arXiv:2601.19236v1 [cs.CV] 27 Jan 2026

VC-BENCH：开创视频连接基准，构建数据集与评估指标

尹志宇[1,3] 刘志鹏[1,3] Kehai Chen[1,2] 刘乐茂 刘锦[3]
李弘东[3] Yang Xiang[1] Min Zhang[1,2]

[1]哈尔滨工业大学（深圳）计算机科学与技术学院
[2]鹏城实验室
[3]中南大学计算机科学与工程学院

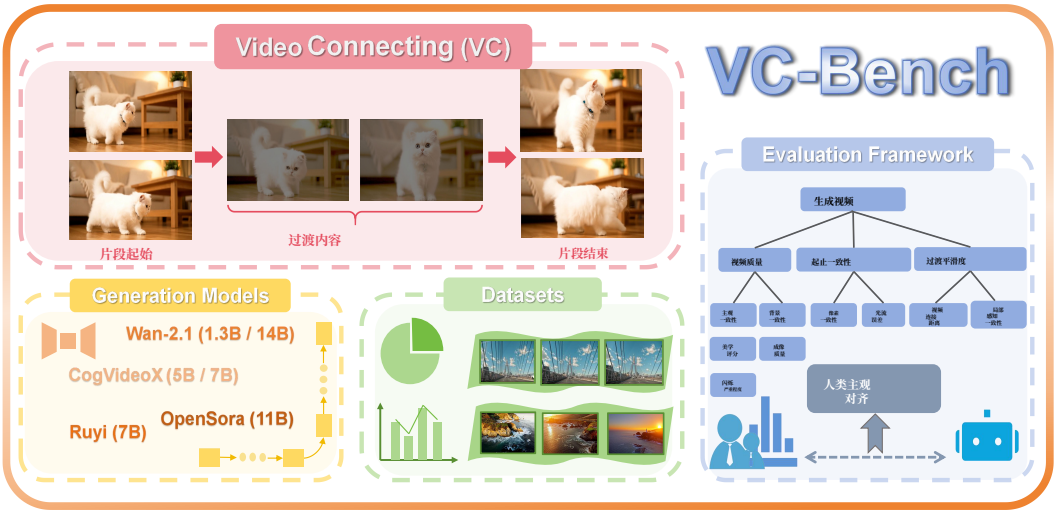


图 1：**VC-Bench 概览。** 我们推出了**VC-Bench**，这是一个专为新兴的**视频连接**任务量身打造的基准测试。我们对该任务进行了精确定义，并调整了多个开源视频生成模型以支持该任务。为建立系统、全面的评估框架，我们整理了一个涵盖 15 个类别的高质量数据集。我们开发了 9 个自动化指标，从三个关键维度评估模型表现：**视频质量**、**起止一致性**和**过渡平滑度**。人工主观验证进一步证实，该框架符合人类偏好。

摘要

当前的视频生成主要依赖文本或图像条件，但视频编辑和视频博客创作等实际应用往往需要无缝衔接彼此独立的片段。在本研究中，我们提出了**视频连接**这一创新任务，旨在生成给定起始片段与结束片段之间流畅的中间视频内容。然而，缺乏标准化评估基准阻碍了这项任务的发展。为弥补这一空白，我们提出了**VC-Bench**，这是一个专为视频连接设计的新型基准测试。它包含从公共平台收集的 1,579 条高质量视频，涵盖 15 个主要类别和 72 个子类别，以确保多样性和结构性。VC-Bench 聚焦于三个核心方面：**视频质量分数** VQS、**首尾一致性分数** $SECS$和**过渡平滑度分数** TSS。三者共同构成了一个全面的框架，超越了传统仅关注质量的指标。我们使用 VC-Bench 对多个最先进的视频生成模型进行了评估。实验结果揭示，这些模型在保持首尾一致性和过渡平滑度方面存在明显局限，导致整体连贯性和流畅度较低。我们希望 VC-Bench 能成为开创性的基准测试，为未来的发展提供启发和指引

视频连接领域的研究。评估指标和数据集已公开发布，获取地址：
https://anonymous.4open.science/r/VC-Bench-1B67/。

1 引言

生成式 AI 的最新进展改变了视频创作方式，Sora（OpenAI，2024年）、Runway-Gen（Runway，2024年）和 Pika（PikaLabs，2024年）等模型在文本到视频和图像到视频生成方面展现出前所未有的进步（Li 等人，2024年；Sun 等人，2024年b）。这些系统大幅提升了创作效率，使电影制作人和创作者能够在数小时内制作出概念视频。然而，现有工作主要集中于生成独立片段，对连续性感知生成的探索仍然不足。

