SKYREELS-V2：无限长度电影生成模型

SkyReels 团队
Skywork AI

摘要

近年来，视频生成的最新进展主要由扩散模型和自回归框架推动，但在保持提示一致性、视觉质量、运动动态和持续时间方面仍存在关键挑战：为了提升时间上的视觉质量而在运动动态上做出妥协；为了优先保证分辨率而限制视频时长（5-10秒）；以及由于通用型多模大语言模型（MLLM）无法理解电影语法（如镜头构图、演员表情和摄像机运动），导致拍摄感知生成不足。这些相互交织的限制阻碍了逼真的长片合成和专业电影风格的生成。为了解决这些问题，我们提出SkyReels-V2，一种无限长度的电影生成模型，结合了多模态大语言模型（MLLM）、多阶段预训练、强化学习和扩散强制框架。首先，我们设计了一种结合多模态LLM的通用描述和子专家模型的详细镜头语言的全面视频结构表示。在人工标注的帮助下，我们训练了一个统一的视频字幕器，命名为SkyCaptioner-V1，以高效标注视频数据。其次，我们建立了逐步分辨率的预训练方法，用于基础视频生成，随后进行四个阶段的后训练增强：初始概念平衡的有监督微调（SFT）提升基础质量；结合人工标注和合成畸变数据的运动特定强化学习（RL）训练，解决动态伪影问题；采用非递减噪声调度的扩散强制框架，实现高效搜索空间中的长视频合成；最终高质量的SFT微调，提升视觉逼真度。实验结果显示，我们在提示遵循性（尤其是镜头语言）、运动质量（具有充分动态）以及电影风格长视频生成能力方面达到了最先进水平，支持多种应用，如故事生成、图像到视频合成、摄像指导和元素到视频生成。所有代码和模型均可在https://github.com/SkyworkAI/SkyReels-V2获取。

1 引言

视频生成已成为生成式人工智能中的一个关键领域，推动了从创意内容制作到虚拟仿真的多种应用。虽然闭源扩散模型在商业应用中取得了成功，例如 Sora [1]、Keling1.6 [2]、Hailuo [3] 和 Veo2 [4]，但开源模型在缩小与闭源模型性能差距方面仍面临挑战，其中 Wan2.1 [5] 在公共基准测试中表现出显著提升，并在 V-bench 1.0 [6] 中排名，截止到 2025-02-24，位居第一。然而，这些视频生成模型在电影制作中的商业应用仍面临巨大挑战，包括随后的镜头语言提示中的文本对齐较差、缺乏高质量的运动动态以及时长有限（通常为 5 秒到 10 秒）。这些限制由多种因素造成。首先，大多数现有方法利用通用的多模态大语言模型（MLLM）对视频数据进行字幕，但在处理具有详细镜头语言的电影或影片场景时，文本对齐较差，导致生成结果失去专业电影的表现力。其次，这些模型的优化目标仍未充分探索，导致运动质量较差，而运动动态对电影制作尤为重要。标准的去噪损失主要关注逐帧外观学习，难以实现时间上的连贯性，正如 [7] 中分析的那样。尽管近期方法尝试采用偏好对齐技术以同时改善语义、美学和运动动态等所有指标，但每个指标的权重定义不明确，导致结果未达最佳。此外，视频生成框架主要由两种方法主导，包括扩散模型和自回归（AR）模型。扩散模型通过迭代去噪在视觉质量上设定了新标杆，而自回归模型在时间连贯性方面表现出色，但现有方法难以融合这两者的优势。例如，纯扩散模型常常生成视觉效果惊艳但时间上碎片化的输出，而 AR 模型则在时间一致性方面表现优异。

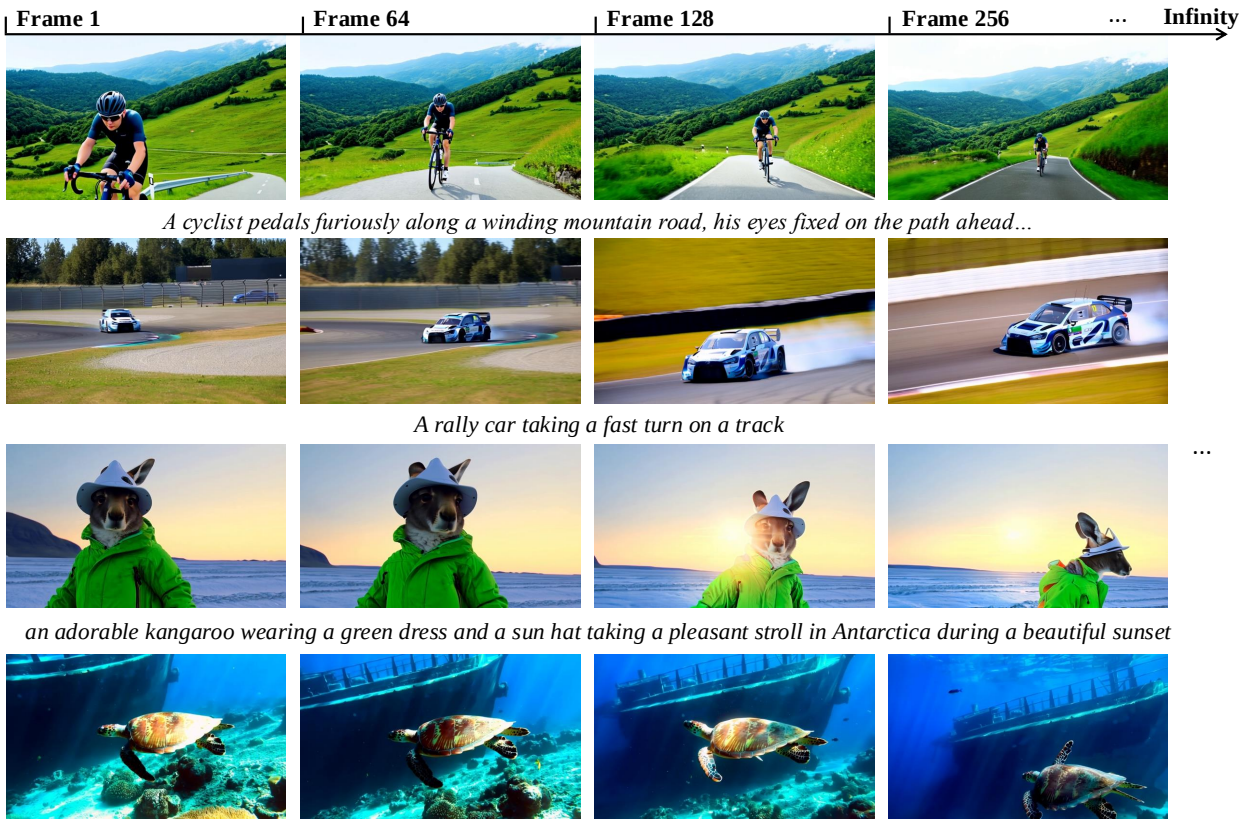

图1：SkyReels-V2 生成令人惊叹的逼真且具有电影感的高分辨率视频，长度几乎无限。该模型擅长在所有帧中保持主体的视觉一致性，确保没有失真，并在延长的视频序列中提供卓越的质量。

受到误差累积和分辨率降低的影响。由于这些限制，这两种方法都无法生成长时间的视频。为了结合高保真扩散方法和随意自回归方法的优点，一些研究人员提出了扩散强制变换器（DFoT）以弥合这一差距，但面临关键限制：DFoT的组合噪声调度导致训练不稳定。

为克服这些限制，我们提出了SkyReels-V2，它结合了多模态大语言模型（MLLM）、多阶段预训练、强化学习和扩散强制框架。我们的方法从对视频字幕的精心集成设计开始。我们为训练视频片段提出了一种结构化表示，包括主体类型、主体外观、表情、动作、位置等。其中一些字段可以被像Qwen2.5-VL [8]这样的通用MLLM模型很好地理解。这些字段需要专家模型，例如镜头类型、镜头角度、镜头位置、表情和相机运动。为了增强对这些字段的理解，我们训练了多个专家模型以实现准确描述。为了高效标注，我们将通用字幕模型和专家模型的知识蒸馏到一个统一的MLLM模型——SkyCaptioner-V1中。最终的文本提示由LLM细化，形成符合原始结构信息的多样化视频字幕描述。通过这些细致的文本描述，我们在逐步分辨率下预训练了基础扩散模型。之后，我们使用概念平衡数据进行第一次高质量的SFT阶段，为进一步优化奠定良好的初始化基础。然后，受到GPT-o1 [9]和deepseek-R1 [10]等LLM推理模型中强化学习成功应用的启发，我们通过偏好优化提升预训练模型的运动质量。为应对RL中数据标注的高成本，我们提出了一种半自动流程以生成偏好对。此外，为了实现长视频合成并减少收敛不确定性，我们没有从零开始预训练扩散强制模型，而是提出了一种扩散强制后训练方法，将预训练的全扩散模型微调为扩散强制模型。为了减少[11]中所述的去噪调度搜索空间，我们采用连续帧中的非递减噪声调度，大大缩小了组成空间的规模，从$O(1e48)$到$O(1e32)$。最后，我们在更高分辨率下训练模型，并应用蒸馏技术以实现高质量的商业应用。

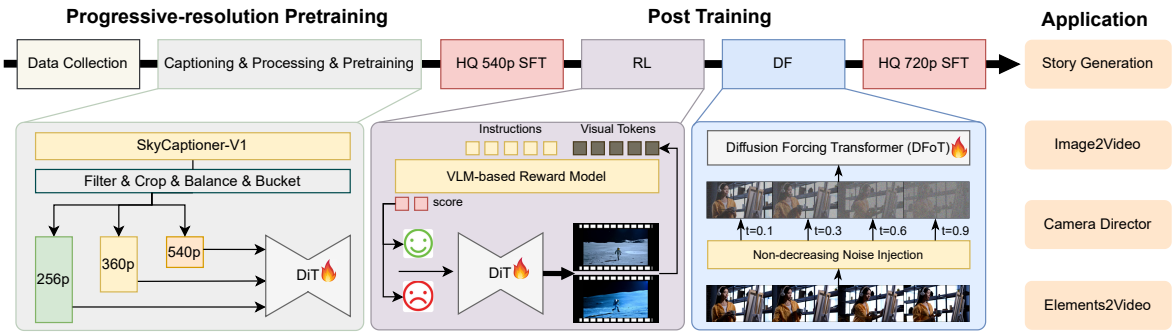

图2：所提出方法的概述。

应用。所提出的模型还支持多种应用，包括故事生成、图像到视频的合成、摄像机导演和元素到视频的生成。

大量实验展示了SkyReels-V2在性能上优于当前最先进的方法。据我们所知，它是第一个采用扩散强制架构的开源视频生成模型，在公开可用的模型中实现了最高的V-Bench分数。值得注意的是，我们的解决方案实现了前所未有的无限长度生成能力，如图1所示。通过SkyReels-Bench基准进行的人类评估进一步显示，我们的模型优于多个闭源替代方案，并在该领域的领先视频生成模型中表现出相当的结果。我们的主要贡献总结如下：

- 全面的视频字幕器，能够理解镜头语言，同时捕捉视频的整体描述，大大提高了提示的符合度。
- 运动特定偏好优化通过半自动数据收集流程增强运动动力学。
- 有效的扩散驱动适应能够生成超长视频和故事生成能力，为延长时间连贯性和叙事深度提供了一个强大的框架。
- SkyCaptioner-V1 和 SkyReels-V2 系列模型，包括扩散强制、文本转视频、图像转视频、摄像机导演和元素转视频模型，具有多种尺寸（1.3B、5B、14B），已开源。

2 相关工作

2.1 视频生成模型

过去一年，视频生成领域取得了显著的进展。虽然像 OpenAI Sora [1]、快手 Kling [2]、MiniMax 海洛 [3]、RunwayML Gen-4 [12] 和 Google Veo2 [4] 这样的闭源模型已取得商业成功，但开源替代方案正迅速缩小性能差距。早期的架构主要采用 2D 空间 + 和 1D 时间框架，如 Make-A-Video [13]、AnimateDiff [14]、Stable Video Diffusion [15] 等，这些架构逐渐演变成以 Video Diffusion Models [16] 和 CogVideoX [17] 为代表的复杂的 3D 全注意力系统。近期的开源实现，包括 HunyuanVideo [18]、StepVideo [19]、SkyReels-V1 [20]、OpenSora-2.0 [21] 和 Wan2.1 [5]，显示出其与专有系统在质量上的差异逐步缩小。

这些改进源于多方面的创新：从 U-Net [22] 到 DiT [23] 或 MMDiT [24] 结构的架构转变，增强的 VAE 实现 [25, 26, 27, 28, 29, 18, 5]，升级的文本编码器 [30, 31, 18, 32]，以及从 DDPM [33, 34] 到流匹配 [35, 24] 优化的范式转变。同时，优化的数据处理流程和视频字幕能力的提升（如 GPT-4o [36]、Qwen2.5-VL [8]、Gemini 2.5 [37]、Tarsier2 [38] 等）也极大地促进了质量的提升。当前的研究前沿扩展到强化学习的创新整合、混合自回归-扩散方法以及长视频生成技术。这些新兴方向有望弥合实现电影级视频合成的剩余差距。

2.2 扩散模型的对齐

强化学习与人类反馈（RLHF）在使大型语言模型与人类偏好对齐方面的成功[39]，激发了其在视觉生成任务中的应用。主要有两种代表性优化算法：(1) 奖励加权回归（RWR）方法[40, 41]，通过强化学习利用显式奖励模型对策略模型进行加权，从而优化策略；(2) 直接偏好优化（DPO）策略[42, 43, 44, 45, 19, 46]，通过直接优化偏好数据，绕过显式奖励建模。这些方法已被证明能够成功提升带有人类偏好的扩散模型的性能，增强美学和语义一致性。按照[46]中的框架，我们在流匹配中应用DPO方法以融入人类反馈。与之前的工作不同，我们主要关注运动质量，忽略文本对齐和视觉质量的优化，这些将由其他训练阶段改进。

