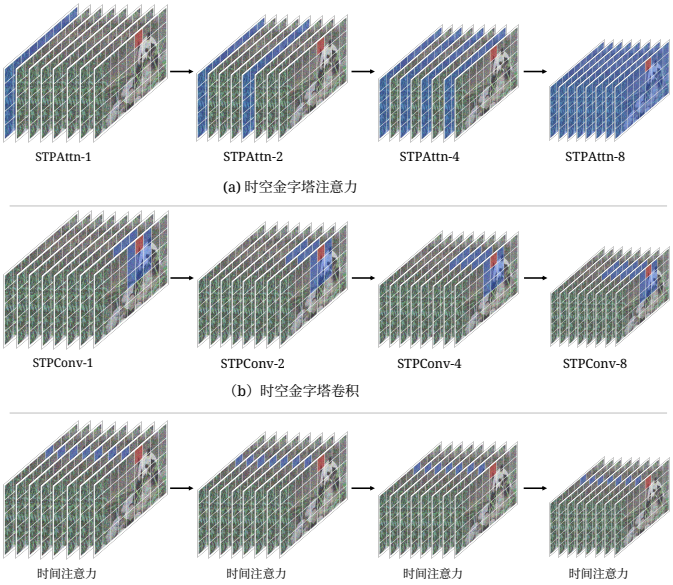

图 1. 我们提出的金字塔时空注意力（PSTAttn-f）、金字塔时空卷积（PSTConv-f）和时间注意力示意图，其中$f \in \{1, 2, 4, 8\}$ 表示 U-Net 各层的下采样因子。帧中的红色区域表示注意力查询，蓝色区域表示注意力键/值。（我们也以类似方式展示了 PSTConv-f 的感受野。）

并提出将金字塔时空注意力和卷积模块引入扩散U-Net的基本块中。通过将空间金字塔架构 [8]–[10]扩展到时间维度，我们的金字塔时空注意力以由粗到细的时间节奏，逐步学习跨帧的运动线索。具体而言，在 U-Net 的中间块中，我们对高层帧特征进行密集的时空语义对齐，而在外围块中则进行稀疏对齐。

为了有效编码图像中的外观细节和文本中的语义描述，以指导生成过程，我们提出将 Stable Diffusion 中常规的交叉注意力模块扩展为双流交叉注意力模块。与文本分支类似，我们使用 CLIP [11] 图像编码器对参考图像进行编码，并在输入特征与视觉图像块嵌入序列之间执行视觉交叉注意力。视觉交叉注意力与文本交叉注意力以双路方式运行，并将两者的结果融合为

输出特征。

总而言之，我们的主要贡献如下：

- 我们开发了一种高质量视频生成模型，即 *UniVid* ，可灵活组合文本提示和参考图像作为控制输入。
- 我们提出将金字塔时空注意力和卷积模块融入二维图像生成骨干网络，使预训练文本到图像扩散模型具备对时间运动进行建模的能力。
- 我们提出一种双流交叉注意力，以双路径方式对齐视频特征的语义与用户提供的文本提示和参考图像。
- 在 UCF-101 和 MSR-VTT 上进行的零样本生成评估表明，*UniVId* 的生成性能优于现有方法。

II. 预备知识

扩散模型 [5], [12]–[14] 是一类基于似然的生成模型，由扩散过程和去噪过程组成。形式上，数据样本 $x_0 \sim q(x_0)$ 会经过一系列高斯噪声逐步模糊：

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

其中 $\{\alpha_t\}_{t=1}^T$ 为噪声幅度。根据马尔可夫链和高斯噪声的性质，当链长 T 足够大时， x_T 将近似服从标准高斯分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$。随后，生成过程定义为从标准高斯噪声反向去噪，即 即， $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$。去噪扩散概率模型 [12] 揭示，训练扩散模型可以简化为学习估计每个时间步 t添加到 x_0 中的噪声 ϵ ：

$$\mathcal{L}_{\text{DM}}(\theta) = \mathbb{E}_{t,x_0,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon_\theta(x_t, t) - \epsilon\|^2\right], \tag{2}$$

其中 $\epsilon_\theta(x_t, t)$ 是一个由 U-Net [15] 实现的噪声预测器，其参数为 θ。

潜在扩散模型 [5] 是扩散模型的一种变体，它在变分自编码器（VAE） [16] 的压缩潜在空间中执行扩散过程，而非在原始像素空间中进行。以 Stable Diffusion [5] 为例，文本到图像潜在扩散模型的训练损失可表示为：

$$\mathcal{L}_{\text{LDM}}(\theta) = \mathbb{E}_{t,z_0,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon_\theta(z_t, \mathcal{C}_t, t) - \epsilon\|^2\right], \tag{3}$$