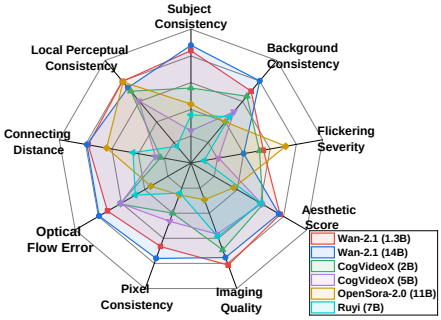


图2：**VC-Bench 上的视频得分**。基于 VC-Bench，我们评估了 6 个开源模型在视频连接任务上的性能。实验结果见表2

我们提出视频连接（VC）任务：给定两个片段（一个起始片段和一个结束片段），目标是合成在时空上连贯且与原始片段保持一致的过渡内容。不同于通用视频生成，VC 要求在不连续的片段之间保持严格的语义和视觉对齐。这一定义反映了实际需求——例如衔接视频博客中的离散镜头、填补设备故障造成的监控空档，或在电影中生成叙事性过渡。视频转换任务的更多应用见附录 C.2。通过将视频连接界定为一个独立问题，我们突显了它在弥合孤立生成与现实世界连续性需求之间差距方面的潜力。

重要的是，视频连接并非随意连接任意两个毫不相关的片段。当片段之间存在逻辑、语义或叙事关联，且本应具有连续性却缺失时，这项任务才有意义。为推动系统性研究，我们构建了一个涵盖多种场景的基准测试，并提出沿三个维度展开的多维评估指标：整体视频质量、起止一致性和过渡平滑度。我们希望这项工作能为推进具备连续性感知能力的视频生成奠定基础。

主要贡献有三点：

- 我们首次系统性地提出了**视频连接（VC）**这一新颖任务，拓展了视频生成的边界，并为这一新兴研究方向奠定了理论基础。
- 我们构建了一个全面的视频连接基准测试，称为**VC-Bench**，其中包括：涵盖多个类别和多种场景的测试视频数据集，以及三维视频评估范式。
- 基于这一基准测试，我们对多个视频生成模型进行了全面评估，深入分析了它们的性能，确定了当前的技术局限，并为未来的发展提出了建设性的研究建议。

2 相关工作

2.1 视频生成

文本到视频生成。 文本生成视频（T2V）是一种利用人工智能模型将自然语言描述转化为视频内容的技术。通过理解文本中的语义、场景和动作，它能够生成高质量的视频序列。近年来，扩散模型和Transformer架构的进步显著提升了生成视频的保真度和一致性。Chen等人使用Stable Diffusion生成图像。Sora（OpenAI，2024年）能够生成高清视频，在复杂场景和多角色互动方面表现出色，而

可灵 AI（快手，2024年）提供运动控制和风格预设，适用于创意生成。Vidu（ShengShu-AI，2024年）和 Haiper（DeepMind，2024年）专注于叙事连贯性和故事生成能力。Wan2.1（Wang等人，2025年）和混元视频（Kong等人，2024年）进一步推动了高效且具有电影质感的视频制作，以满足多样化的创意需求。

图像到视频生成。 图像到视频（I2V）生成能够将静态图像转化为动态视频。它适用于动画制作和叙事，通常会结合文本提示以增强灵活性。Runway Gen-3 Alpha（Runway，2024年）和 Luma AI（LumaLabs，2024年）支持高保真度的图像到视频生成，而即梦 AI（Jianying，2024年）则专注于中国场景的视觉效果，能够生成短视频。Ruyi（Team，2024年）和 CogVideoX（Yang 等人，2024年）支持基于首帧和末帧的视频生成。LTX-Video（ HaCohen 等人，2024年）能够实时生成流畅的 10 秒短视频。Open-Sora 2.0（Peng 等人，2025年）整合图像和文本输入以生成高质量的动态内容，显著拓展了视觉叙事的可能性。

2.2 视频生成评估

近年来，视频生成技术迅速发展，催生了各种指标，用于从不同维度评估生成视频。相关评估指标可归纳为以下几个维度：

运动动态评估： 在运动动态方面，多个基准测试（Ling 等人，2025年）利用长时视频数据集分析运动自然度。MiraBench（Ju 等人，2024年）构建了一个包含高质量长时视频和细粒度字幕的数据集，以更好地评估视频生成中的时间一致性和运动强度。（Liao 等人，2024年）提出了 DEVIL 方法，通过定义细粒度的时间动态评分以及多个动态级别的文本提示，评估文本到视频模型的生成能力。