奖励模型在对齐生成模型中扮演着重要角色，因为它们用于收集偏好数据。早期的工作采用了像 CLIP 分数 [30] 和图像质量分数 [47] 这样的指标作为奖励模型，以改善对视觉质量和文本对齐的评估。最近的方法 [46, 48, 49] 开始使用带有人类标注的数据集训练奖励模型，从而获得更准确和直接的结果。然而，这些方法中使用的生成数据相对较旧，导致训练得到的奖励模型在运动质量的人类对齐方面表现较差。考虑到收集和标注运动质量数据的高成本，我们提出了一种半自动数据采集流程，以扩展运动质量数据规模，从而在对齐性能上实现了显著提升。

2.3 扩散强迫框架

虽然现有的扩散模型在视频生成方面已取得显著成功，但它们仍然局限于生成固定长度的序列，缺乏大规模语言模型（LLMs）通过自回归令牌预测实现的无限序列扩展能力。以往试图将视频建模为下一个令牌预测的自回归方法受到误差累积问题的影响，性能不及基于扩散的方法。新兴的扩散强制（Diffusion Forcing）[50]范式通过建立一个基于独立噪声水平的下一个令牌预测机制，形成部分掩码，从而结合了扩散模型的高质量生成能力与自回归方法的无限扩展潜力。然而，该框架中扩展的搜索空间带来了显著的训练挑战。为此，AR-Diffusion [51]引入了一种新颖的非递减时间步约束，系统性地缩小搜索空间并稳定训练过程。此外，基于历史引导的视频扩散（History-Guided Video Diffusion）[52]通过扩展无分类器引导（CFG）[53]以适应可变长度的上下文帧条件，显著提升了对历史信息的利用效率。CausVid [54]提出了一种高效的适应策略，通过DMD蒸馏（DMD distillation）[55, 56]，实现了预训练双向扩散变换器到自回归扩散强制架构的直接转换，无需完全重新训练。此外，长上下文调优（Long Context Tuning）框架[57]采用了一种广泛的方法：在场景层面应用扩散强制，同时在镜头层面保持全序列扩散，从而实现无限扩展的故事生成，同时保持局部视觉质量。这些创新共同推动了长格式视频合成的前沿，通过扩散与自回归范式的协同融合。

3 种方法

在本节中，我们将全面介绍我们的方法论。图2展示了训练框架。我们首先在第3.1节详细介绍数据处理流程，然后在第3.2节解释视频字幕器的架构。接下来，我们在第3.3节描述我们的多任务预训练策略。随后，我们在第3.4节详细说明训练后优化技术，包括第3.4.1节的强化学习、第3.4.2节的扩散强制训练，以及第3.4.3节的高质量监督微调（SFT）阶段。我们还在第4节概述了用于训练和推理的计算基础设施。为了验证我们的方法，我们在第5节进行了与最先进基线的系统性比较。最后，我们在第6节展示了所提出模型的实际应用，包括故事生成、图像到视频合成、摄像指导和元素到视频的生成。

3.1 数据处理

数据处理是视频模型训练的基石，我们的框架集成了三个关键组件——*Data Sources*、*Processing Pipeline* 和 *Human-In-The-Loop Validation*——以确保稳健的质量控制。处理流程如图3所示，采用逐步过滤策略，从宽松到严格的标准，系统地在整个训练过程中减少数据量同时提升质量。

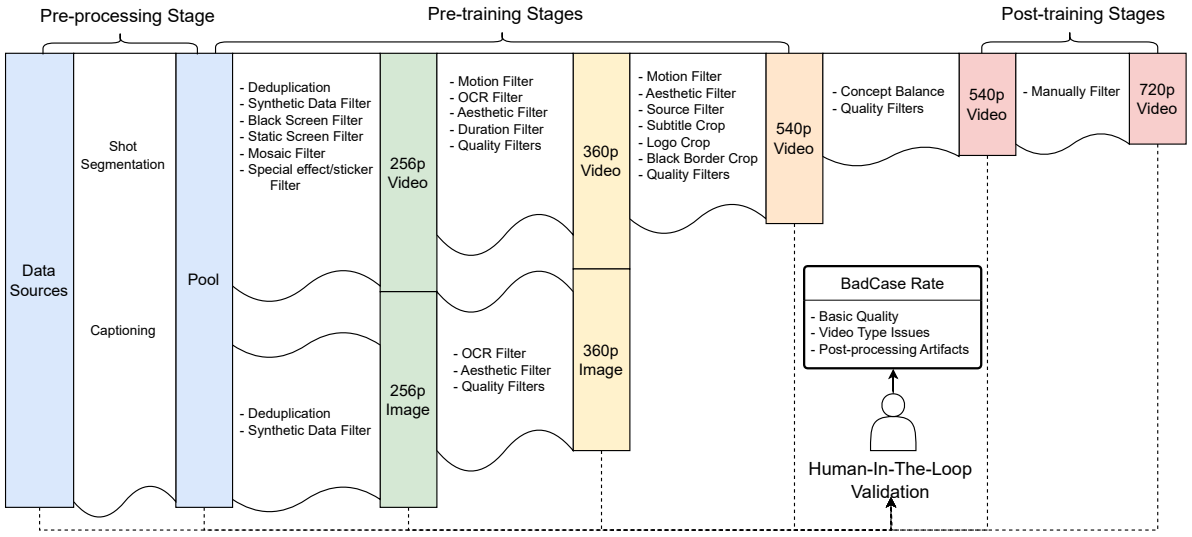

图 3：数据处理流程。

管道始于来自多样数据源的原始输入，然后通过一个自动化管道进行处理，旨在通过不同的过滤阈值控制样本的质量。我们管道的一个关键支柱是人机交互验证（Human-In-The-Loop Validation），它专注于对来自原始数据源和训练样本在不同阶段的抽样数据进行人工评估。通过在关键阶段——从初始数据导入到管道输出——进行系统的抽样检查，这一步确保了模糊、错误或不合规的数据被识别和纠正，最终保障了对稳健模型训练至关重要的最终数据质量。

3.1.1 数据来源

鉴于我们开发电影生成模型的目标，我们的多阶段质量保证框架整合了来自三个主要来源的数据：(1) 通用数据集整合了包括 Koala-36M [58]、HumanVid [59] 在内的开源资源，以及来自互联网的额外网页爬取视频资源。(2) 自行收集的媒体，包括280,000+部电影和800,000+集电视剧，覆盖120+个国家（估计总时长：6.2M+小时）。(3) 艺术存储库，包含来自互联网的高质量视频资产。原始数据集规模达到$O(100M)$，在不同的训练阶段根据质量要求使用不同的子集。我们还收集了$O(100M)$个概念平衡的图像数据，以在早期训练中加速生成能力的建立。

3.1.2 处理流程

如图3所示，为了获得训练数据池，对原始数据进行了两项预处理：镜头分割和字幕生成。之后，我们在不同的训练阶段使用一系列数据过滤器来处理数据质量问题。通过系统分析，我们将数据质量问题分为三类：1) *Basic quality*：低分辨率源、低帧率、黑白/静态屏幕、相机抖动、不稳定的运动以及任意镜头切换。2) *Video type issues*：监控录像、游戏录像、动画、无意义内容和静态视频。3) *Post-processing artifacts*：字幕、标志、图片编辑、分屏、黑边/模糊边界、画中画、速度变化以及特效/马赛克。这些问题的详细定义见表1。此外，我们还使用一些裁剪工具来修复特定的质量问题，并进行数据平衡，以确保模型的泛化能力。预训练阶段生成用于第3.3节多阶段预训练的数据。后训练阶段生成用于第3.4节后训练的数据。

预处理阶段 预处理阶段包括两个过程：1) *Shot Segmentation*：所有原始视频通过使用 PyDetect 和 TransNet-V2 [60] 进行镜头边界检测，并被分割成单镜头视频片段。 2) *Captioning*：分割好的单镜头片段使用我们在第3.2 节中描述的层次化字幕系统进行标注。 在预处理阶段之后，训练数据池将经过一系列数据过滤器，这些过滤器为不同的训练阶段设置不同的阈值。同时，引入数据裁剪器以解决若干数据质量问题。

数据过滤器的详细信息　在本部分，我们将解释数据过滤器的分类和细节。数据过滤器由元素过滤器和质量过滤器组成，用于在不同的训练阶段筛选数据。元素过滤器用于判断特定质量问题的严重程度。这些过滤器可以是基于分类的，用于检测问题的存在或类别，或者是基于分数的，用于为不同的质量要求设置不同的阈值。*Element Filters* 包括：1) *Black Screen Filter*：使用启发式规则检测黑屏数据。2) *Static Screen Filter*：计算基于流的得分以检测静态屏幕数据。3) *Aesthetic Filter*：依赖美学模型[61]获取得分。4) *Deduplication*：为了增强预训练集的多样性，我们利用复制检测嵌入空间[62]中的相似性，消除感知上冗余的片段。5) *OCR Filter*：分析文本的存在情况，计算文本的占用比例，并根据训练阶段裁剪数据。6) *Mosaic Filter*：使用训练好的专家模型检测马赛克区域。7) *Special effect/sticker Filter*：使用训练好的专家模型识别特效/贴纸。此外，我们还包括一些*Quality Filters*，如视频质量评估（VQA）模型[63, 64, 65]、图像质量评估（IQA）模型[66]和视频训练适用性评分（VTSS）[58]。在不同的训练阶段，我们将使用这些模型并设置不同的阈值来筛选数据。图3展示了在不同训练阶段应用不同数据过滤器的情况。

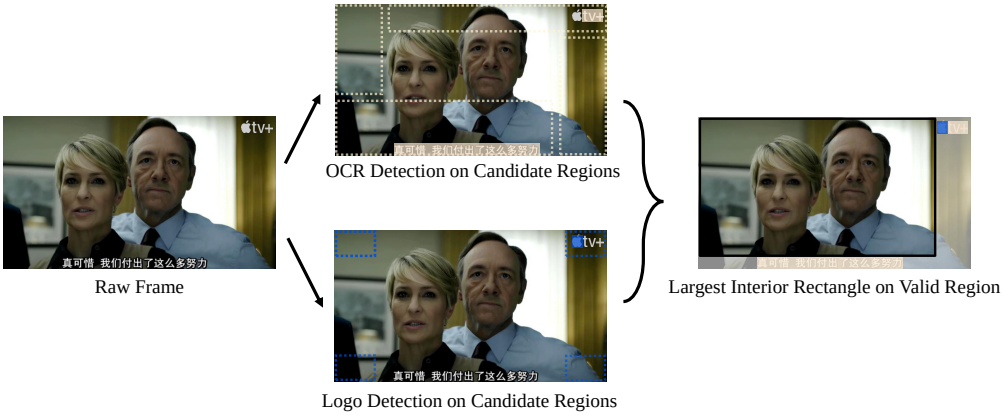

图4：字幕和标志裁剪的流程。字幕和标志在候选区域（虚线）中被检测到。然后，通过算法A1在检测边界框之外获得最大的内部矩形。

字幕和Logo裁剪的详细信息　我们的大部分训练数据来自电影和电视剧，这些视频可能包含字幕和频道logo，这会影响最终视频生成模型的质量。直接丢弃这些数据是浪费的。为了解决这个问题，我们对每个视频片段依次进行字幕检测、Logo检测和视频裁剪，以去除覆盖物，同时保持数据量。在字幕检测之前，我们使用启发式方法进行黑边裁剪，以裁剪黑边，确保更合理的字幕位置检测，从而提供更干净的数据。对于字幕检测，我们定义了四个潜在区域（画面边缘的顶部20%、底部40%、左侧20%和右侧20%）作为候选区域。然后，利用CRAFT模型[67]对这些区域在所有视频帧中进行OCR检测，并记录OCR边界框的坐标。同样，对于Logo检测，我们关注四个角区域（每个覆盖画面宽度/高度的15%），并采用MiniCPM-o模型[68]进行Logo检测和坐标记录。在视频裁剪阶段，我们首先构建一个与视频帧尺寸相匹配的二值矩阵，将检测到的字幕/Logo区域标记为0，其他区域标记为1。然后，应用单调栈算法（详见算法A1）识别包含全部1的最大内部矩形。如果该矩形覆盖原始画面面积的80%以上，并且其宽高比接近原始画面，则所有帧将根据该矩形的坐标进行裁剪，并保存为新的视频片段，不符合条件的数据将被丢弃。整个视频处理流程如图4所示。

后训练中的数据平衡 在后训练阶段，我们开始使用字幕者的主题类别进行详细的概念平衡，导致数据量减少了50%。未平衡和已平衡的概念按主要主题类型分组的比较如图5所示。在概念平衡后，我们还计算了每个一级类型下各个子类型的分布情况。表2提供了前五个主要类型的子类型统计的详细细节。

3.1.3 人类在环验证

人机交互验证涉及在数据生产的每个阶段进行人工视觉检查——*Data Sources*、*Shot*、*Segmentation*、*Pre-training* 和 *Post-training*——以确保用于模型训练的高质量数据。对于数据源，