其中， $z_0 = \mathcal{E}(x_0)$ 是 x_0 的编码潜变量， z_t 是其在时间步 t 的含噪版本， $\epsilon_\theta(z_t, \mathcal{C}_t, t)$ 是带有文本条件 $\mathcal{C}_t$ 的噪声预测器。

III. 方法

本节将介绍我们如何提出一种级联潜在扩散模型，即 *Univid*，用于可控的高质量视频生成，使其能够通过文本描述和参考图像进行双流控制。*UidVid* 的整体框架如图 2 所示。*UidVid* 的核心模块是一个基础潜在视频扩散模型，它能够根据输入文本和可选的图像提示学习生成帧序列。

A. 双流控制下的视频生成

本文旨在利用现成的文本到图像（T2I）潜在扩散模型，例如Stable Diffusion [5]，将其作为内容先验，并将其扩展为一种灵活的视频生成器，使其能够同时受文本提示和参考图像控制。Stable Diffusion 中的 U-Net 架构由多个基本模块堆叠而成，每个模块都包含二维卷积、自注意力和交叉注意力模块：

$$\texttt{SDBlock} = \texttt{Conv2D} \circ \texttt{SelfAttn} \circ \texttt{CrossAttn}. \tag{4}$$

SDBlock 利用每帧内的像素相关性，并建立像素与输入文本之间的对齐关系，从而更新特征。然而，由于缺少必要的时空建模能力，它并不适用于生成具有运动感知内容相关性的连续帧。因此，我们的首要目标是将空间 SDBlock 扩展为时空版本，即 STBlock，使其能够捕捉时间信息，从而实现跨帧语义对齐并支持双模态输入控制。STBlock 的形式化描述如下：

$$\begin{aligned}\texttt{STBlock} = \texttt{PSTConv} \circ \texttt{SelfAttn} \circ \texttt{PSTAttn} \circ \\ \texttt{CrossAttn}_{\texttt{Dual}} \circ \texttt{TemAttn},\end{aligned} \tag{5}$$

其中，自注意力 是自注意力模块，PSTConv 和PSTAttn 分别是金字塔时空卷积和注意力，交叉注意力$_{\texttt{Dual}}$ 将 Stable Diffusion 中的交叉注意力 扩展为对文本和图像提示进行双流引导，而TemAttn是时间注意力层。

1）金字塔时空模块：SDBlock 中的自注意力和 Conv2D 会独立更新每一帧中的空间特征，而不显式考虑它们在语义上的相关性，这往往会导致不同帧之间的外观不一致。为增强跨帧一致性并纳入运动建模，我们引入了一种时空注意力模块，即PSTAttn，并通过在每个 Conv2D 模块后堆叠时间卷积，将其扩展为 $1D$PSTConv $2D$ 。

受到视频理解中利用多层次视觉节奏建模运动动态这一成功经验的启发， [6], [7], [17], [18], 我们提出在 PSTAttn 和PSTConv 中进行动态节奏调制，以捕捉视频中的复杂运动结构。借鉴空间金字塔 [8]–[10] 的理念并将其应用于时间结构，我们的金字塔时空模块以由粗到细的时间方式，逐步学习跨帧运动线索。如图 1 所示，我们在 U-Net 的中间块（较低分辨率）中实现了密集的跨帧语义相关性，而在外部层中则实现了稀疏相关性。具体而言，输入视频的编码潜变量 $z = \mathcal{E}(x) \in \mathbb{R}^{F\times C\times H\times W}$ 将在 U-Net 中以因子$f \in \{1, 2, 4, 8\}$进行下采样。[1] 我们将处理分辨率为