文本对齐评估： 生成模型需要严格遵循文本提示中的语义约束。StoryBench（ Bugliarello 等人，2023）通过设置三个层级递进的任务，系统评估模型从基础指令理解到复杂故事构建的迁移能力。StoryEval（Wang 等人，2024b）专门用于评估模型处理事件级故事呈现的能力。该方法使用视觉语言模型对生成视频进行事件级分解和验证，并结合投票机制提高可靠性。Zhu 等人精心构建了一个专为大型视觉语言模型设计的基准测试，旨在提升其视觉理解和推理能力。

时间连贯性评估：时间连贯性是视频生成的核心挑战。ChronoMagic-Bench（Yuan 等人，2024年）引入 MTScore 和 CHScore，评估 T2V 模型生成高动态延时视频时渐进过程的合理性。T2VBench（Ji 等人，2024年）利用时间动态特征，从运动合理性和事件时序逻辑方面量化生成视频的表现。TC-Bench（Feng 等人，2024年）定义场景的初始状态和目标状态，并通过 TCR指标衡量转变完整性。

综合性能评估：视频生成模型的综合评估体系正朝着多维度、细粒度的方向发展（Liu 等人，2024年）。AIGCBench（Fan 等人，2023）通过 11 项指标，为图生视频生成任务建立了首个可扩展的评估体系。VBench（Huang 等人，2024年）将视频评估从表层真实性扩展到内在真实性。T2V-CompBench（Sun 等人，2024a）整合了 LLaVA-1.5 和目标跟踪技术，在复杂提示词评估方面展现出显著优势。

3 视频连接

3.1 任务定义

假设起始视频片段记为 $VS = \{I_t\}_1^{N_S}$，结束视频片段记为$V_E = \{I_t\}_1^{N_E}$，其中 N_* 表示视频的帧数， $I_t \in \mathbb{R}^{H\times W\times C}$ 表示每一帧。 V_S 和 V_E 具有相同的空间分辨率和帧率。给定一对视频片段 V_S和 V_E，以及可选的文本提示词 T，视频连接任务旨在生成一个满足以下条件的视频V：

基准名称	Year	开放领域	文本-视频	图像-视频	视频连接	指标
LFDM 评测（Ni 等人，2023）	2023	✘	✔	✘	✘	3
StoryBench（Bugliarello 等人，2023）	2023	✔	✔	✘	✘	10
CATER-GEN（Hu 等人，2024年）	2023	✘	✔	✔	✘	7
SVD 评测（Blattmann 等人，2023）	2023	✔	✔	✔	✘	5
EvalCrafter（Liu 等人，2024年）	2023	✔	✔	✔	✘	17
VideoPhy（Bansal 等人，2024年）	2024	✔	✔	✘	✘	2
PhyGenBench（Meng 等人，2024年）	2024	✔	✔	✘	✘	5
MiraBench（Ju 等人，2024年）	2024	✔	✔	✘	✘	17
DEVIL（Liao 等人，2024年）	2024	✔	✔	✘	✘	10
StoryEval（Wang 等人，2024年b）	2024	✔	✔	✘	✘	7
ChronoMagic-Bench（Yuan 等人，2024年）	2024	✔	✔	✘	✘	6
T2VBench（Ji 等，2024年）	2024	✔	✔	✘	✘	16
TC-Bench（Feng 等，2024年）	2024	✔	✔	✔	✘	7
AIGCBench（Fan 等，2023）	2024	✔	✔	✔	✘	11
Video-Bench（Han 等，2025年）	2024	✔	✔	✔	✘	10
Step-Video-T2V-Eval（Huang 等，2025年）	2024	✔	✔	✘	✘	4
VMBench（Lin 等，2024年）	2025	✔	✔	✘	✘	5
VBench 1.0 / 2.0（Huang 等人，2024年）	2025	✔	✔	✔	✘	21
T2V-CompBench（Sun 等人，2024a）	2025	✔	✔	✘	✘	7
VC-Bench（我们的方法）	**2025**	✔	✔	✔	✔	**9**

表 1：**视频生成基准测试概述**。我们总结了过去三年与视频生成相关的基准测试工作，并提取了相关信息。

- V 的起始和结束片段分别与 V_S 和 V_E 保持一致。
- V 具备视频的各项属性，例如一致的空间分辨率和帧率、平滑的运动等。

第一个条件确保生成视频忠实于给定的起始和结束片段，分别记为 V_S 和 V_E。这一约束在实际应用中至关重要，因为这些片段通常由内容创作者预先指定。引入第二个条件是为了在整个生成过程中保持输出视频的高视觉质量。

与 FLF2V 的关系 首尾帧生成视频（FLF2V）（Wang 等人，2025年；Zhang 和 Agrawala，2025年）的生成方式与我们的视频连接（VC）任务相似，因为两者都通过起始和结束状态控制视频内容。然而，FLF2V 属于图像到视频任务，而 VC 属于视频到视频任务，因此面临的挑战各不相同。这两项任务之间的差异详见附录 B.1。