表1：数据质量问题类别及定义

Category	Issue	Definition
Basic Quality	Low-resolution sources	Video sources with insufficient pixel density, typically below 720p resolution
	Low frame rates	Videos with frame rates below 16fps causing choppy motion
	Black/white/static screens	Frames containing blank screens or frozen images
	Camera shake	Unintentional camera movement causing unstable footage
	Unstable motion	Irregular object/camera movement creating visual discomfort
	Arbitrary shot transitions	Abrupt or mismatched scene changes without logical continuity
Video Type Issues	Surveillance footage	CCTV-style recordings with fixed angles and timestamps
	Game recordings	Screen captures of video game gameplay
	Animation	Computer-generated or hand-drawn non-live-action content
	Meaningless content	Videos lacking coherent narrative or visual purpose
	Static videos	Footage with minimal motion (e.g., still images with audio)
Post-processing Artifacts	Subtitles	Text overlays added during editing
	Logos	Watermarks or channel identifiers superimposed on video
	Image editing	Color grading, filters, or digital alterations
	Split screens	Multiple video streams shown simultaneously
	Black/blurred borders	Non-content areas added during post-production
	Picture-in-picture	Secondary video inset within main footage
	Speed variations	Altered playback speed (slow/fast motion)
	Special effects/mosaics	Added visual elements or pixelation overlays

人类主观评估原始数据是否适合使用。在镜头分割过程中，审查员会检查样本，以确保错误镜头比例低于1%，如错误转场等。在预训练阶段，数据会被过滤，并且会对0.01%的样本（每10,000个样本中1个）进行人工检查，以满足严格的限制：整体不良案例（如质量差、内容类型错误或处理错误）必须低于15%，其中基础质量问题 <占3%，视频类型问题 <占5%，后期处理缺陷 <占7%。在后训练阶段，采用相同的0.1%样本比例（每1000个样本中1个），但规则更为严格：不良案例总数必须低于3%，包括基础质量 <占0.5%，视频类型问题 <占1%，后期处理缺陷 <占1.5%。我们通过利用人工检查得出的不良案例比例，来判断数据源批次的可用性。如果某一批次的不良比例超过预设阈值，将采取相应措施，如丢弃或进一步优化该批次。此外，过滤参数会根据不同数据源的特性进行调整。例如，对于质量问题较多的数据源，会收紧与质量相关的过滤条件。每个阶段的逐步人工评估确保数据质量保持高水平，帮助模型有效训练。

3.2 视频字幕生成器

我们的影片字幕生成器旨在通过整合结构化字幕格式与专业子专家字幕员，生成精准的视频字幕。其目标包括：1）纠正多模态大语言模型（MLLM）中的错误或虚构信息。2）持续优化动态视频元素（例如：镜头信息、表情和摄像机运动）。3）根据应用场景（文本转视频或图像转视频）动态调整字幕长度。为了实现这些目标，我们设计了一种结构化字幕，如图6所示，提供多维度的描述信息，从不同角度包括：1）*Subjects*：具有外观、动作、表情、位置和层级类别/类型（例如：动物→哺乳动物）的主要和次要实体。2）*Shot Metadata*：镜头类型、角度、位置、摄像机运动、环境、照明等。我们使用基础模型Qwen2.5-VL-32B生成这些初始结构信息。然而，部分信息将由专家字幕员的结果替换，以获得更精确的描述。最后，我们生成最终字幕。

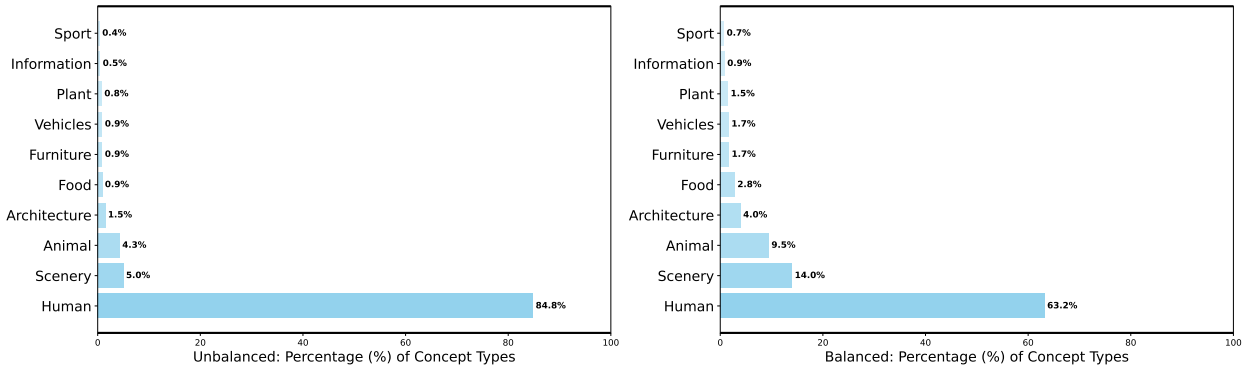

图5：不平衡（左）与平衡（右）概念分布的比较。

表2：前5大主要类型及其子类型比例统计

Primary type	Sub-type	Ratio	Primary type	Sub-type	Ratio
Human	Man	55.2%	Animal	Mammal	55.4%
	Woman	40.8%		Bird	18.5%
	Girl	1.9%		Aquatic Life	13.4%
	Boy	1.4%		Insect	7.5%
	Child	0.5%		Reptile	5.2%
	Baby	0.2%			
Scenery	Mountain	17.9%	Architecture	Historical	41.3%
	Seascape	16.4%		Commercial	19.1%
	Urban	12.0%		Industrial	16.3%
	River	10.6%		Residential	12.3%
	Beach	7.6%		Religious	11.1%
	Road	6.1%	Food	Snack	35.8%
	Lake	6.1%		Dessert	19.4%
	Sky	6.1%		Fruit	11.3%
	Forest	5.7%		Meat	10.7%
	Volcano	4.1%		Vegetable	8.6%
	Desert	2.6%		Seafood	6.9%
	Valley	2.0%		Dairy	4.7%
	Canyons	1.7%		Poultry	3.7%
	Cloud	1.2%			

通过不同模型的结构化数据融合。1) *Text-to-video*：生成密集描述。2) *Image-to-video*：关注“主体 + 时间动作/表达 + 相机运动”。每个字幕字段遵循10%的丢失率，以适应不同用户的情况，因为用户可能不会为每个字段提供精确的描述。字幕融合的详细信息显示在附录C中。

3.2.1 子专家字幕员

镜头字幕器 镜头字幕器由三个子字幕器组成，用于描述镜头的不同方面。它包括镜头类型、镜头角度和镜头位置。我们将这些方面定义为分类问题。 1) *Shot type*：特写镜头、极特写镜头、中景镜头、远景镜头和全景镜头。 2) *Shot angle*：仰视角镜头、俯视角镜头、低角度镜头。 3) *Shot position*：背面视图、正面视图、头顶视图、肩上视图、第一人称视角和侧面视图。

我们的训练方法采用精心设计的两阶段策略，以开发稳健的镜头分类器。在第一阶段，我们使用网络图片训练一个初步分类器，以建立基线性能（我们使用类别标签作为触发词，从网络爬取数据）。这个低精度模型主要用于从我们的电影数据集中提取跨所有目标类别的平衡真实场景数据。第二阶段则专注于通过手动标注真实电影数据，开发高精度的专家分类器，每个类别包含2,000个经过仔细标注的样本。

结构 标题

```
{
  "subjects": [
    {
      "appearance": "long platinum blonde hair styled in two braids, ...",
      "action": "The woman remains still, looking forward as ...",
      "expression": "The woman exhibits a neutral facial expression...",
      "position": "Centrally positioned in the frame.",
      "is_main_subject": true
    },
    {
      "appearance": "dark-colored with multiple masts, docked or anchored.",
      "action": "",
      "expression": "",
      "position": "In the background, blurred and indistinct.",
      "is_main_subject": false
    }
  ],
  "shot_angle": "eye level",
  "shot_position": "front view",
  "camera_motion": "use a handheld static shot",
  "environment": "Overcast sky with clouds, maritime setting ...",
  "lighting": "Soft and diffused lighting with no harsh shadows ..."
}
```

```
{
  "TYPES": {
    "type": "Human",
    "sub_type": "Woman"
  },
  "TYPES": {
    "type": "Vehicles",
    "sub_type": "Ship"
  },
  "shot_type": "extreme close-up shot",
}
```

图6：我们结构性标题的设计与演示

这些带注释的样本构成了我们最终高精度分类器的训练集，这些分类器专门针对真实电影视频中的镜头类型、镜头角度和镜头位置的准确分类进行优化。这种多阶段的训练方法确保了我们的训练数据集在类别上的平衡，以及在生产应用中的高分类准确率。

为了评估我们三个分类器的性能：镜头类型、镜头角度和镜头位置。我们构建了一个平衡的测试集，每个标签包含100个人工标注的样本。评估结果显示，镜头类型分类的平均准确率为82.2%，镜头角度分类的平均准确率为78.7%，而镜头位置分类的平均准确率为93.1%。虽然镜头位置分类器表现出色，但镜头类型和镜头角度分类器在未来仍有提升空间，特别是通过增强数据的平衡性以及提供更高质量的场景和角度标注。

表达字幕器　表达字幕器提供关于人类面部表情的详细描述，重点关注几个关键维度1) *Emotion Label*：情感被分类为七种常见类型，*i.e.*、*neutral*、*anger*、*disgust*、*fear*、*happiness*、*sadness* 和 *surprise*。2) *Intensity*：情感的强度被量化，例如“轻微愤怒”、“中等喜悦”或“极度惊讶”，表示情感的程度。3) *Facial Features*：影响情感表达的身体特征，包括眼型、眉毛位置、嘴角弯曲、皱纹和肌肉运动。4) *Temporal Description*：捕捉情感表现随时间的动态变化，关注情感的演变以及这些变化在视频中的时间点。

表情字幕生成包括两个阶段：1）我们首先检测并裁剪人脸，使用情感分类器对其情感进行分类。2）然后将情感标签和视频帧输入到VLM模型中，以生成详细的表情字幕。具体而言，我们采用S2D [69]的框架，并使用∼10k的内部数据集进行训练，重点关注人类和非人类角色。对于VLM模型，我们使用InternVL2.5，利用情感标签作为先验生成逐帧描述，并采用链式思维提示策略来优化描述并生成最终的表情字幕。

为了验证我们的情感分类器的性能，我们编制了一个包含1200个视频的测试集。该分类器在所有情感类别中的平均精确度达到了85%。对于表情字幕生成器，我们收集了560个视频样本，并邀请人工标注员评估模型在四个关键维度上的效果。该字幕器在情感标注方面的准确率为88%，情感强度评估为95%，面部特征识别为85%，以及时间描述的准确性为93%。这些结果突显了我们模型在捕捉视频内容中细腻情感和表现细节方面的鲁棒性。

相机运动字幕器　我们的框架采用分层分类策略，通过一个包含运动复杂度过滤、单类型运动建模和单类型运动数据整理的三阶段处理流程。 1) *Motion Complexity Filtering*：此阶段通过双重检测机制消除琐碎和过于复杂的运动。首先使用二元静态镜头检测器（准确率95%）筛查静止片段，然后利用针对不规则模式（手持抖动、目标跟踪、突变）训练的专用分类器对手工标注数据进行检测。存留的片段作为标准单类型运动处理。 2) *Single-type Motion Modeling*：我们使用6自由度（6DoF）坐标（平移x/y/z；旋转滚转/俯仰/偏航）对运动进行参数化，每个轴离散为负（-）、中性（0）或正（+）状态。结合三种速度等级（慢：<5%，中：5-20%，快：>20%每帧位移/秒），形成2197种不同的运动组合。训练数据结合手工标注和合成样本。 3) *Single-type Motion Data Curation*：我们实施五轮主动学习以高效扩展标注规模。从大约10k的人类标注样本作为基础训练开始，迭代预测10万未标注数据的标签，平衡采样1万预测结果进行验证，然后通过微调优化模型。此过程产生93k个高置信度样本，并补充16k个在运动轴上平衡的合成数据。合成数据确保每个自由度轴的正负状态比例相等。所有数据用于训练基于分类的字幕器以实现运动识别。

在一个包含15,000个视频的测试集上的评估结果显示：单一类型运动的预测准确率为89%，复杂运动的预测准确率为78%（手持）、83%（跟随主体）和81%（突发切换），静态镜头检测的准确率为95%。

3.2.2 SkyCaptioner-V1：一种结构化视频字幕生成方法　　模型

SkyCaptioner-V1 作为我们用于数据标注的最终视频字幕模型。该模型是在基础模型 Qwen2.5-VL-32B 的字幕结果以及子专家字幕员在平衡视频数据上的基础上训练而成。平衡视频数据是一个经过精心策划的数据集，包含大约 200 万个视频——从最初的 1000 万个样本中精选而出，以确保概念的平衡和标注的质量。

基于Qwen2.5-VL-7B-Instruction基础模型，SkyCaptioner-V1经过微调，以提升在特定领域视频字幕任务中的性能。为了与最新技术（SOTA）模型进行性能比较，我们使用包含1,000个样本的测试集，对不同字幕领域的准确性进行了人工评估。表3展示了详细内容。

结构化字幕中每个字段的准确率指标。提出的SkyCaptioner-V1在基线模型中实现了最高的平均准确率，并在与镜头相关的领域显示出显著的结果。

表3：在视觉理解测试集上的综合模型性能比较（所有模型均使用附录C中生成视频结构化字幕的相同系统提示。对于*Tarsier2-recap-7B*基线，我们实现了一个字幕转JSON的转换器，因为它不能直接输出结构化格式，适用于它们的有监督微调方法。）

model	Qwen2.5-VL-7B-Ins.	Qwen2.5-VL-72B-Ins.	Tarsier2-recap-7B	**SkyCaptioner-V1**
Avg accuracy	51.4%	58.7%	49.4%	**76.3**%
shot type	76.8%	82.5%	60.2%	**93.7**%
shot angle	60.0%	73.7%	52.4%	**89.8**%
shot position	28.4%	32.7%	23.6%	**83.1**%
camera motion	62.0%	61.2%	45.3%	**85.3**%
expression	43.6%	51.5%	54.3%	**68.8**%
TYPES_type	43.5%	49.7%	47.6%	**82.5**%
TYPES_sub_type	38.9%	44.9%	45.9%	**75.4**%
appearance	40.9%	52.0%	45.6%	**59.3**%
action	32.4%	52.0%	**69.8**%	68.8%
position	35.4%	48.6%	45.5%	**57.5**%
is_main_subject	58.5%	68.7%	69.7%	**80.9**%
environment	70.4%	**72.7**%	61.4%	70.5%
lighting	77.1%	**80.0**%	21.2%	76.6%