[1]这里我们用 $f = 1$ 表示第一次下采样操作之前或最后一次上采样操作之后的特征。

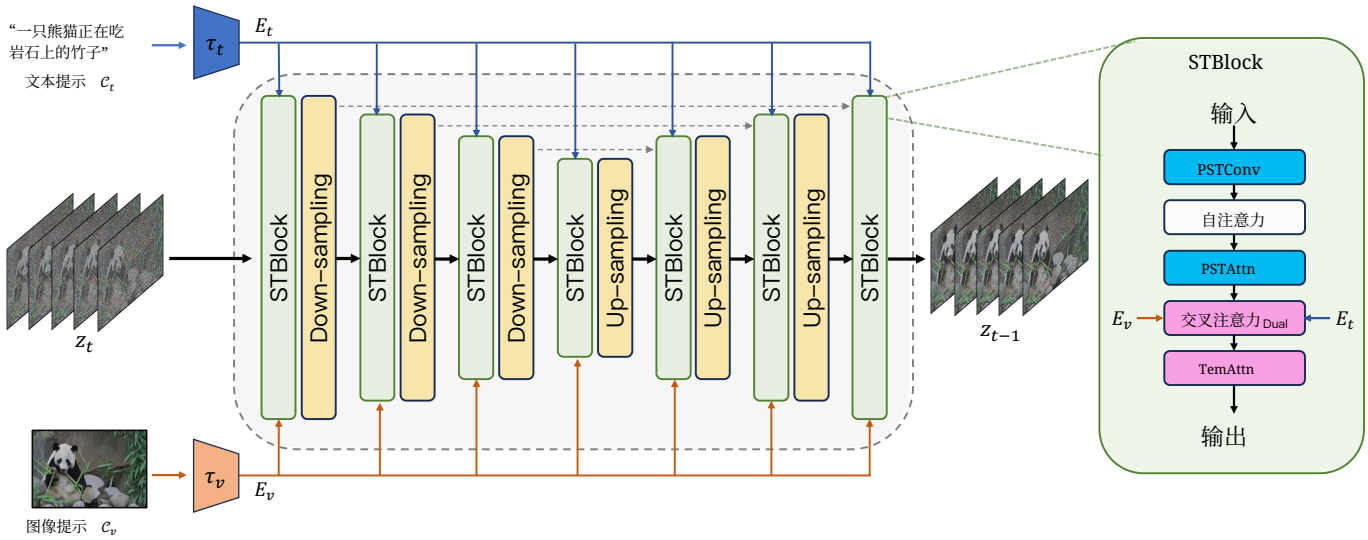

图 2. UniVid 概览。对于潜变量 z_t，我们使用文本提示 $\mathcal{C}_t$ 和图像提示 $\mathcal{C}_v$ 作为混合条件，引导去噪（生成）过程。具体而言，我们使用 CLIP文本编码器 τ_t 和 CLIP图像编码器 τ_v 分别获取语言标记嵌入 E_t 和视觉图块嵌入 E_v。UniVid 的 U-Net 架构由多个基本 STBlock 堆叠而成，每个块包含金字塔时空卷积（PSTConv）、自注意力（SelfAttn）、金字塔时空注意力（PSTAttn）、双流交叉注意力（CrossAttn）和时间注意力（TemAttn）。彩色版本效果最佳。

$H \times W$ f 称为 PSTAttn-f 和 PSTConv-f（f 越大，模块越靠近中间层）。在 PSTAttn-f 中，每一帧都会选择性地关注通过动态时间步长采样得到的一个帧子序列r_f，并探索帧间的像素级相关性。为此，我们将时间区间 $[1, 2, ..., F]$ 均匀划分为$n_f = \frac{F}{r_f}$个片段，然后从每个片段中采样n_f个参考帧，组成参考帧集合 Γ。形式上，可将其定义为 $\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V$，其中：

$$Q = W^Q z_{v_i}, K = W^K z_\Gamma, V = W^V z_\Gamma, \tag{6}$$

其中，z_{v_i} 表示帧 v_i 的特征，z_Γ 表示采样参考帧的拼接特征 Γ，投影矩阵 W^Q、W^K和W^V均为可学习参数。在 PSTConv-f中，我们在每个 $1D$ 卷积后添加一个具有动态核大小的时间卷积k_f。对于 PSTAttn-f 和 PSTConv-f，我们的时序调制金字塔时空机制表现为：随着 U-Net 层中的值 r_f 和 k_f 随 f 增大而增大。这意味着，我们通过选择更多参考帧并扩大时空相关性感知范围，逐步细化每帧的运动感知特征。

2）双流交叉注意力：尽管文本提示已成为视频合成中一种灵活的用户界面，但如何生成外观与用户提供的参考图像相似的视频，仍是一个极具实际意义的问题。本文旨在将文本描述和参考图像提供的控制方式都融入我们的 *Univid*，使输入的文本和视觉信息相互补充，从而提供详细的视频描述。我们的核心思路是将视觉控制引入原始的交叉注意力