（1）信息复杂度： 在 FLF2V 任务中，视频生成主要由首帧和末帧所包含的信息引导，这些帧提供静态视觉线索。相比之下，VC 任务依赖起始和结束视频片段中的时空信息，其中包含动态运动、时间上下文和更丰富的内容。尽管这提供了更多信息，但也带来了如何有效提取和利用这些复杂时空数据的挑战，因为处理视频片段比处理静态帧的难度大得多。

(2) 内容差异： 在 FLF2V 任务中，两个帧通常属于同一个视频序列，在内容、风格和场景方面具有高度一致性。模型只需专注于生成中间帧，以确保视觉效果和运动过渡平滑。相比之下，在我们的视频连接任务中，起始视频和结束视频可能来自不同的视频序列，内容上可能存在显著差异，例如光照条件、配色方案等。

(3) 时间一致性： 两项任务都要求确保时间一致性。然而，视频连接的难点在于，起始和结束片段可能具有不同的运动模式和节奏。例如，起始视频可能是慢动作，而结束视频则节奏较快。模型必须生成一段过渡视频，使其平滑地适应结束视频的节奏，这比简单的帧插值更复杂。

3.2 迁移方法

为解决现有模型缺乏对视频连接任务支持的问题，我们调整了扩散 Transformer（DiT）架构，使其能够在起始片段和结束片段的条件下生成视频，从而增强时间一致性、语义连贯性和内容可控性。具体方法如下：

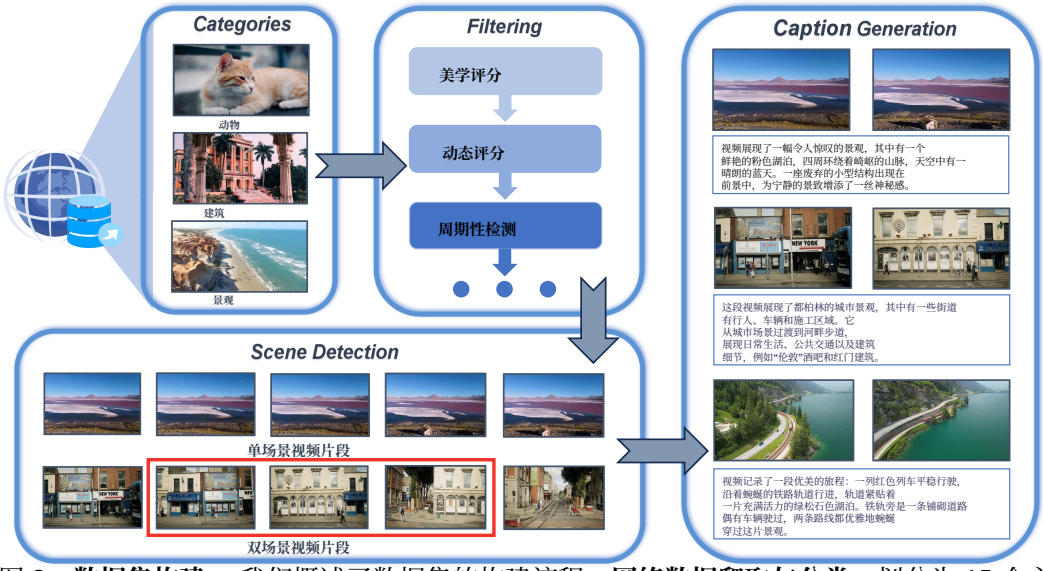


图 3：**数据集构建**。我们概述了数据集的构建流程：**网络数据爬取与分类**：划分为 15 个主要类别和 72 个子类别。**数据筛选**：通过美学评分、动态评分和周期性检测，筛选出高质量测试数据。**场景检测**：使用 PySceneDetect 进行场景分割。**字幕生成**：使用 Qwen2-VL 生成高质量字幕。

（1）映射到潜在空间：起始片段和结束片段分别映射到潜在空间的初始位置和最终位置。中间部分填充随机噪声，并与经过条件化的片段一同输入 DiT 模型处理，以引导去噪过程对时间依赖关系进行建模，并确保时空一致性。

(2) 用于条件控制的球面线性插值： 与 TVG（Zhang等人，2024年）类似，我们采用球面线性插值（SLERP）混合起始和结束片段的特征，以生成平滑的过渡部分，从而改善跨场景视频衔接，实现自然过渡。