训练细节 我们采用 Qwen2.5-VL-7B-Instruct 作为基础模型，并在 64 个 NVIDIA A800 GPU 上以全局批量大小为 512 进行训练，使用 4 微批次和 2 梯度累积步骤。该模型使用 AdamW 进行优化，学习率为 1e-5，训练 2 个周期，并根据我们测试集的综合评估指标选择表现最佳的检查点。此训练配置确保了稳定收敛，同时保持了大规模视频字幕任务的计算效率。

3.3 多阶段预训练

我们采用Wan2.1[5]中的模型架构，并仅从零开始训练DiT，同时保留包括VAE和文本编码器在内的其他组件的预训练权重。然后，我们还使用Flow Matching框架[35, 24]来训练我们的视频生成模型。这种方法通过连续时间概率密度路径将复杂的数据分布转化为简单的高斯先验，从而通过常微分方程（ODEs）实现高效采样。

训练目标：给定潜在表示 $\mathbf{x}_1$ (图像或视频)，我们从对数正态分布 [24] 中采样一个时间步 $t \in [0, 1]$。然后，初始化噪声 $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$，并通过线性插值构建中间潜在 $\mathbf{x}_t$：

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0. \tag{1}$$

计算真实速度向量 $\mathbf{v}_t$，方法如下：

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0. \tag{2}$$

模型预测速度场 $\mathbf{u}_\theta(\mathbf{x}_t, \mathbf{c}, t)$，它引导样本 $\mathbf{x}_t$ 朝向样本 $\mathbf{x}_1$，并以文本嵌入 $\mathbf{c}$ (（例如，512维的umT5特征)) 为条件，通过最小化损失函数 L：

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\mathbf{c}}\left[\|\mathbf{u}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \mathbf{v}_t\|^2\right], \tag{3}$$

遵循此训练目标，我们首先设计了一个双轴分桶框架和FPS归一化方法，以对所有数据进行归一化。之后，我们进行三阶段预训练，分辨率逐步提高。

双轴分桶框架与FPS归一化　遵循第3.1节中描述的数据处理方法，我们通过双轴分桶框架解决视频数据的时空异质性。该框架沿两个正交维度组织训练样本：时间持续时间箱（B_T 分割）和空间宽高比类别（B_{AR} 分割），形成一个相互排斥的 $B_T \times B_{AR}$ 矩阵桶。为了优化GPU

为了在防止OOM失败的同时优化内存利用率，我们通过经验分析实现了自适应批次大小——每个桶根据其持续时间和宽高比分配不同的最大批次容量。在数据预处理阶段，样本被映射到其最近的桶中。在模型训练过程中，分布式计算节点采用随机桶采样，动态组装小批量，确保输入分辨率和时间跨度的持续变化。在双轴分桶系统（B_T 时间桶 × B_{AR} 空间宽高比类别）的基础上，我们扩展了框架，加入了时间频率自适应。视频通过残差感知的降采样协议进行FPS归一化：对于每个样本，我们计算相对于目标频率（16/24 FPS）的模余数，选择余数最小的频率作为重采样基础。这一数学表达式：$f_{\text{target}} = \arg\min_{f \in \{16,24\}} (\text{original\_fps} \bmod f)$，确保了最佳的时间对齐，同时保持运动语义。重采样后的视频使用已建立的持续时间-宽高比矩阵进行分桶。为了消除帧率依赖性，我们在DiT架构中加入了可学习的频率嵌入，与时间步嵌入进行加性交互。这些可学习的频率嵌入将在我们仅使用FPS-24视频数据进行高质量SFT阶段后被废弃。

预训练阶段1 我们首先在低分辨率数据（256p）上进行预训练，以捕捉基本的生成能力。在此阶段，我们提出了联合图像-视频训练，并支持不同的宽高比和帧长度。我们实施严格的数据过滤，去除低质量和合成数据，并进行去重以确保数据的多样性。这个低分辨率阶段帮助模型从大量样本中学习低频概念。此阶段训练的模型展示了基本的视频生成能力，尽管生成的视频仍然相对模糊。

预训练阶段2 在此阶段，我们继续进行联合图像-视频训练，但将分辨率提高到360p。我们采用了更为复杂的数据过滤策略，包括时长过滤、运动过滤、OCR过滤、美学过滤和质量过滤。在这一训练阶段之后，生成视频的清晰度显著提高。

预训练阶段3 在最后的预训练阶段，我们将分辨率进一步提升到540p，专注于视频目标。我们实施了更严格的运动、美学和质量筛选标准，以确保高质量的训练数据。此外，我们引入源过滤，去除用户生成的内容，同时保留电影数据。这种方法提升了生成视频的视觉质量，并显著增强了模型生成具有优越纹理和电影品质的逼真人类视频的能力。

预训练设置　在优化方面，我们在所有预训练阶段都采用AdamW 优化器。在第1阶段，我们将学习率初始化为 1e-4，权重衰减设置为 0。一旦损失收敛到稳定范围，我们将学习率调整为 5e-5，并引入 1e-4 的权重衰减。在第2和第3阶段，我们进一步将学习率降低到 2e-5。

3.4 训练后

训练后阶段是提升模型整体性能的关键环节。我们的训练后阶段包括四个阶段：高质量的540p SFT、强化学习、扩散强制训练以及高质量的720p SFT。为了提高效率，前三个阶段在540p分辨率下进行，而最后一个阶段在720p分辨率下完成。540p的高质量SFT利用平衡数据提升整体性能，为后续阶段提供更好的初始化。为了增强运动质量，我们将依赖强化学习而非标准的扩散损失。在此阶段，我们提出了一种半自动化流程，从人类和模型两方面收集偏好数据。此外，我们还提出了扩散强制训练阶段，在该阶段中，我们将全序列扩散模型转变为扩散强制模型，该模型应用帧特定的噪声水平，从而实现多长度视频的生成能力。之后，我们进行720p的高质量SFT阶段，将生成分辨率从540p提升到720p。

3.4.1 强化学习

受到之前在大规模语言模型[9, 10]成功的启发，我们提出通过强化学习来提升生成模型的性能。具体来说，我们关注运动质量，因为我们发现我们的生成模型的主要缺点是：1）生成模型在处理大规模、可变形的运动时表现不佳（图7.a，图7.b）；2）生成的视频可能违反物理定律（图7.c）。

为了避免在其他指标上的退化，例如文本对齐和视频质量，我们确保偏好数据对具有可比的文本对齐和视频质量，而只有运动质量不同。这一要求在获取偏好标注时带来了更大的挑战，因为人工标注本身成本较高。为了解决

在这个挑战中，我们提出了一种半自动化的流程，策略性地结合了自动生成的运动对和人工标注的结果。这种混合方法不仅扩大了数据规模，还通过精心筛选的质量控制改善了与人类偏好的对齐。利用这一增强的数据集，我们首先训练一个专门的奖励模型，以捕捉成对样本之间的通用运动质量差异。随后，这个学习到的奖励函数指导直接偏好优化（DPO）的样本选择过程，提升生成模型的运动质量。

(a) V2V失真示例：生成视频中角色的脸部出现轻微的扭曲。

(b) I2V失真示例：生成视频中的人物经历严重的身体变形。

(c) T2V失真示例：生成的视频显示篮球向上升起而非向下落下（违反重力）。

图7：我们逐步扭曲创建过程产生的各种扭曲类型的示例。

偏好数据由人工注释 通过对生成视频中的运动伪影进行严格分析，我们建立了一个系统的常见失败模式分类：过度/不足的运动幅度、主体变形、局部细节损坏、物理定律违反以及不自然的运动。此外，我们还记录了与这些失败模式对应的提示，并由大型语言模型（LLMs）生成相同类型的提示。这些生成的提示多样，涵盖人类与动物互动、物体运动等所有上述运动失败类型。然后，使用每个提示在我们的预训练模型的历史检查点池中生成四个样本。

在样本收集阶段之后，针对同一提示的样本被系统地配对成样本对。邀请专业的人类标注员对这些样本对的偏好进行评分。我们的标注流程主要包括两个步骤：1) *Data filtering*：样本将在以下两种情况下被排除：首先，内容/质量不匹配——如果两个样本描述的文本内容不同或在视觉质量上存在显著差异，以确保关注运动

质量分析；第二，标注标准失效——如果样本对中的任意一个未能满足三个标准中的任何一项——主要对象的清晰度、图像框内的主体大小是否足够，或背景构图的简洁性。根据我们的经验，这一过程将在进行进一步标注之前，剔除几乎80%的数据对。 2) *Preference selection*：人工标注员根据运动质量标准为每对样本分配三种标签之一（更好/更差/平局）。人工标注的运动标准详情列在表A2中，提供了所有运动质量失效类型的描述。每种失效类型都被赋予一个加权分数，两个视频的总评分将被计算，以便进行比较。

偏好数据自动生成　在我们严格的质量标准下，人类标注的资源密集型特性极大地限制了数据集的规模。为了扩充偏好数据集，我们设计了一个自动偏好数据生成流程，包括两个核心步骤:

1) 真实数据收集 我们使用生成的提示词来查询现有的数据集，通过计算它们的 CLIP 特征[30]之间的余弦相似度，获得相似的提示词。与语义匹配的提示词相关联的经过筛选的真实参考视频作为选中的样本。被拒绝的样本通过以下步骤生成，以形成偏好对。

2) 逐步扭曲创建　基本观察是，最先进的视频生成模型在运动质量方面仍然不及真实视频。我们通过故意向真实视频添加可控的扭曲，创建系统性运动缺陷的模拟。每个真实视频都配有一段文本描述（内容说明）和其第一帧（静态参考），实现动态缺陷分析，同时保持视觉结构。我们创建了三种不同的损坏样本变体：V2V：对噪声潜在表示的直接反转（最低扭曲）；I2V：使用第一帧引导的重建（中等扭曲）；T2V：根据文本描述重新生成（最高扭曲）。此外，我们还使用不同的生成模型（[5, 18, 17]）和模型参数（例如时间步长）来构建不同水平的运动质量，同时保持样本的多样性。图7展示了通过我们的流程自动化构建的三个案例。

超越标准程序，我们的研究还探索了创新技术以引发特定的视频质量问题。我们可以在时间域中操控帧采样率，增加或减少它们以产生过度或不足的运动幅度效果，或交替使用不同的采样率以制造不规则的运动。利用Tea-Cache [70] 方法，我们可以调节参数并注入噪声，以破坏视频帧中的局部细节。对于汽车行驶或鸟类飞行等场景，我们通过倒放视频创建配对，挑战模型区分正确与错误的物理运动。这些方法在模拟视频生成中的各种不良情况方面非常有效。它们能够准确复制异常运动、局部细节丢失和物理异常等场景，这些都是在视频生成过程中可能发生的违背常理的动作。

奖励模型训练 继 VideoAlign [46] 之后，我们使用 Qwen2.5- VL-7B-Instruct [8] 实现了我们的运动质量奖励。训练数据来自上述数据收集过程，形成了总共 $30k$ 个样本对。由于运动质量与上下文无关，样本对不包括提示。模型采用带平局的 BradleyTerry 模型（BTT）[71] 进行训练，这是 BT 的扩展，考虑了平局偏好:
$\mathcal{L} = -\sum_{(i,j)} [y_{i>j} \ln P(i > j) + y_{i<j} \ln P(i < j) + y_{i=j} \ln P(i = j)]$。其中 $i > j$、$i < j$、$i = j$ 表示样本 i 在样本对中优于/劣于/等于样本 j。

DPO 训练 我们采用来自 [46] 的流式直接偏好优化（Flow-DPO）来提升我们的生成模型的运动质量。损失函数可以定义为:

$$\mathcal{L}_{\text{DPO}} = -\frac{1}{N}\sum_{i=1}^{N} \cdot \log\sigma\left(-\frac{\beta}{2}\Big[\underbrace{(L^w_{\text{model}} - L^l_{\text{model}})}_{\Delta_{\text{model}}} - \underbrace{(L^w_{\text{ref}} - L^l_{\text{ref}})}_{\Delta_{\text{ref}}}\Big]\right) \tag{4}$$

其中:

$$L^w_{\text{model}} = \frac{1}{2}\|\hat{\mathbf{y}}^w_{\text{model}} - \mathbf{y}\|_2^2, \quad L^l_{\text{model}} = \frac{1}{2}\|\hat{\mathbf{y}}^l_{\text{model}} - \mathbf{y}\|_2^2$$
$$L^w_{\text{ref}} = \frac{1}{2}\|\hat{\mathbf{y}}^w_{\text{ref}} - \mathbf{y}\|_2^2, \quad L^l_{\text{ref}} = \frac{1}{2}\|\hat{\mathbf{y}}^l_{\text{ref}} - \mathbf{y}\|_2^2$$
$$\Delta_{\text{model}} = L^w_{\text{model}} - L^l_{\text{model}}, \quad \Delta_{\text{ref}} = L^w_{\text{ref}} - L^l_{\text{ref}}$$

其中 β 是温度系数。$\hat{\mathbf{y}}^{w/l}_{\text{model}}$ 是当前模型对被选择/拒绝样本的预测。以及 $\hat{\mathbf{y}}^{w/l}_{\text{ref}}$ 是参考模型对被选择/拒绝样本的预测。