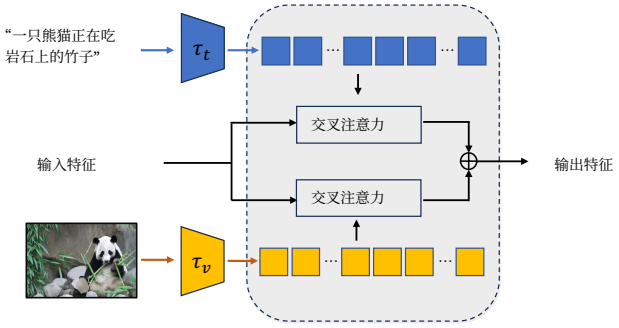

图 3. 我们提出的双流交叉注意力（$\text{CrossAttn}_{\text{Dual}}$）示意图。我们使用 CLIP文本编码器 τ_t 获取文本提示 $\mathcal{C}_t$的语言标记嵌入 E_t。对于图像提示 $\mathcal{C}_v$，我们使用对应的 CLIP图像编码器 τ_v 获取视觉图像块嵌入 E_v。然后，我们对输入特征与这些双模态嵌入执行双路径交叉注意力。输出特征由双流结果与系数 λ_t 和 λ_v融合得到。

模块，从而形成双流交叉注意力模块，即交叉注意力$_{\text{Dual}}$。

为缓解跨模态对齐的挑战，我们采用来自 CLIP [11] 的配对文本 τ_t 和图像编码器 τ_v，作为文本提示 $\mathcal{C}_t$ 和图像提示 $\mathcal{C}_v$ 的特征提取器。给定一组语言标记嵌入 $E_t = \tau_t(\mathcal{C}_t) = [e_t^1, e_t^2, ..., e_t^{N_t}]$ 和一组视觉图像块嵌入 $E_v = \tau_v(\mathcal{C}_v) = [e_v^1, e_v^2, ..., e_v^{N_v}]$，其中 N_t 和 N_v分别是语言标记和视觉图像块标记的数量，我们的双流交叉注意力模块可表示为：

$$\begin{aligned} z'_{v_i} &= \lambda_t \, \texttt{Attention}(Q, K_t, V_t) + \lambda_v \, \texttt{Attention}(Q, K_v, V_v) \\ &= \lambda_t \, \texttt{Softmax}(\frac{QK_t^T}{\sqrt{d}})V_t + \lambda_v \, \texttt{Softmax}(\frac{QK_v^T}{\sqrt{d}})V_v, \end{aligned} \tag{7}$$

其中 $Q = W^{Q} z_{v_i}$、 $K_t = W^{K_t} E_t$、 $V_t = W^{V_t} E_t$、 $K_v = W^{K_v} E_v$、 $V_v = W^{V_v} E_v$、 $W^{Q}, W^{K_t}, W^{V_t}, W^{K_v}, W^{V_v}$ 是投影矩阵， z_{v_i} 是第 v_i 帧的特征， z'_{v_i} 是输出特征。与一种仅使用全局图像令牌的类似技术 [19]不同，我们使用图像编码器生成的所有图像块令牌，因为我们的目标是以细粒度方式将每一帧的外观信息与参考图像精确对齐。

由于我们的交叉注意力$_{\texttt{Dual}}$会独立地将每一帧的语义与输入提示对齐，因此我们还引入了时间注意力层 `TemAttn` ，以在所有帧之间传播语义信息。该时间注意力层充当自注意力机制，专门沿时间轴进行运算，其形式如下（使用 Einops [20]表示法）：

$$
\begin{aligned}
z' &\leftarrow \texttt{rearrange}(z, \texttt{b t c h w} \rightarrow \texttt{(b h w) f c}) \\
z' &\leftarrow \texttt{TemAttn}(z') \\
z' &\leftarrow \texttt{rearrange}(z', \texttt{(b h w) f c} \rightarrow \texttt{b t c h w}),
\end{aligned} \tag{8}
$$

其中，b、t、c、h、w 分别表示批次、时间、通道、高度和权重的维度。

IV. 实验

本节将介绍我们的训练数据集、训练策略和实现细节。随后，我们将在零样本视频生成设置下，通过与现有方法进行比较，从定性和定量两方面评估我们的*Univid* 。

A. 实验设置

数据集 为了训练我们的模型，我们采用了两个公开可用的数据集，即 Laion 和 Pandas-70M。完整版本的 Laion 数据集包含 50 亿个图像-文本对。由于我们的模型继承了 Stable Diffusion 主干网络的二维参数，为节省训练时间，我们仅选取了 6 亿个带有较高美学评分的数据点，这些评分由 Laion 网站标注[2]。Pandas-70M [21] 数据集包含 7000 万个视频-文本对。我们从 Pandas-70M 数据中选取了 600 万个质量较高的样本 [22] 用于模型训练。

评估指标 对于 UCF-101，我们采用弗雷歇视频距离（FVD）和初始分数（IS） [23] 来评估所生成视频的质量和多样性。对于 MSR-VTT，我们使用弗雷歇 Inception 距离（FID） [24], 、CLIP-FID 和 CLIP-SIM [25]分数来评估我们的模型。