说明 用于视频生成的 DiT 架构在实现硬约束生成方面存在局限：**(1) VAE 编码器尺度效应：** VAE 编码器中的固定比例压缩以及舍入或零填充等技术，会导致原始内容不可逆地丢失。**(2) 基于图块的序列化：** 将视频划分为带有位置编码的固定大小时空图块，会改变原始视频的特性，尤其是边界图块。**(3) 潜空间去噪：**将视频压缩到低维潜在空间以实现高效去噪，再对其进行重建，会引入噪声并降低保真度。

4 VC-BENCH

本节将详细介绍 VC-Bench，包括数据集的构建和评估指标的制定。表 1 汇总了与视频生成相关的基准测试。与以往主要关注文本生成视频和图像生成视频任务评估的工作相比，我们的基准测试侧重于对视频生成模型进行全面评估。

4.1 数据集构建

VC-Bench 数据集经过精心整理，来源于多个公共视频资源平台，例如 Pexels、Pixabay、Mixkit 和 YouTube。其构建流程如图 3 所示。
除美学筛选、动态运动筛选、水印去除和字幕生成等常规步骤外，我们还针对视频连接任务引入了额外流程：场景检测、周期性运动检测和视频片段提取。

场景检测：使用 PySceneDetect 场景检测器对原始视频进行自动分析，以确保每个视频仅包含一个场景或恰好包含两个场景。这一做法基于以下理由：任意数量场景之间的视频连接，都可以通过连接包含两个场景的视频来实现。

周期性运动检测：我们排除了包含明显周期性运动（例如弹簧振荡、哑铃举重）的视频，因为这类内容无法充分体现视频连接任务的性质。我们使用结构相似性指数（SSIM）计算第一帧与后续各帧之间的相似度。通过 SciPy（Virtanen 等人，2020）检测峰值，并计算相邻峰值之间的平均帧间隔来估算周期。

视频片段提取：我们将视频总时长设为 5 秒，起始和结束片段的时长则在 2 至 4 秒之间变化。对于包含两个场景的视频，我们确保这些片段不包含过渡帧。这样设计旨在评估现有视频生成模型重建视频过渡部分的能力。

我们详细介绍数据集的统计信息，以便全面了解该数据集。数据集共包含 1,579 个视频。视频时长从 4 秒到 43 秒不等，其中 7.5 至 20 秒的视频最为常见。视频质量方面，所有视频的分辨率均超过 720p，并经过美学评分预测器筛选，平均美学评分高达 0.55，如图4所示。此外，我们使用 Qwen2-VL（Bai 等人，2023；Wang 等人，2024a）为每个视频生成详尽且生动的文本描述。字幕长度为 18 至 57 个字符。

在视频分类方面，我们采用基于BERT的文本分类模型，根据视频的文本描述对视频进行分类，大幅缩短了数据集构建时间。如图4所示，该数据集涵盖15个主要类别和72个子类别，确保内容多样。

4.2 评估指标

鉴于视频连接任务的特点，有必要设计专门的指标来评估视频内容的质量。需要指出的是，受当前条件视频生成模型的局限，起始视频片段与结束视频片段之间仍难以实现无缝过渡，因此传统评估方法并不足够。为此，我们提出**VC-Bench** ，以更准确地评估生成视频内容，为视频生成技术的研究与优化提供更可靠的依据。

我们的工作提出了一套全面的视频生成质量评估指标，涵盖以下三个方面：**视频质量分数** VQS、**首尾一致性分数**$SECS$ 和**过渡平滑度分数** TSS。具体如下。

视频质量分数 VQS**：** 受VBench（Huang等，2024年）的启发，我们提出了一种综合视频质量评估方法，基于五项关键指标评估生成视频的真实感和整体表现。

(1) 主体一致性 Q_S: 该指标用于评估视频中的主要主体（例如人物、动物或物体）在整个序列中是否保持身份、外观和结构一致。
(2) 背景一致性 Q_B: 该指标用于衡量背景随时间变化的稳定性，确保不会出现不自然的偏移或抖动。

(3) 闪烁严重程度 Q_F: 量化局部细节（例如颜色或亮度）中的闪烁伪影，以此表明生成过程是否稳定。
(4) 美学评分 Q_A: 使用预训练美学评估模型，自动预测视频的艺术质量和视觉吸引力。

(5) 成像质量 Q_I: 评估单帧中的低级视觉伪影（例如模糊、失真或噪声）。

The 上述指标分数的数值范围为 0 到 1，随后将 VQS 计算为 t
他曾
这些指标的年龄。

$$VQS = [Q_S + Q_B + (1 - Q_F) + Q_A + Q_I]/5. \tag{1}$$

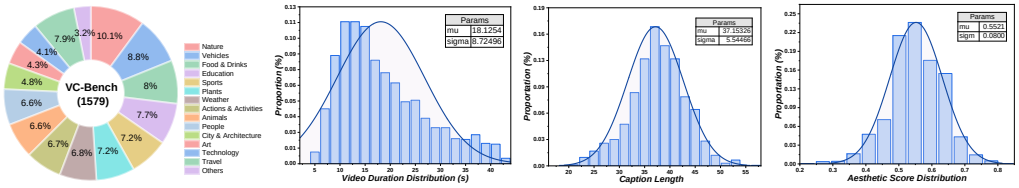


图4：**精选数据的分布**。我们分析了该数据集，包括类别统计、视频时长统计（24 fps）、字幕长度统计和美学评分统计。统计分析表明，我们的数据集具有高质量、多样性和复杂度，满足基准测试的严格要求。