为了收集这些训练样本，我们构建了两种类型的提示集：概念平衡提示（用于多样性）和运动特定提示（用于运动质量）。每个提示用于使用我们的生成模型生成8个视频。然后，我们使用运动质量奖励模型对视频进行排序，选择最佳视频和最差视频，形成一个样本三元组（选择的视频、被拒绝的视频、提示）。请注意，我们的DPO训练是分阶段进行的。当模型开始能够轻松区分选择的样本和被拒绝的样本（表明性能达到平台期）时，我们会用最新的迭代版本刷新参考模型。更新后的参考模型随后生成新数据，这些数据由奖励模型进行排序，形成下一阶段的训练数据。每个阶段需要20k个训练数据，我们总共进行3个阶段的DPO训练。

3.4.2 扩散驱动力

在本节中，我们介绍扩散强制变换器（Diffusion Forcing Transformer），它解锁了我们模型生成长视频的能力。扩散强制[50]是一种训练和采样策略，其中每个标记被赋予一个独立的噪声水平。这允许使用训练好的模型根据任意的、每个标记的时间表对标记进行去噪。从概念上讲，这种方法类似于部分掩码：噪声为零的标记完全未掩码，而完全噪声则完全掩码它。扩散强制训练模型“解掩”任何组合的不同噪声标记，利用较干净的标记作为条件信息引导噪声标记的恢复。在此基础上，我们的扩散强制变换器可以基于前一段的最后几帧无限延伸视频生成。注意，同步全序列扩散是扩散强制的一个特殊情况，其中所有标记共享相同的噪声水平。这种关系使我们能够从全序列扩散模型微调扩散强制变换器。

受到 AR-Diffusion [11] 的启发，我们采用帧导向概率传播（FoPP）时间步调度器进行扩散强制训练。该过程包括以下步骤：

1. 均匀采样：首先，我们均匀采样一个帧索引 $f \sim U(1, F)$ 和一个对应的时间步 $t \sim U(1, T)$。这确保了时间步在所有视频帧中均匀分布。 2. 动态规划用于概率传播：利用动态规划，我们计算在条件 $t_f = t$ 下，采样帧之前和之后的帧的时间步的概率。 3. 转移方程的定义：我们将 $d^s_{i,j}$ 定义为在非递减约束下，从帧 i 开始，时间步为 j 的有效时间步序列的计数。我们使用转移方程计算 $d^s_{i,j}$：

$$d_{i,j} = d_{i,j-1} + d_{i-1,j}$$

边界条件 $d_{*,T} = 1$ 和 $d_{F,*} = 1$。

4. 访问概率计算：对于 f 之后的帧，访问时间步 k 的概率为：

$$\frac{d^s_{i,k}}{\sum_{j=K}^{T} d^s_{i,j}}$$

类似地，对于 f 之前的帧，我们定义 $d^e_{i,j}$ 并计算概率为：

$$\frac{d^e_{i,k}}{\sum_{j=1}^{K} d^e_{i,j}}$$

5. 时间步采样：最后，根据计算出的概率，逐个采样前一帧或后一帧的时间步。

在推理过程中，我们采用支持自适应视频生成的自适应差分（AD）时间步调度器 [11]，能够兼容异步自回归和同步生成。

AD调度器将相邻帧之间的时间步差异视为一个自适应变量s。对于连续的帧，其时间步为t_i和t_{i-1}，条件为：

$$t_i = \begin{cases} t_i + 1, & \text{if } i = 1 \text{ or } t_{i-1} = 0, \\ \min(t_{i-1} + s, T), & \text{if } t_{i-1} > 0 \end{cases}$$

当前一帧为空或干净时，当前帧专注于自我去噪。否则，它会以与上一帧的时间步差为s进行去噪。值得注意的是，同步扩散（$s = 0$）和自回归生成（$s = T$）是特殊情况。较小的s会产生更相似的邻近帧，而较大的s则增加内容的多样性。

我们的条件机制通过利用更干净的历史样本作为条件，实现了自回归帧的生成。在这个框架中，信息流本质上是有方向的：噪声样本依赖于前面的历史以确保一致性。这种有方向的特性意味着不需要双向注意力，可以用更高效的因果注意力来替代。在用双向注意力训练扩散强制变换器后，可以用上下文因果注意力对模型进行微调，以提高效率。在推理过程中，这种架构可以缓存来自历史样本的 K、V 特征，消除冗余计算，显著降低计算开销。

3.4.3 高质量有监督微调（SFT）

我们在540p和720p分辨率下分别实施了两个连续的高质量有监督微调（SFT）阶段，初始的SFT阶段在预训练之后立即进行，但在强化学习（RL）阶段之前。第一阶段的SFT作为一个概念平衡训练器，建立在仅使用fps24视频数据进行预训练的基础模型之上，同时有策略地移除FPS嵌入组件以简化架构。该阶段采用第3.1.2节中详细描述的高质量概念平衡样本进行训练，为后续训练过程建立了优化的初始化参数。在此之后，我们在完成扩散强制阶段后，在720p进行第二次高分辨率SFT，采用相同的损失函数和通过人工筛选的更高质量的概念平衡数据集。这一最终的细化阶段专注于提高分辨率，从而进一步提升整体视频质量。

4 基础设施

在本节中，我们介绍　在训练和推理期间利用基础设施优化　　参考阶段。

4.1 训练优化

训练优化旨在确保高效且稳健的训练，包括 *memory optimization*、*training*、*stability* 和 *parallel strategy*，这些将在以下段落中详细说明。

内存优化　　注意力块的fp32 内存绑定操作占据了 GPU 内存的主要部分。我们通过高效的算子融合来解决这一问题，减少内核启动开销，同时优化内存访问和利用率，从而提升性能。梯度检查点（GC）通过仅在 fp32 中存储变换器块的输入来最小化内存使用；将这些转换为 bf16 可以将内存需求降低 50%，且几乎不影响精度。激活转移（Activation offloading）通过异步将临时张量移动到 CPU，进一步节省 GPU 内存，同时保持吞吐量。然而，由于 8 个 GPU 共享 CPU 内存，以及过度的转移导致的计算重叠有限，我们策略性地将 GC 与选择性激活转移相结合，以实现最佳效率。

训练稳定性 我们提出了一个智能自愈框架，通过三阶段修复实现自主故障恢复：实时检测和隔离受损节点，使用备用计算单元进行动态资源重新分配，以及通过检查点恢复进行任务迁移，以确保模型训练的连续性。

并行策略 我们预先计算VAE和文本编码器的结果。使用FSDP在所有节点上分布式存储DiT的权重和优化器状态，以应对由模型规模庞大引起的GPU内存压力。在以720p分辨率进行训练时，由于临时张量较大，我们遇到了严重的GPU内存碎片化问题，即使内存仍然充足，也会触发torch.empty_cache()。因此，我们使用Sequence Parallel [72] 来解决激活引起的内存压力。

4.2 推理优化

推理优化的关键目标是降低视频生成的延迟，同时不影响质量。虽然基于扩散的模型在生成高保真度视频方面取得了成功，但在推理过程中需要多步采样，通常需要30到50步，可能需要超过5分钟才能生成5秒的视频。在我们的实际部署中，我们通过 *VRAM optimization*、*quantization*、*multi-GPU parallel* 和 *Distillation* 实现了优化。

VRAM 优化 我们的部署利用 RTX 4090 GPU（24GB VRAM）为 14B 参数模型提供服务。通过结合 FP8 量化和参数级卸载技术，我们成功实现了 720p 视频生成，同时在单个 GPU 实例上保持了完整的模型能力。

量化我们的分析确定注意力和线性层是DiTs中的主要计算瓶颈。为了优化性能，我们在整个架构中实现了FP8量化。具体而言，我们将FP8动态量化与FP8 GEMM加速相结合，用于线性层，在RTX 4090硬件上实现了*1.10*×的加速，相较于bf16基线。在注意力操作方面，我们部署了sageAttn2-8bit[73]，在同一RTX 4090平台上实现了比bf16实现更快的*1.30*×推理速度。

并行策略 为了加快单视频生成速度，我们采用了三种关键的并行策略：内容并行、CFG并行和VAE并行。在实际部署中，当从4个到8个RTX 4090 GPU扩展时，这种方法将整体延迟降低了*1.8*×。

蒸馏 为了加快视频生成速度，我们采用了 DMD 蒸馏技术 [55, 56]。我们去除了回归损失，使用高质量的视频数据代替纯噪声作为学生生成器的输入，以加快模型收敛。此外，我们还采用了两时间尺度的更新规则，以确保伪评分生成器跟踪学生生成器的输出分布，以及来自 DMD 的多步调度。类似地，如公式所示，梯度被用来更新学生生成器 G。

$$\nabla_\theta D_{KL} \simeq \mathbb{E}_{t,x}\left[\left(s_{\text{fake}}(x,t) - s_{\text{real}}(x,t)\right)\frac{dG}{d\theta}\right]$$

其中 x 表示由学生生成器生成的视频，s_{fake} 和 s_{real} 分别代表由假分数生成器和真实分数生成器产生的评估分数。我们将假分数生成器和学生生成器的更新比例设为 5，并使用具有特定调度的 4 步生成器，针对流匹配框架进行调优。在蒸馏阶段，我们发现较小的学习率结合较大的批量大小对于稳定训练非常重要。通过上述蒸馏过程，我们可以显著减少视频生成所需的时间。

5 性能

为了全面评估我们提出的方法，我们构建了 SkyReels-Bench 供人工评估，并利用开源的 V-Bench 进行自动评估。这使我们能够将我们的模型与最先进（SOTA）基线进行比较，包括开源和专有模型。

5.1 SkyReels-Bench

为了人工评估，我们设计了包含1020个文本提示的SkyReels-Bench，系统地评估三个维度：指令遵循、运动质量、一致性和视觉质量。该基准旨在评估文本到视频（T2V）和图像到视频（I2V）生成模型，提供跨不同生成范式的全面评估。

指令遵循评估生成视频与提供的文本提示的符合程度。 1) *Motion instruction adherence*：执行指定动作或动作的准确性。 2) *Subject instruction adherence*：对描述的主体和属性的正确表现。 3) *Spatial relationships*：主体之间的正确定位和互动。 4) *Shot adherence*：对指定镜头类型（特写、广角等）的正确实现。 5) *Expression adherence*：对情感状态和面部表情的准确表现。 6) *Camera motion adherence:*：相机运动（平移、倾斜、缩放等）的正确执行。 7) *Hallucination*：没有包含提示中未指定的内容。

运动质量评估视频中主体的时间动态。1) *Motion dynamism*：动作的多样性和表现力。2) *Fluidity and stability*：运动的流畅性，无抖动或不连续。3) *Physical plausibility*：符合自然物理和逼真的运动模式。

一致性 衡量视频帧之间的连贯性。 1) *Subject consistency*：视频中主要主体的外观稳定。 2) *Scene consistency*：背景、位置和环境元素的连贯性。对于图像到视频（I2V）模型，我们还评估： 3) *First-frame fidelity*：生成视频与提供的输入图像之间的一致性，包括颜色调色板的保持、主体身份的维护以及第一帧中场景元素的连续性。

视觉质量评估生成内容的空间保真度。1) *Visual clarity*：视觉元素的清晰度和定义。2) *Color accuracy*：色彩平衡适当，无过度饱和。3) *Structural integrity*：主体和背景无失真或损坏。

这个全面的评估框架使我们能够系统地比较不同模型的视频生成能力，并识别视频质量在各个方面的具体优缺点。

为了评估，一组由20名专业评估员组成的评审团使用1-5的评分标准对每个维度进行评估，评分标准如下表4所示：

表4：视频质量评估评分标准

Score	Label	Assessment Criteria
1	Fail	Complete failure to meet evaluation criteria
2	Marginal	Partial compliance with significant deficiencies
3	Adequate	Basic compliance with non-critical flaws
4	Proficient	Full satisfaction of all requirements
5	Excellent	Exceptional quality exceeding baseline requirements

最终结果总结在表5中。评估显示，我们的模型在遵循指令方面相较于基线方法取得了显著的进步，同时在运动表现上保持了竞争力，且未牺牲一致性。为了确保公平，所有模型均在默认设置和一致的分辨率下进行评估，且未应用后处理过滤。每个评判标准的详细评分指南可在附录A中找到。

Model Name	Average	Instruction Adherence	Consistency	Visual Quality	Motion Quality
Runway-Gen3 Alpha [74]	2.53	2.19	2.57	3.23	2.11
HunyuanVideo-13B [18]	2.82	2.64	2.81	3.20	2.61
Kling-1.6 STD Mode [2]	2.99	2.77	3.05	3.39	**2.76**
Hailuo-01 [3]	3.0	2.8	3.08	3.29	2.74
Wan2.1-14B [5]	3.12	2.91	3.31	**3.54**	2.71
SkyReels-V2	**3.14**	**3.15**	**3.35**	3.34	2.74

表5：SkyReels-Bench上的文本到视频（T2V）模型性能。在多个维度上采用1-5的评分进行评估，得分越高表示性能越好。

5.2 模型基准测试与排行榜

为了客观比较SkyReels-V2与其他领先的开源视频生成模型，我们使用公共基准VBench1.0 [6]进行全面评估。

我们的评估特别利用了基准的长版本提示。为了与基线模型进行公平比较，我们严格遵循它们推荐的推理设置。同时，我们的模型在生成过程中采用50个推理步骤和指导尺度为6，符合常见做法。

Model	Total Score	Quality Score	Semantic Score
CogVideoX1.5-5B [17]	80.3 %	80.9 %	77.9 %
OpenSora-2.0 [75]	81.5 %	82.1 %	78.2 %
HunyuanVideo-13B [18]	82.7 %	84.4 %	76.2 %
Wan2.1-14B [5]	83.7 %	84.2 %	**81.4 %**
SkyReels-V2	**83.9 %**	**84.7 %**	80.8 %