实现细节 我们使用公开提供的 Stable Diffusion 2.1 [5] 作为基础文本到图像（T2I）模型，并使用 OpenCLIP [26]作为图像编码器 τ_v。在实施同时利用图像和视频数据的联合微调策略时，我们针对时间层和空间层采用了不同的训练过程。对于图像输入，我们仅更新空间层。相比之下，对于视频输入，我们同时更新这两个层。

为了训练我们的(T+I)2V模型，我们采用从T2V到(T+I)2V的训练策略。为适应数据集的分辨率，我们将视频的输出尺寸设置为 24 x 576 x 320。首先，我们仅使用文本提示条件在Pandas-70M和LAION数据集上训练模型。在这一阶段，我们训练基于Stable Diffusion 2.1扩展而来的T2V模型。随后，我们整合T2V、图像到视频和(T+I)2V等任务，以训练适配器及相应的交叉注意力模块。最后，我们使用文本与图像联合条件对模型进行微调，更新完整U-Net主干网络的参数，从而进一步对齐并联合优化文本和图像之间的条件引导。

[2]https://laion.ai/blog/laion-aesthetics/

表 I UCF-101 和 MSR-VTT 上的零样本文本到视频生成。

方法	UCF-101		MSR-VTT	
	初始分数 (↑)	FVD (↓)	CLIP-SIM (↓)	FID (↓)
CogVideo [27]	25.27	701.59	0.2631	-
Make-A-Video [2]	33.00	367.23	0.3049	-
MagicVideo [28]	-	655	-	36.50
Latent-Shift [29]	-	360.04	0.2773	-
视频潜在扩散模型 [30]	33.45	550.61	0.2929	-
VideoFactory [31]	-	410.00	0.3005	-
PYoCo [3]	**47.76**	355.19	-	22.14
Show-1 [32]	35.42	394	**0.3072**	-
VideoPoet [33]	38.44	355	0.3049	-
Univid	43.23	**352.77**	0.3008	**22.06**

表 II UCF-101 和 MSR-VTT 上的零样本文本与图像以及仅图像视频生成。

方法	Task	UCF-101		MSR-VTT	
		初始分数 (↑)	FVD (↓)	CLIP-FID (↓)	FID (↓)
PYoCo [3]	T2V	**47.76**	355.19	9.73	22.14
EMU [34]	I2V	42.7	606.2	-	-
SVD [35]	I2V	-	242	-	-
DynamiCrafter [36]	(T+I) 2V	-	429.23	-	-
Univid	I2V	44.51	255.33	9.91	21.54
Univid	(T+I) 2V	44.98	**239.57**	**9.52**	**21.17**

B. 主要结果

1）定量比较：我们将本方法与 Make-A-Video [2],、CogVideo [27],以及若干同期工作 [28], [30], [31], [33], [36]进行了定量比较。表 I 显示，我们的方法在 UCF-101 数据集上的表现优于所有基线，并将零样本 FVD 从 701.59 提升至 352.77。这表明我们的方法生成的视频质量很高，且 Inception Score（IS）已达到先进水平，说明生成视频具有足够的多样性。同时，我们的方法在 MSR-VTT 数据集的各项指标上均表现出色，尤其是在 FID 指标 [25]上取得了最先进的结果。这不仅验证了我们的方法在文本到视频（T2V）任务中的有效性，也表明模型生成的视频具有足够高的多样性。

在表 II 中，对于图像到视频和 (T+I)2V 任务，可以看到我们的基线模型取得了新的最先进 FID 分数 21.17，而使用集成模型则进一步提升了 CLIP-FID 和 FID 分数。与其他图像到视频或 (T+I)2V 方法相比，我们的方法在 MSR-VTT 和 UCF-101 数据集上的零样本测试中均展现出显著的性能提升。这一提升表明，统一的框架和训练方法不仅保留了模型性能，