首尾一致性分数 $SECS$**：** 视频转换任务对保持起始和结束视频片段之间的一致性提出了更高要求。该维度主要评估生成视频在起始和结束片段之间的像素级一致性。它包含两个指标：

(1) 像素一致性 C_P：通过比较亮度、对比度和结构相似性，评估生成视频的起始或结束片段与原始视频之间的像素对齐情况。

(2) 光流误差 C_{OF}：光流描述帧间像素的运动信息。该指标衡量生成视频与原始视频之间的运动一致性。

同样，$SECS$ 的计算方式如下。

$$SECS = [C_P + (1 - C_{OF})] / 2. \tag{2}$$

过渡平滑度分数 TSS**：** 评估原始片段与生成部分之间的过渡是否自然流畅，由两个指标组成：

(1) 视频连接距离 T_{CD}：使用DTW方法在时间上对齐生成视频帧与原始视频帧。通过计算原始视频中的帧 f_I 与生成视频中的帧 f_G 之间的结构相似性 $SSIM(f_I, f_G)$ 以及帧间距 d，来衡量过渡一致性。该指标确保过渡部分能够自然连接起始和结束视频片段，同时避免突兀的变化。

(2) 局部感知一致性 T_{LP}：通过预训练VGG网络提取特征并计算LPIPS距离，量化过渡区域中相邻帧之间的感知相似性。这可以判断过渡点是否存在感知跳变，从而评估生成视频过渡的自然流畅度。

TSS 的公式如下。

$$TSS = [(1 - T_{CD}) + T_{LP}] / 2. \tag{3}$$

最后，生成视频的综合评价分数表示为上述三个分数的平均值。各项指标的具体计算公式见附录B.2。

$$Score = (VQS + SECS + TSS)/3. \tag{4}$$

5 实验

在本节中，我们使用 VC-Bench 评估一些最先进的视频生成模型。我们首先介绍评估设置，随后展示定量比较结果。我们还评估不同起始和结束片段时长下衔接视频的质量。最后，我们分析所提出的评估指标与主观人类判断之间的一致性。

5.1 设置

我们先简要介绍本研究所采用的实验设置；更详细的配置请参阅补充材料。

	视频质量				
	主体一致性	背景一致性	闪烁严重程度（↓）	美学评分	成像质量
Wan-2.1（1.3B）	0.921	0.944	0.045	0.577	0.720
Wan-2.1(14B)	0.922	0.946	0.060	0.576	0.713
CogVideoX(2B)	0.914	0.943	0.048	0.562	0.704
CogVideoX(5B)	0.906	0.940	0.079	0.561	0.685
Open-Sora 2.0(11B)	0.911	0.938	0.028	0.537	0.644
Ruyi（7B）	0.909	0.939	0.090	0.560	0.688

	起止一致性		过渡平滑度		总分
	像素一致性	光流误差（↓）	连接距离（↓）	局部感知一致性	
Wan-2.1（1.3B）	0.933	0.042	0.022	0.839	0.892
Wan-2.1 (14B)	0.952	0.031	0.021	0.820	0.893
CogVideoX（2B）	0.880	0.058	0.077	0.810	0.864
CogVideoX（5B）	0.893	0.059	0.073	0.782	0.858
OpenSora-2.0（11B）	0.851	0.098	0.036	0.837	0.859
Ruyi (7B)	0.848	0.078	0.056	0.649	0.827

Tab表 2：**VC-Bench 上的视频评分**。基于 VC-Bench，我们评估了 6 个开放 n source 模型在视频连接任务上的表现。

	起始与结束场景相同					起始与结束场景不同				
条件比例	FLF2V	40%	60%	80%	Avg.	FLF2V	40%	60%	80%	Avg.
Wan-2.1（1.3B）	0.840	0.870	0.872	0.874	0.864	0.801	0.832	0.836	0.835	0.826
Wan-2.1 (14B)	0.851	0.871	0.873	0.875	0.868	0.802	0.822	0.831	0.834	0.822
CogVideoX (2B)	0.820	0.838	0.845	0.847	0.838	0.760	0.786	0.797	0.797	0.785
CogVideoX (5B)	0.785	0.834	0.847	0.849	0.829	0.734	0.786	0.799	0.802	0.780
OpenSora-2.0（11B）	0.756	—	0.844	0.840	0.813	0.765	—	0.763	0.765	0.764
Ruyi（7B）	0.827	0.824	0.816	0.820	0.822	0.767	0.771	0.771	0.766	0.769