表6：Vbench1.0长提示版本的文本到视频（T2V）模型性能

VBench 结果（表6）显示 SkyReels-V2 优于所有基线模型，包括 HunyuanVideo-13B 和 Wan2.1-14B，获得最高的总分（83.9%）和质量分（84.7%）。在此次评估中，语义得分略低于 Wan2.1-14B，而在之前的人类评估中，我们优于 Wan2.1-14B，主要差距归因于 V-Bench 对镜头场景语义一致性的评估不足。

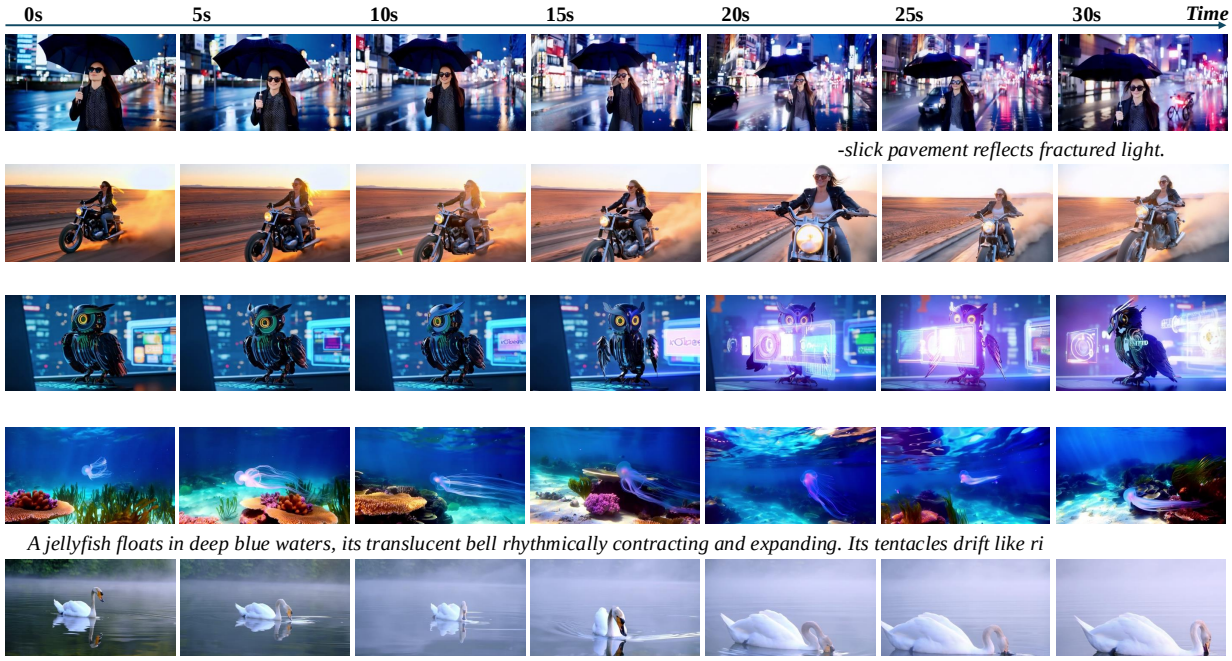

图8：使用我们的SkyReels-V2模型，通过单一提示生成超长视频的示例。

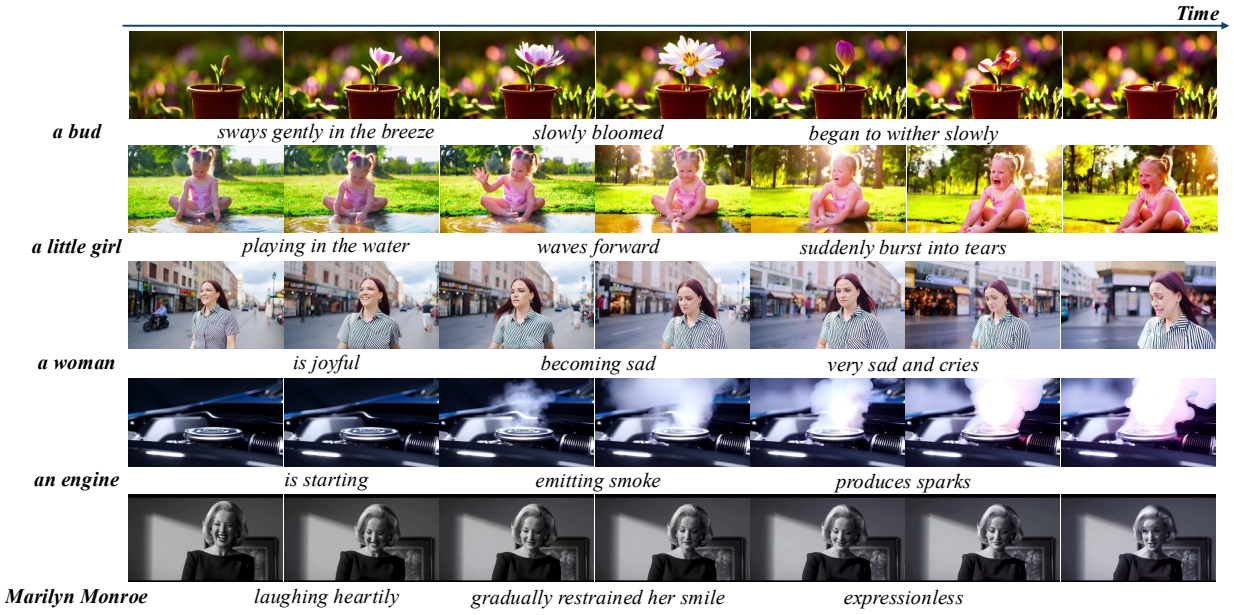

图9：使用我们的SkyReels-V2模型，通过顺序提示生成超长视频的示例。

6 应用

6.1 故事生成

我们训练的扩散强制变换器能够生成理论上无限延长超长视频。该模型采用滑动窗口方法生成长视频。它以最后的 f_{prev} 帧和文本提示作为条件，生成接下来的 f_{new} 帧，除了第一次迭代外，在第一次迭代中，它仅依赖文本提示。然而，延长视频长度可能会导致误差随着时间积累。为了解决这个问题，我们采用了

稳定技术，其中先前生成的帧被标记为轻微噪声水平。这种方法防止了误差积累，并进一步稳定了长时间的滚动过程。在图8中，我们展示了将长镜头视频扩展到超过30秒的示例，展示了在保持视觉连贯性的同时增强时间长度的能力。

我们的模型不仅支持单纯的时间延伸，还能够生成具有引人入胜故事情节的长镜头。通过利用一系列叙事文本提示，我们可以编排出一个连贯的视觉叙事，跨越多个动作，同时在整个视频中保持视觉一致性。这一能力确保了场景之间的平滑过渡，允许进行动态叙事而不影响视觉元素的完整性。模型无缝整合叙事文本提示的能力，使得创建既引人入胜又视觉和谐的扩展视频内容成为可能，非常适合需要复杂多动作序列的应用。在图9中，我们展示了用户可以通过连续的文本提示操控的示例，例如小女孩的动作、女人的表情以及引擎的状态等属性。

6.2 图像到视频的合成

在我们的框架下，有两种方法可以开发图像到视频（I2V）模型：1) *Fine-Tuning full-sequence Text-to-Video (T2V) diffusion Models* (SkyReels-V2-I2V)：遵循Wan 2.1的I2V实现，我们通过注入第一个参考帧作为图像条件，扩展T2V架构。输入图像被填充以匹配目标视频长度，然后通过VAE编码器处理以获得图像潜变量。这些潜变量与噪声潜变量以及4个二值掩码通道（1代表参考帧，0代表后续帧）连接，使模型能够利用参考帧进行后续生成。在适应过程中，为了保持原有的T2V能力，我们对新加入的卷积层以及交叉注意中的图像上下文到值的投影采用零初始化，而其他新组件（如图像上下文到键的投影）则使用随机初始化，以最小化微调期间性能的突变。此外，I2V的训练利用了通过第3.2节描述的字幕生成框架生成的I2V特定提示。值得注意的是，该方法仅用10,000次训练迭代，在384个GPU上就取得了具有竞争力的结果。2) *Text-to-Video (T2V) Diffusion Forcing model with first-frame Conditioning* (SkyReels-V2-DF)：我们的另一种方法直接利用扩散框架的条件机制，将第一帧作为干净的参考条件输入。这绕过了显式的模型再训练，同时通过潜空间约束保持时间一致性。我们使用SkyReels-Bench评估套件（表7）对SkyReels-V2与领先的开源和闭源图像到视频模型进行了比较。我们的结果显示，SkyReels-V2-I2V（3.29）和SkyReels-V2-DF（3.24）在开源模型中达到了最先进的性能，在所有质量维度上都显著优于HunyuanVideo-13B（2.84）[18]和Wan2.1-14B（2.85）[5]。平均得分为3.29的SkyReels-V2-I2V表现出与专有模型Kling-1.6[2]（3.4）和Runway-Gen4[12]（3.39）相当的性能。基于这些令人鼓舞的结果，我们公开发布了我们的SkyReels-V2-I2V模型，以推动社区在图像到视频合成方面的研究。

Model	Average	Instruction Adherence	Consistency	Visual Quality	Motion Quality
HunyuanVideo-13B [18]	2.84	2.97	2.95	2.87	2.56
Wan2.1-14B [5]	2.85	3.10	2.81	3.00	2.48
Hailuo-01 [3]	3.05	3.31	2.58	3.55	2.74
Kling-1.6 Pro Mode [2]	3.4	3.56	3.03	3.58	3.41
Runway-Gen4 [12]	3.39	3.75	3.2	3.4	3.37
SkyReels-V2-DF	3.24	3.64	3.21	3.18	2.93
SkyReels-V2-I2V	3.29	3.42	3.18	3.56	3.01

表7：SkyReels-Bench上的图像到视频（I2V）模型性能。在多个维度上采用1-5的评分进行评估，得分越高表示性能越好。

6.3 摄像机导演

虽然SkyCaptioner-V1在标注相机运动方面表现出稳健的性能，但我们观察到，尽管它实现了平衡的主体分布，但相机运动数据中固有的不平衡性为摄影参数的进一步优化带来了挑战。为了解决这一限制，我们专门从有监督微调（SFT）数据集中筛选了大约100万样本，确保基本相机运动及其常见组合的代表性均衡。在此基础上，我们利用384个GPU进行了超过3,000次迭代的微调实验，针对我们的图像到视频生成模型。这一专门的训练方案显著提升了摄影效果，尤其是在相机运动的流畅性和多样性方面。

6.4 元素到视频生成

当前的视频生成模型主要解决两个任务：文本到视频（T2V）和图像到视频（I2V）。T2V 利用像 T5 [76] 或 CLIP [30] 这样的文本编码器，从文本提示中生成视频，但由于扩散过程的随机性，常常存在不一致的问题。另一方面，I2V 从静态图像和可选文本中生成运动，但通常受到对初始帧过度依赖的限制。在我们之前的工作中，我们引入了 *elements-to-video* (E2V) 任务，并提出了 SkyReels-A2[77]，这是一种可控的视频生成框架，可以将任意视觉元素（如角色、物体和背景）组合成连贯的视频，受文本提示引导，同时确保每个元素与参考图像的高度一致。如图 10 所示，SkyReels-A2 生成高质量、时间上一致的视频，且可以编辑多个视觉元素的组合。此外，我们还提出 A2-Bench，这是一个用于全面评估 E2V 任务的新型基准，显示出与人类主观判断具有统计学显著相关性。

未来，我们计划推出一个统一的视频生成框架，支持包括音频和姿势在内的额外输入模态。在我们之前的工作 SkyReels-A1[78] 基于音频驱动和姿势驱动的人像动画的基础上，这个增强的框架将支持更丰富、更多样的输入形式。通过这样做，它旨在显著拓宽应用范围，包括但不限于短剧、音乐视频和虚拟电商内容创作。

图10：我们提出的SkyReels-A2模型的***elements-to-video***结果示例。给定带有多张图像和文本提示的参考，我们的方法可以生成逼真且自然构图的视频，同时保持特定身份的一致性。

7 结论

我们提出了SkyReels-V2模型，一种新颖的框架，能够生成无限长度的视频，同时保持对镜头场景提示的遵循、高质量的视频效果以及强大的运动质量。关键改进通过以下几个方面实现：1) *Prompt Adherence*：通过SkyCaptioner-V1模块增强，该模块利用来自通用多模态大语言模型（MLLMs）和专业镜头专家模型的知识蒸馏，确保与输入提示的精确对齐。2) *Video Quality*：通过利用多样化的数据源和多阶段训练流程显著提升，确保输出具有视觉连贯性和高保真度。3) *Motion Quality*：通过强化学习后训练进行优化，配合半自动化的数据生成流程，提升动态运动的一致性和流畅性。4) *Infinite-Length Generation*：由

扩散-强制框架，允许无缝扩展视频内容而没有明确的长度限制。尽管取得了这些进展，扩散-强制框架仍然受到在长时间生成中误差累积的影响，这目前限制了高质量视频输出的实际长度。未来的工作将集中于解决这一挑战，以进一步提升模型的可扩展性和可靠性。

8 位贡献者

我们衷心感谢所有贡献者的辛勤努力。以下列表按其主要贡献角色对参与者进行识别：

- 项目赞助人：周雅慧
- 项目负责人：陈桂宾† (guibin.chen@kunlun-inc.com)
- Contributors:贡献者：
  - *Infrastructure:* Hao Zhang†，雄伟，徐志恒，金宇哲 – *Data & Captioning:* 范明远†，陈正，马成成，赵鹏，许博远 – *Model Training:* 林迪轩†，杨江平†，林春泽†，朱俊成†，陈胜，王伟，庞诺，康康，梁玉鹏，宋玉冰，邱迪，李德邦，费正聪