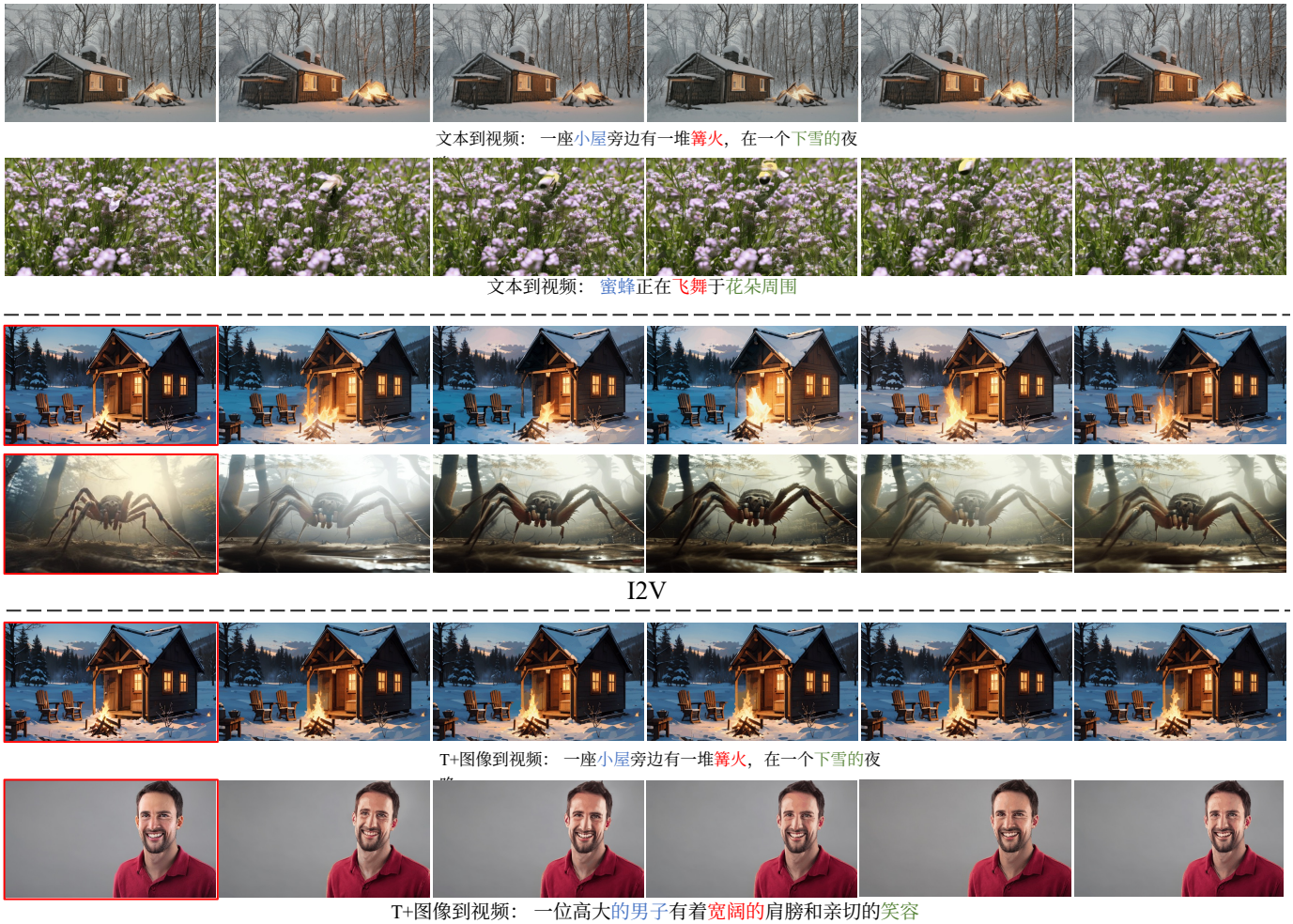

图 4. 图中展示了三个任务生成的示例：文本到视频、图像到视频以及 (T+I)2V。前两行展示文本到视频的结果，中间两行展示图像到视频的结果，最后两行展示 (T+I)2V 的结果。红色高亮的图像为参考图像。

还进一步提升了模型性能，从而改善了指令遵循、生成质量以及视频生成的多样性。这一发现具有重要意义。

2）可视化分析：我们在图 4 中展示了 *Univid* 在文本到视频、图像到视频和 (T+I)2V 三项任务上的定性可视化结果。在图 4 的第一个示例中，我们的 *Univid* 能够连贯地生成“雪夜中篝火旁的小屋”这一场景，并保持篝火持续跳动。在图 4 中间的两行里，输入一张参考图像以生成视频，生成的视频既能有效保留图像的原始结构，也呈现出逼真的运动效果。在图 4 底部的两行里，结合文本和图像来引导视频生成；生成的视频不仅保留了图像的信息内容，也能有效遵循文本提供的运动指令。

V. 结论

本文提出了一种统一视频生成模型（*Univid*），在单一框架中整合了文本到视频、图像到视频和 (T+I)2V 等多项任务。为支持这些任务，我们设计了一种双流交叉注意力机制，通过重新加权的控制分数，使文本和图像能够同时引导视频生成。此外，我们将原有的空间注意力和卷积模块改造为金字塔时空模块，以增强模型的时空建模能力，从而生成质量更高、连贯性更好的视频。实验结果和可视化结果表明，我们的方法在多项任务中表现出色，尤其善于保持语义一致性和时间一致性。这些结果证明了所提出的统一视频生成框架的有效性。