表3：**时长比例对比**。 我们在 6 个测试模型上进行了对比实验，以评估起止时长比例对生成性能的影响。实验设置包括 FLF2V、40%（各 24 帧）、60%（各 36 帧）和 80%（各 48 帧），生成视频时长固定为 5 秒（120 帧）。

模型。 我们在视频连接任务上评估了六种主流视频生成模型：Wan2.1（Wang 等人，2025年）（1.3B 和 14B）、CogVideoX（Yang 等人，2024年）（2B 和 5B）、Open-Sora 2.0（Peng 等人，2025年）（11B）以及 Ruyi（Team，2024年）（7B）。

框架。 我们采用了 DiT 架构和 3D 因果变分自编码器，并根据第 3.2 节所述的方法进行了修改。通过对起始和结束视频片段进行潜在空间映射，并采用基于插值的条件控制，我们使该框架适用于视频连接任务。

条件设置。 我们选择了以下条件设置：使用单帧、1 秒（24 帧）、1.5 秒（36 帧）和 2 秒（48 帧）作为起始和结束片段。文本提示为可选项。所有情况下，均生成了 5 秒的过渡视频。

5.2 主要结果

我们使用 VC-Bench 评估了六种适用于视频连接任务的视频生成模型。为统一评估标准，我们将负向指标用 1 减去原值的方式转换为正向指标，并将具有范围的指标归一化到 0–1 区间。结果见表2和图2。

如表2所示，Wan-2.1（1.3B 和 14B）优于其他模型，总分最高。它在主体一致性、背景一致性、光流误差和连接距离方面表现出色，展现出卓越的时空连贯性和过渡自然度。相比之下，CogVideoX（2B 和 5B）在美学评分和成像质量方面表现强劲，但在闪烁严重程度和过渡平滑度方面效果较差，因此适用于优先考虑视觉吸引力的短视频。Open-Sora-2.0（11B）在闪烁严重程度方面领先，凭借精细调整的噪声控制，有效抑制了颜色或亮度波动等不稳定现象，但其起止一致性表现一般。由于缺乏针对视频连接任务的充分训练，Ruyi（7B）在过渡平滑度方面表现不佳，导致过渡突兀，降低了整体流畅度。

总而言之，当前的视频生成模型在我们的 VC-Bench 上展现出很强的判别能力。然而，它们在起止一致性和过渡平滑度方面仍明显不足。这些缺陷凸显了视频连接任务中维持时间连贯性、确保视频片段之间过渡流畅所面临的持续挑战。继续

得分	客观平均值	主观平均值	相关系数	主观一致性
VQS	0.815	0.807	0.839	0.968
SECS	0.911	0.916	0.905	0.980
TSS	0.867	0.824	0.812	0.979

表 4：**验证 VC-Bench 与人类的一致性**。我们邀请了 30 名志愿者，对视频生成结果在视频质量、起止一致性和过渡平滑度方面进行主观评分。随后，我们计算了客观评分与主观评分之间的相关系数，以证明客观评分能够有效反映人类偏好。此外，我们还计算了志愿者之间的 ICC 指标，以确认评分具有较高的一致性。

未来的研究将优先关注条件视频生成，探索先进技术和优化的训练策略，以提升该领域的整体性能和有效性。

5.3 分析

单场景与双场景的比较。如表3所示，我们评估了模型在两种条件下生成的过渡视频质量：起始和结束场景相同，以及二者不同。结果表明，在前一种情况下生成的过渡质量始终高于后一种情况。这表明，对于当前的视频生成模型而言，双场景视频衔接仍然极具挑战性。模型难以在不同的起始和结束片段之间“构想”合理的过渡，生成的视频中常常出现明显的“幻觉”——例如，结束片段中的人物毫无逻辑地突然出现——这显然偏离了真实且连贯的运动连续性。

条件比例的影响。如表3所示，我们开展了实验，以分析随着起始和结束片段中条件帧比例的增加，衔接视频的质量如何变化。总体而言，随着条件帧比例提高，视频质量逐步提升，其中 Wan-2.1 (14B) 始终优于其他模型。值得注意的是，使用更高条件帧比例（40%、60%和80%）生成的视频质量显著优于仅使用起始帧和结束帧时的质量。这一改进可归因于模型在获得更多上下文帧后，能够捕捉到更丰富的时空特征和更强的运动一致性。