† 表示贡献相等的作者。

参考文献

[1] OpenAI。视频生成模型作为世界模拟器，2024年。 [2] 快手。Kling，2024年。 [3] MiniMax。海洛，2024年。 [4] DeepMind。Veo 2，2024年。

[5] 万团队，安旺，宝乐爱，文彬，茅超杰，谢陈伟，陈迪，余飞武，赵海明，杨建孝，曾建远，王嘉瑜，张景峰，周景仁，王金凯，陈吉轩，朱凯，赵康，严科瑜，黄良华，冯孟阳，张宁怡，李攀登，吴平宇，楚锐航，冯睿丽，张世伟，孙思阳，方涛，王天行，桂天一，翁廷宇，沈彤，林伟，王伟，周文萌，王文腾，沈文婷，余文远，史显忠，黄晓明，徐新，寇彦，吕杨宇，李一菲，刘怡景，王一鸣，张颖雅，黄一通，李永，吴友，刘宇，潘玉林，郑云，洪云涛，史玉鹏，冯宇通，江泽仁，韩振，吴志凡，刘子宇。万：开放且先进的大规模视频生成模型。*arXiv preprint arXiv:2503.20314*，2025年。

[6] 黄子奇, 何一男, 于佳硕, 张凡, 司晨阳, 江玉明, 张远瀚, 吴天行, 金庆阳, 陈帆, 王耀辉, 陈新源, 王黎明, 林大华, 乔宇, 刘子威. VBench：视频生成模型的综合基准套件. 载于 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[7] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, 和 Shelly Sheynin。Videojam：用于增强视频模型中运动生成的联合外观-运动表示，2025年。 [8] 白帅，陈克勤，刘雪晶，王家林，葛文彬，宋思博，邓凯，王鹏，王世杰，唐俊，钟虎门，朱元志，杨明坤，李昭海，万建强，王鹏飞，丁伟，傅哲仁，许亦恒，叶嘉博，张曦，谢天宝，程泽森，张航，杨志博，许海洋，林俊阳。Qwen2.5-vl技术报告。*arXiv preprint arXiv:2502.13923*，2025年。 [9] OpenAI。Gpt-o1，2024年。

[10] DeepSeek-AI. Deepseek-r1：通过强化学习激励大语言模型的推理能力，2025年。
已知：{S*}，且{D*}是一个常数。假设：存在一个常数{a}，使得对于所有的{x}，有 {f(x) = a \cdot x + b}。 那么，{f(x)}

[11] 孙明玲，王维宁，李根，刘鹏伟，程义松，张一凡，冯玉堂，赵子琳，刘静。Ar-diffusion：基于自回归扩散的异步视频生成。*arXiv preprint arXiv:2503.07418*，2025年。
已知：{v*}，且{v*}是一个常数。假设：存在一个常数{a}，使得对于所有的{x}，有 {f(x) = a \cdot x + b}。 那么，{f(x)}

[12] Runway ML. Gen-4, 2024.
已知：{w*}，且{w*}是一个常数。 假设：存在一个常数 {a}，使得对于所有的 {x}，有 {f(x) = a \cdot x + b}。 那么，{f(x)}

[13] 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。*arXiv preprint arXiv:2209.14792*，2022年。 [14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, 和 Bo Dai. Animatediff：无需特定调优即可动画化个性化文本到图像扩散模型，2023年。 [15] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, 等. 稳定视频扩散：将潜在视频扩散模型扩展到大型数据集。*arXiv preprint arXiv:2311.15127*，2023年。 [16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, 和 David J Fleet. 视频扩散模型。*Advances in Neural Information Processing Systems*，35:8633–8646，2022年。 [17] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, 等. Cogvideox：带有专家变换器的文本到视频扩散模型。*arXiv preprint arXiv:2408.06072*，2024年。 [18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, 和 Caesar Zhong. 混元视频：大型视频生成模型的系统框架，2025年。 [19] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, 和 Daxin Jiang. 步视频-t2v技术报告：视频基础模型的实践、挑战与未来，2025年。 [20] SkyworkAI. Skyreels-v1，2025年。 [21] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, 等. Open-sora 2.0：在20万美元内训练商用级视频生成模型。*arXiv preprint arXiv:2503.09642*，2025年。 [22] Olaf Ronneberger, Philipp Fischer, 和 Thomas Brox. U-net：用于生物医学图像分割的卷积网络。在 *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*，第234–241页。Springer，2015年。 [23] William Peebles 和 Saining Xie. 具有变换器的可扩展扩散模型。*arXiv preprint arXiv:2212.09748*，2022年。 [24] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, 等. 扩展修正流变换器以实现高分辨率图像合成。在 *Forty-first international conference on machine learning*，2024年。 [25] Patrick Esser, Robin Rombach, 和 Bjorn Ommer. 高分辨率图像合成的变换器调优。在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第12873–12883页，2021年。 [26] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, 和 Yang You. Open-sora：让所有人都能高效制作视频的民主化平台。*arXiv preprint arXiv:2412.20404*，2024年。 [27] Zongjian Li, Bin Lin, Yang Ye, Liuhan Chen, Xinhua Cheng, Shenghai Yuan, 和 Li Yuan. Wf-vae：通过小波驱动的能量流增强视频VAE以实现潜在视频扩散模型。*arXiv preprint arXiv:2411.17459*，2024年。

[28] 赵思杰，张勇，存晓东，杨少书，牛慕瑶，李晓宇，胡文博，山颖 Shan。Cv-vae：一种兼容的视频VAE，用于潜在生成视频模型。*https://arxiv.org/abs/2405.20279*，2024年。 [29] NVIDIA 等。Cosmos世界基础模型平台，用于物理AI。*arXiv preprint arXiv:2501.03575*，2025年。 [30] 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的视觉模型。在 *International conference on machine learning*，第8748–8763页。PmLR，2021年。 [31] Colin Raffel，Noam Shazeer，Adam Roberts，Katherine Lee，Sharan Narang，Michael Matena，Yanqi Zhou，Wei Li，Peter J Liu。探索具有统一文本到文本变换器的迁移学习极限。*Journal of machine learning research*，2020年第21卷（第140期）：1–67页。 [32] Hyung Won Chung，Noah Constant，Xavier Garcia，Adam Roberts，Yi Tay，Sharan Narang，Orhan Firat。Unimax：为大规模多语言预训练提供更公平、更有效的语言采样。*arXiv preprint arXiv:2304.09151*，2023年。 [33] Jonathan Ho，Ajay Jain，Pieter Abbeel。去噪扩散概率模型。*Advances in neural information processing systems*，2020年第33期：6840–6851页。 [34] Robin Rombach，Andreas Blattmann，Dominik Lorenz，Patrick Esser，Björn Ommer。使用潜在扩散模型进行高分辨率图像合成。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*，2022年6月，第10684–10695页。 [35] Yaron Lipman，Ricky TQ Chen，Heli Ben-Hamu，Maximilian Nickel，Matt Le。用于生成建模的流匹配。*arXiv preprint arXiv:2210.02747*，2022年。

[36] OpenAI. Gpt-4o，2024年。

[37] DeepMind。Gemini 2.5 pro，2025。

[38] 袁丽萍，王嘉伟，孙浩淼，张宇辰，林远。Tarsier2：从详细视频描述到全面视频理解的推进大型视觉-语言模型，2025年。 [39] 欧阳龙，吴杰，江旭，阿尔维拉·库马尔，卡罗尔·L·韦恩赖特，帕梅拉·米什金，张冲，阿格瓦尔·桑迪尼，斯拉玛·卡塔琳娜，雷克斯·亚历克斯，约翰·舒尔曼，雅各布·希尔顿，弗雷泽·凯尔顿，卢克·米勒，西蒙斯·玛迪，阿斯克尔·阿曼达，韦林德·彼得，克里斯蒂亚诺·保罗，莱克·简，莱克·瑞安。用人类反馈训练语言模型以遵循指令，2022年。 [40] 彭雪彬，阿维拉尔·库马尔，张Grace，莱文Sergey。优势加权回归：简单且可扩展的离策略强化学习，2021年。 [41] 刘涛，匡华峰，林显明。在没有人类反馈的情况下对齐文本到图像扩散模型。在
*ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*中。 [42] 布拉姆·华莱士，邓梅花，拉斐尔·拉法伊洛夫，周林奇，卢安·艾伦，普鲁什沃克姆·森蒂尔，斯特凡诺·埃尔蒙，熊才明，乔蒂·沙菲克，奈克希尔·奈克。使用直接偏好优化进行扩散模型对齐。*IEEE*，2023年。 [43] 杨凯，陶建，吕佳菲，葛春江，陈家欣，李启迈，沈伟汉，朱晓龙，李秀，2023年。利用人类反馈在没有奖励模型的情况下微调扩散模型。*IEEE*，2023年。 [44] 梁展豪，袁玉辉，顾书阳，陈博涵，杭天凯，李济，郑亮。步级偏好优化：在每一步对齐偏好与去噪性能。2024年。 [45] 刘润涛，吴浩宇，张志强，陈伟，何颖青，皮仁杰，陈启锋。Videodpo：视频扩散生成的全偏好对齐，2024年。 [46] 刘杰，刘公业，梁家俊，袁子阳，刘晓坤，郑明武，吴谢乐，王秋林，秦文宇，夏孟涵，王新涛，王晓红，杨飞，万鹏飞，张迪，盖坤，杨宇九，欧阳万里。通过人类反馈改善视频生成。*arXiv preprint arXiv:2501.13918*，2025年。 [47] 柯俊杰，王启飞，王伊林，米兰法尔·佩伊曼，杨峰。Musiq：多尺度图像质量变换器，2021年。 [48] 王一彬，臧宇航，李浩，金成，王佳琪。多模态理解与生成的统一奖励模型，2025年。 [49] 童海波，王兆阳，陈昭润，季浩南，邱世伟，韩思维，耿克新，薛中凯，周怡阳，夏鹏，丁明宇，拉斐尔·拉法伊洛夫，克莱莎·芬恩，姚华秀。Mj-video：在视频生成中细粒度的基准测试与偏好奖励，2025年。 [50] 陈博远，蒙索·马尔蒂，杜伊伦，马克斯·西姆乔维茨，泰德雷克·拉斯，席茨曼·文森特。扩散强制：下一词预测与全序列扩散的结合，2024年。

[51] 孙明震、王维宁、李根、刘佳伟、孙嘉辉、冯万全、老珊珊、周思宇、何谦、刘靖。Ar-diffusion：基于自回归扩散的异步视频生成，2025年。 [52] 宋基焕、陈博远、Max Simchowitz、杜伊伦、Russ Tedrake、Vincent Sitzmann。以历史为导向的视频扩散，2025年。 [53] Jonathan Ho 和 Tim Salimans。无分类器扩散引导，2022年。 [54] 尹天伟、张强、张锐、William T. Freeman、Fredo Durand、Eli Shechtman 和 Xun Huang。从缓慢的双向到快速的自回归视频扩散模型，2025年。 [55] 尹天伟、Michaël Gharbi、Richard Zhang、Eli Shechtman、Fredo Durand、William T. Freeman 和 Taesung Park。一阶扩散与分布匹配蒸馏，2024年。 [56] 尹天伟、Michaël Gharbi、Taesung Park、Richard Zhang、Eli Shechtman、Fredo Durand 和 William T. Freeman。改进的分布匹配蒸馏用于快速图像合成，2024年。 [57] 郭玉伟、杨策远、杨子彦、马志贝、林志杰、杨振恒、林大华、姜璐。长上下文调优用于视频生成，2025年。 [58] 王秋恒、史玉凯、欧佳荣、陈锐、林科、王佳豪、姜博远、杨浩天、郑明武、陶新、杨飞、万鹏飞、张迪。Koala-36m：一个大规模视频数据集，提升细粒度条件与视频内容的一致性，2024年。 [59] 王振志、李一轩、曾艳红、方有庆、郭玉伟、刘文然、谭靖、陈凯、薛天凡、戴博、林大华。Humanvid：揭秘可控摄像机的人体图像动画训练数据。在 *NeurIPS*，2024年。 [60] Tomá Souček 和 Jakub Lokoč。Transnet v2：一种用于快速镜头转换检测的有效深度网络架构。*arXiv preprint arXiv:2008.04838*，2020年。 [61] 动词。aesthetic-predictor-v2-5。https://github.com/discus0434/aesthetic-predictor-v2-5，2024年。访问时间：2024.11.12。 [62] Ed Pizzi、Sreya Dutta Roy、Sugosh Nagavara Ravindra、Priya Goyal 和 Matthijs Douze。用于图像复制检测的自监督描述符。*Proc. CVPR*，2022年。 [63] Wu Haoning、Chen Chaofeng、Hou Jingwen、Liao Liang、Wang Annan、Sun Wenxiu、Yan Qiong、Lin Weisi。Fast-vqa：一种高效的端到端视频质量评估方法，采用片段采样。在 *Proceedings of European Conference of Computer Vision (ECCV)*，2022年。 [64] Wu Haoning、Zhang Erli、Liao Liang、Chen Chaofeng、Hou Jingwen、Wang Annan、Sun Wenxiu、Yan Qiong、Lin Weisi。从美学和技术角度探索用户生成内容的视频质量评估。在 *International Conference on Computer Vision (ICCV)*，2023年。 [65] Wu Haoning。开源深度端到端视频质量评估工具箱，2022年。 [66] Lorenzo Agnolucci、Leonardo Galteri、Marco Bertini 和 Alberto Del Bimbo。Arniqa：学习畸变流形以进行图像质量评估。在 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*，第189–198页，2024年。 [67] Baek Youngmin、Lee Bado、Han Dongyoon、Yun Sangdoo 和 Lee Hwalsuk。字符区域感知的文本检测。在 *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*，第9365–9374页，2019年。 [68] Yao Yuan、Yu Tianyu、Zhang Ao、Wang Chongyi、Cui Junbo、Zhu Hongji、Cai Tianchi、Li Haoyu、Zhao Weilin、He Zhihui 等。Minicpm-v：一款在手机上实现的GPT-4v级别的多模长语言模型。*arXiv preprint arXiv:2408.01800*，2024年。 [69] Chen Yin、Li Jia、Shiguang Shan、Wang Meng 和 Hong Richang。从静态到动态：将地标感知图像模型应用于视频中的面部表情识别。*IEEE Transactions on Affective Computing*，第1–15页，2024年。 [70] Liu Feng、Zhang Shiwei、Wang Xiaofeng、Wei Yujie、Qiu Haonan、Zhao Yuzhong、Zhang Yingya、Ye Qixiang 和 Wan Fang。时间步嵌入告诉你：是时候为视频扩散模型缓存了。*arXiv preprint arXiv:2411.19108*，2024年。 [71] Pejaver V Rao 和 Lawrence L Kupper。配对比较实验中的 ties：Bradley-Terry 模型的推广。*Journal of the American Statistical Association*，第62卷(317期)：194–204，1967年。 [72] Sam Ade Jacobs、Tanaka Masahiro、Zhang Chengming、Zhang Minjia、Song Shuaiwen Leon、Rajbhandari Samyam 和 He Yuxiong。Deepspeed Ulysses：实现极长序列变换器模型训练的系统优化，2023年。