参考文献

[1] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、Ruiqi Gao、Alexey Gritsenko、Diederik P Kingma、Ben Poole、Mohammad Norouzi、

David J Fleet 等，《Imagen 视频：利用扩散模型生成高清晰度视频》，arXiv 预印本 arXiv:2210.02303，2022年。[2] Uriel Singer、Adam Polyak、Thomas Hayes、Xi Yin、Jie An、Songyang Zhang、Qiyuan Hu、Harry Yang、Oron Ashual、Oran Gafni 等，《Make-a-video：无需文本-视频数据的文本到视频生成》，载于《第十一届国际学习表征会议》，2022年。[3] Songwei Ge、Seungjun Nah、Guilin Liu、Tyler Poon、Andrew Tao、Bryan Catanzaro、David Jacobs、Jia-Bin Huang、Ming-Yu Liu 和 Yogesh Balaji，《保留自身相关性：视频扩散模型的噪声先验》，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第22930–22941页。[4] Andreas Blattmann、Robin Rombach、Huan Ling、Tim Dockhorn、Seungjun Wook Kim、Sanja Fidler 和 Karsten Kreis，《对齐你的潜变量：利用潜在扩散模型进行高分辨率视频合成》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2023年，第22563–22575页。[5] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer，《利用潜在扩散模型进行高分辨率图像合成》，载于 CVPR，2022年，第10684–10695页。[6] Ceyuan Yang、Yinghao Xu、Jianping Shi、Bo Dai 和 Bolei Zhou，《用于动作识别的时间金字塔网络》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第591–600页。[7] Chen-Lin Zhang、Jianxin Wu 和 Yin Li，《Actionformer：利用 Transformer 定位动作时刻》，载于《欧洲计算机视觉会议》。施普林格，2022年，第492–510页。[8] Tatsuya Harada、Yoshitaka Ushiku、Yuya Yamashita 和 Yasuo Kuniyoshi，《判别式空间金字塔》，载于 CVPR 2011。电气电子工程师学会，2011年，第1617–1624页。[9] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，《深度卷积网络中的空间金字塔池化及其视觉识别应用》，《IEEE模式分析与机器智能汇刊》，第37卷第9期，第1904–1916页，2015年。[10] Anna Bosch、Andrew Zisserman 和 Xavier Munoz，《利用空间金字塔核表示形状》，载于《第六届美国计算机协会图像与视频检索国际会议论文集》，2007年，第401–408页。[11] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，《从自然语言监督中学习可迁移视觉模型》，载于《国际机器学习会议》。机器学习研究论文集，2021年，第8748–8763页。[12] Jonathan Ho、Ajay Jain 和 Pieter Abbeel，《去噪扩散概率模型》，《神经信息处理系统大会》，第33卷，第6840–6851页，2020年。[13] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen，《利用 CLIP潜变量进行分层文本条件图像生成》，arXiv 预印本 arXiv:2204.06125，2022年。[14] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily L Denton、Kamyar Ghasemipour、Raphael Gontijo Lopes、Burcu Karagol Ayan、Tim Salimans 等，《结合深度语言理解的照片级真实感文本到图像扩散模型》，《神经信息处理系统进展》，第35卷，第36479–36494页，2022年。[15] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox，《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干预——2015年医学图像计算与计算机辅助干预会议：第18届国际会议，慕尼黑，德国，2015年10月5日至9日，论文集，第III部分18》。施普林格，2015年，第234–241页。[16] Diederik P Kingma 和 Max Welling，《自编码变分贝叶斯》，arXiv 预印本 arXiv:1312.6114，2013年。[17] Christoph Feichtenhofer、Haoqi Fan、Jitendra Malik 和 Kaiming He，《用于视频识别的 SlowFast网络》，载于《IEEE/CVF国际计算机视觉会议论文集》，2019年，第6202–6211页。[18] Bolei Zhou、Alex Andonian、Aude Oliva 和 Antonio Torralba，《视频中的时间关系推理》，载于《欧洲计算机视觉会议论文集（ECCV）》，2018年，第803–818页。[19] Hu Ye、Jun Zhang、Sibo Liu、Xiao Han 和 Wei Yang，《IP-Adapter：兼容文本的图像提示适配器，用于文本到图像扩散模型》，arXiv 预印本 arXiv:2308.06721，2023年。[20] Alex Rogozhnikov，《Einops：采用爱因斯坦式记法实现清晰可靠的张量操作》，载于《国际学习表征会议》，2021年。[21] Tsai-Shien Chen、Aliaksandr Siarohin、Willi Menapace、Ekaterina Deyneka、Hsiang-wei Chao、Byung Eun Jeon、Yuwei Fang、Hsin-Ying Lee、Jian Ren、Ming-Hsuan Yang 等人，“Panda-70m：利用多个跨模态教师为 7000 万个视频生成字幕”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第 13320–13331 页。[22] PKU-Yuan实验室和 Tuzhan AI 等，“Open-sora-plan”，2024年4月。[23] Jonathan Ho、Tim Salimans、Alexey Gritsenko、William Chan、Mohammad Norouzi 和 David J Fleet，“视频扩散模型”，*arXiv:2204.03458*，2022年。[24] Gaurav Parmar、Richard Zhang 和 Jun-Yan Zhu，“关于混叠缩放及 GAN 评估中令人惊讶的细微之处”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第 11410–11420 页。[25] Tuomas Kynk¨a¨anniemi、Tero Karras、Miika Aittala、Timo Aila 和 Jaakko Lehtinen，“ImageNet类别在 Fréchet 初始距离中的作用”，载于*ICLR*，2023年。[26] Gabriel Ilharco、Mitchell Wortsman、Ross Wightman、Cade Gordon、Nicholas Carlini、Rohan Taori、Achal Dave、Vaishaal Shankar、Hongseok Namkoong、John Miller、Hannaneh Hajishirzi、Ali Farhadi 和 Ludwig Schmidt，“OpenCLIP”，2021年7月。如果您使用此软件，请按以下格式引用。[27] Wenyi Hong、Ming Ding、Wendi Zheng、Xinghan Liu 和 Jie Tang，“CogVideo：通过 Transformer 实现文本到视频生成的大规模预训练”，*arXiv* 预印本 *arXiv:2205.15868*，2022年。[28] Daquan Zhou、Weimin Wang、Hanshu Yan、Weiwei Lv、Yizhe Zhu 和 Jiashi Feng，“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年。[29] Jie An、Songyang Zhang、Harry Yang、Sonal Gupta、Jia-Bin Huang、Jiebo Luo 和 Xi Yin，“Latent-Shift：通过时间偏移实现高效文本到视频生成的潜在扩散”，*arXiv* 预印本 *arXiv:2304 08477*，2023年。。[30] Andreas Blattmann、Robin Rombach、Huan Ling、Tim Dockhorn、Seung Wook Kim、Sanja Fidler 和 Karsten Kreis，“对齐你的潜变量：利用潜在扩散模型实现高分辨率视频合成”，载于*CVPR*，2023年。[31] Wenjing Wang、Huan Yang、Zixi Tuo、Huiguo He、Junchen Zhu、Jianlong Fu 和 Jiaying Liu，“VideoFactory：在时空扩散中采用交换注意力，实现文本到视频生成”，*arXiv* 预印本 *arXiv:2305 10874*，2023年。。[32] David Junhao Zhang、Jay Zhangjie Wu、Jia-Wei Liu、Rui Zhao、Lingmin Ran、Yuchao Gu、Difei Gao 和 Mike Zheng Shou，“Show-1：结合像素扩散模型与潜在扩散模型，实现文本到视频生成”，国际计算机视觉期刊，第 1–15 页，2024年。[33] Dan Kondratyuk、Lijun Yu、Xiuye Gu、José Lezama、Jonathan Huang、Grant Schindler、Rachel Hornung、Vighnesh Birodkar、Jimmy Yan、Ming-Chang Chiu 等人，“VideoPoet：用于零样本视频生成的大语言模型”，*arXiv* 预印本 *arXiv:2312.14125*，2023年。[34] Rohit Girdhar、Mannat Singh、Andrew Brown、Quentin Duval、Samaneh Azadi、Sai Saketh Rambhatla、Akbar Shah、Xi Yin、Devi Parikh 和 Ishan Misra，“Emu Video：通过显式图像条件分解文本到视频生成”，*CoRR*，第 abs/2311.10709 卷，2023年。[35] Andreas Blattmann、Tim Dockhorn、Sumith Kulal、Daniel Mendelevitch、Maciej Kilian、Dominik Lorenz、Yam Levi、Zion English、Vikram Voleti、Adam Letts 等人，“Stable Video Diffusion：将潜在视频扩散模型扩展至大规模数据集”，*arXiv* 预印本 *arXiv:2311.15127*，2023年。[36] Jinbo Xing、Menghan Xia、Yong Zhang、Haoxin Chen、Xintao Wang、Tien-Tsin Wong 和 Ying Shan，“DynamiCrafter：利用视频扩散先验为开放域图像制作动画”，*arXiv* 预印本 *arXiv:2310.12190*，2023年。