[73] 张金涛，黄浩峰，张鹏乐，魏佳，朱俊，陈建飞。Sageattention2：高效注意力机制，结合彻底的异常值平滑和每线程的int4量化，2025年。 [74] RunwayML。Gen-3 alpha，2025年。 [75] 彭翔宇，郑臻伟，沈晨辉，汤姆·杨，郭欣颖，王彬洛，徐航，刘宏新，江明燕，李文俊，王玉辉，叶安邦，任刚，马千然，梁婉莹，连翔，吴希文，钟玉婷，李壮岩，龚超宇，雷国军，程雷军，张丽敏，李明浩，张瑞杰，胡思兰，黄世杰，王晓康，赵元恒，王玉琪，魏子昂，游杨。Open-sora 2.0：用20万美元训练一个商业级视频生成模型。*arXiv preprint arXiv:2503.09642*，2025年。 [76] Colin Raffel，Noam Shazeer，Adam Roberts，Katherine Lee，Sharan Narang，Michael Matena，周彦奇，李伟，Peter J. Liu。探索统一文本到文本变换器在迁移学习中的极限，2019年。 [77] 费正聪，李德邦，邱迪，王家华，窦一坤，道睿，徐景涛，范明远，陈桂斌，李阳，等。Skyreels-a2：在视频扩散变换器中实现任意内容的创作。*arXiv preprint arXiv:2504.02436*，2025年。 [78] 邱迪，费正聪，道睿，白佳林，于昌谦，范明远，陈桂斌，文翔。Skyreels-a1：在视频扩散变换器中实现富有表现力的肖像动画。*arXiv preprint arXiv:2502.10841*，2025年。

SkyReels-Bench 评分指南

SkyReels-Bench 详细评分指南

Evaluation Scoring Guidelines (1-5 scale)	
Instruction Adherence	**Score 1:** Complete failure to follow instructions, with severe deviations in video theme and key elements from the prompt **Score 2:** Partial adherence with significant deviations; key elements or themes are somewhat present but incomplete and inaccurate **Score 3:** Basic adherence with complete representation of prompt content; main content and key elements align with instructions but may contain minor inaccuracies or omissions in details **Score 4:** High adherence with accurate representation of all key content and elements from the instructions without any deviations **Score 5:** Perfect adherence with extensions; video completely follows all aspects of instructions (theme, elements, style) and enhances them with appropriate extensions for superior results
Consistency	**Score 1:** Complete inconsistency with severe discrepancies in subjects, style, and scenes; extreme frame-to-frame differences **Score 2:** Multiple significant inconsistencies that substantially impact overall video coherence **Score 3:** Partial inconsistencies limited to minor details (extremely localized areas); slight abnormalities in specific elements **Score 4:** Complete consistency with stable subjects, style, and scenes; minor imperfections only in non-primary elements that don't affect the overall viewing experience **Score 5:** Perfect consistency with extensions; all aspects of the video (subjects, style, scenes) are completely aligned with instructions, creating exceptional visual harmony
Visual Quality	**Score 1:** Severe quality issues with significant blurriness, pixelation, or other visual artifacts making the video nearly unwatchable **Score 2:** Poor quality with noticeable blurriness and obvious issues; content is barely recognizable and significantly impacts viewing experience **Score 3:** Average quality with minor visual flaws such as slight blurriness or minimal noise; does not affect comprehension of main content **Score 4:** Good quality reaching normal viewing standards; clear imagery without notable flaws, presenting well across various devices **Score 5:** Perfect quality at professional standards; impeccable resolution, color, contrast, and detail presentation suitable for high-quality exhibitions and distribution
Motion Quality	**Score 1:** Severely flawed movements with extreme jerkiness and discontinuity, resulting in extremely poor viewing experience **Score 2:** Poor motion with obvious stuttering and disconnected transitions; movements appear unnatural with abrupt shifts between scenes **Score 3:** Adequate motion with occasional stuttering or discontinuities; generally follows movement rhythm without significantly affecting content comprehension **Score 4:** Good motion quality with smooth, natural movement throughout; no obvious stuttering and overall fluid viewing experience **Score 5:** Exceptional motion quality with perfect fluidity and naturalness; completely free of stuttering or discontinuities, achieving human-like movement without any AI artifacts

表A1：综合评估标准。本评分指南由人工评估者用以评估视频生成模型的四个维度。每个维度的评分范围为1到5，其中1表示完全失败，5代表超出基本要求的卓越质量。

B 数据处理流程

在数据处理中过滤最大内部矩形的伪代码如下所示。

Algorithm A1 Pseudo-code of finding largest interior rectangle

Require: A 0-1 matrix $M \in \mathbb{R}^{m \times n}$
Ensure: The coordinate of largest interior rectangle $rect\_coords$

Initialize: $max\_area \leftarrow 0$, $rect\_coords \leftarrow (0, 0, 0, 0)$, $heights \leftarrow [0] \times n$
for $i \leftarrow 0$ **to** $m - 1$ **do**
    for $j \leftarrow 0$ **to** $n - 1$ **do**
        if matrix$[i][j] = 1$ **then**
            $heights[j] \leftarrow heights[j] + 1$
        else
            $heights[j] \leftarrow 0$
        end if
    end for
    $stack \leftarrow \emptyset$, $left \leftarrow [-1] \times n$
    for $j \leftarrow 0$ **to** $n - 1$ **do**
        while $stack \neq \emptyset$ and $heights[j] \leq heights[\text{top}(stack)]$ **do**
            pop$(stack)$
        end while
        if $stack \neq \emptyset$ **then**
            $left[j] \leftarrow \text{top}(stack)$
        else
            $left[j] \leftarrow -1$
        end if
        push$(stack, j)$
    end for
    $stack \leftarrow \emptyset$, $right \leftarrow [n] \times n$
    for $j \leftarrow n - 1$ **downto** 0 **do**
        while $stack \neq \emptyset$ and $heights[j] \leq heights[\text{top}(stack)]$ **do**
            pop$(stack)$
        end while
        if $stack \neq \emptyset$ **then**
            $right[j] \leftarrow \text{top}(stack)$
        else
            $right[j] \leftarrow n$
        end if
        push$(stack, j)$
    end for
    for $j \leftarrow 0$ **to** $n - 1$ **do**
        $height \leftarrow heights[j]$
        $width \leftarrow right[j] - left[j] - 1$
        $area \leftarrow height \times width$
        if $area > max\_area$ **then**
            $max\_area \leftarrow area$
            $top \leftarrow i - height + 1$
            $bottom \leftarrow i$
            $left\_col \leftarrow left[j] + 1$
            $right\_col \leftarrow right[j] - 1$
            $rect\_coords \leftarrow (top, left\_col, bottom, right\_col)$
        end if
    end for
end for
return $rect\_coords$

C 系统提示 SkyCaptioner-V1

用于 SkyCaptioner-V1 生成结构化字幕的系统提示如表 3 所示。在评估过程中，我们也对基线模型使用了相同的系统提示。

生成视频结构化字幕的系统提示

I need you to generate a structured and detailed caption for the provided video. The structured output and the requirements for each field are as shown in the following JSON content:

```
{
  “subjects": [
    {
      “TYPES": {
        “type": “Main category (e.g., Human)",
        “sub_type": “Sub-category (e.g., Man)"
      },
      “appearance": “Main subject appearance description",
      “action": “Main subject action",
      “expression": “Main subject expression (Only for human/animal categories, empty otherwise)",
      “position": “Subject position in the video (Can be relative position to other objects or spatial description)",
      “is_main_subject": true
    },
    {
      “TYPES": {
        “type": “Main category (e.g., Vehicles)",
        “sub_type": “Sub-category (e.g., Ship)"
      },
      “appearance": “Nonmain subject appearance description",
      “action": “Nonmain subject action",
      “expression": “Nonmain subject expression (Only for human/animal categories, empty otherwise)",
      “position": “Position of nonmain subject 1",
      “is_main_subject": false
    }
    ],
    “shot_type": “Shot type(Options: long_shot/full_shot/medium_shot/close_up/extreme_close_up/other)",
    “shot_angle": “Camera angle(Options: eye_level/high_angle/low_angle/other)",
    “shot_position": “Camera position(Options: front_view/back_view/side_view/over_the_shoulder/ overhead_view/point_of_view/aerial_view/overlooking_view/other)",
    “camera_motion": “Camera movement description",
    “environment": “Video background/environment description",
    “lighting": “Lighting information in the video"
}
```

在通过 SkyCaptioner-V1 模型生成结构化字幕后，我们设计了一个字幕融合流程，用于获得文本到视频（T2V）和图像到视频（I2V）模型训练的最终字幕。我们的流程利用 Qwen2.5-32B-Instruct 模型智能地结合结构化字幕字段，根据应用需求生成密集或稀疏的最终字幕。

系统提示用于T2V提示融合

你是视频字幕方面的专家。你会得到一个结构化的视频字幕，需要将其改写得更自然、更流畅。## *Structured Input* {structured_caption} ## *Notes* - 根据动作字段信息，将其名称字段改为动作中的主语代词。 - 如果有空字段，则忽略它，不在输出中提及。

- 不要对原始字段进行任何语义上的更改。请确保遵循原意。## *Output Principles and Orders* - 首先，声明镜头类型、镜头角度、镜头位置（如果这些字段不为空）。 - 其次，删除动作字段中与时间动作无关的信息，例如背景或环境信息。 - 第三，描述每个主体的纯动作、外观、表情、位置（如果这些字段存在）。 - 最后，如果字段不为空，声明环境、照明和摄像机运动。## *Output* 请直接输出最终组合的字幕，不要包含任何额外信息。

系统提示用于 I2V 提示融合

你是一位视频字幕专家。你将获得一个结构化的视频字幕，需要将其润色成更自然流畅的英语。## *Structured Input* {structured_caption} ## *Notes* - 如果某个字段为空，请忽略它，不要在输出中提及。 - 不要对原始字段进行任何语义上的更改。请确保遵循原意。 - 如果动作字段不为空，去除与时间动作无关的无关信息（如穿着、背景和环境信息），使动作字段纯粹。## *Output Principles and Orders* - 首先，去除与时间动作无关的静态信息，如背景或环境信息。 - 其次，用纯粹的动作和表情描述每个主体（如果这些字段存在）。 - 最后，如果相机运动字段不为空，添加到最终润色的字幕中。## *Output* 请直接输出最终润色后的字幕，不要包含任何额外信息。

D 视频运动质量评分标准用于人工标注

Video Motion Quality Assessment Criteria	
Insufficient Motion Amplitude (1 point/instance)	Subject's motion range significantly less than reasonable real-world range
	Rigid limb movements (insufficient arm swing)
	Facial expressions lack detail (insufficient smile uplift)
	Object trajectories too gentle (vehicles moving too slowly)
Excessive Motion Amplitude (2 points/instance)	Motion speed/range exceeds realistic physical laws or human behavior
	Abnormal object speeds (car instantly accelerating to extreme speed)
	Exaggerated limb movements (arm bending beyond physiological limits)
Subject Distortion (3 points/instance)	Unnatural distortion of subject shape/structure during movement
	Body parts stretching during walking (waist becoming thin)
	Legs separating from torso while running
	Surface wrinkles or tears when objects rotate
	Facial muscles causing misaligned features
Local Detail Distortion (1 point/instance)	Local areas appear blurry, incorrect, or missing
	Hair breaks or color blocks disappear during movement
	Blurry object surface textures (unclear fabric wrinkles)
	Unrecognizable text or identifiers in background
	Vehicle wheels remaining stationary while driving
Basic Physics Violations (3 points/instance)	Content violates basic laws of physics
	No physical feedback after collisions (ball passing through wall)
	Fluid movements inconsistent with fluid dynamics
Interaction Violations (2 points/instance)	Subject interactions with environment/other subjects defy reality
	Unreasonable object collisions/penetrations (person passing through closed door)
	Objects colliding without reaction (balls not bouncing after collision)
	Hand-object interaction misalignment (hand not touching cup handle when holding)
Unnatural/Monotonous Movement (1 point/instance)	Movements lack fluidity or diversity
	Single actions repeated back and forth
	Movements not following conventional paths
	Objects moving in single trajectories (straight lines only)
	Abrupt movement transitions (sudden acceleration/deceleration)

表A2：视频运动质量评估标准。该评分系统用于识别和量化生成视频中的运动相关问题，根据运动问题的严重程度和类型分配分值。