案例展示与分析。 我们从多个类别中精心挑选案例进行展示和分析，具体内容见附录 D.3。这些案例生动地展示了视频连接中的常见问题，包括主体失真（涉及动画、动物和人类）、伪影闪烁（如植物和体育）、图像失真（可见于 citeps）以及不自然的运动轨迹（如体育）。这表明现有模型的适应能力仍显不足，未来仍需进一步研究并推动技术进步。

VC-Bench 与人类的一致性。 为验证 VC-Bench 测试集与人类主观感知的一致性，我们邀请了 30 名志愿者对生成视频进行人工评估。表 4 中的相关系数显示，客观评估与主观评估高度相似，证实 VC-Bench 能有效反映人类感知，可用作自动化视频评估工具。为确保主观一致性并发现潜在的恶意评分，我们采用了 ICC(2,K) 统计方法（Koo & Li, 2016）。结果证实评分者信度较高，且没有证据表明存在故意给出高分或低分的情况。

6 结论与未来工作

本文介绍了视频连接这一条件视频生成领域的新任务，并提出了专为评估该领域性能而设计的基准测试 VC-Bench。不同于专注于整体视频质量的现有基准测试，VC-Bench 使用包含 1,579 个高质量样本的多样化数据集，对视频质量、首尾一致性和过渡平滑度进行评估。对开源模型的评估揭示了它们在时空连贯性和过渡流畅度方面面临的挑战，尤其是在跨场景设置中。其局限性包括仅测试时长为 5 秒的视频，未探索更长的序列；此外，由于缺少闭源模型的

支持，评估依赖于开源模型。未来的工作应纳入更长的视频和闭源模型，以进行稳健验证。我们希望 VC-Bench 能推动视频连接生成模型的发展。

7 道德声明

本研究遵循 ICLR 道德准则。所提出的 VC-Bench 完全由公开可用的视频构建，这些视频不含个人身份信息或敏感信息。所有数据来源均遵循其原始许可协议，处理步骤见第 4.1 节。我们的基准测试和评估框架旨在推动可控且连贯的视频生成研究，并在电影制作、短视频创作、广告和虚拟现实等领域发挥积极作用。我们承认生成式视频模型可能被滥用于制作误导性或有害内容；但我们的贡献仅限于基准测试和评估指标，旨在推进学术研究。所有程序均经机构伦理审查委员会批准，所有参与者在参与前均提供了知情同意。向人类受试者收集的数据仅限于标注反馈和对视频的主观质量评分，未收集任何个人身份信息或敏感信息。我们未发现生成模型研究通常涉及的伦理风险之外的其他伦理风险。

8 可复现性声明

我们采取了多项措施，以确保本研究具有可复现性。VC-Bench 数据集的构建和预处理步骤见第 4.1 节。所提出方法和评估指标的详细说明见论文正文第 4.2 节，并在附录 B.1 中作了进一步阐述。此外，我们提供了源代码和评估脚本的匿名链接：https://anonymous.4open.science/r/VC-Bench-1B67/，以便复现报告的结果。

参考文献

Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang 等。通义千问技术报告。*arXiv* 预印本 *arXiv:2309.16609*，2023。

Hritik Bansal、Zongyu Lin、Tianyi Xie、Zeshun Zong、Michal Yarom、Yonatan Bitton、Chenfanfu Jiang、Yizhou Sun、Kai-Wei Chang 和 Aditya Grover。VideoPhy：评估视频生成的物理常识。*arXiv* 预印本 *arXiv:2406.03520*，2024年。

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至大型数据集。*arXiv* 预印本 *arXiv:2311.15127*，2023。

Emanuele Bugliarello、H Hernan Moraldo、Ruben Villegas、Mohammad Babaeizadeh、Moham- mad Taghi Saffar、Han Zhang、Dumitru Erhan、Vittorio Ferrari、Pieter-Jan Kindermans 和 Paul Voigtlaender。StoryBench：用于连续故事可视化的多面向基准测试。神经信息处理系统进展，36:78095–78125，2023。

Andong Chen、Yuchen Song、Kehai Chen、Xuefeng Bai、Muyun Yang、Liqiang Nie、Jie Liu、Tiejun Zhao 和 Min Zhang。让想象更清晰！基于稳定扩散的多模态机器翻译视觉想象。在计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 26567–26583 页，2025年。

DeepMind。Haiper ai。*https://haiper.ai/home*，2024年。

Fanda Fan、Chunjie Luo、Wanling Gao 和 Jianfeng Zhan。AIGCBench：对人工智能生成的图像转视频内容进行全面评估。*BenchCouncil*基准、标准与评估汇刊，3(4):100152，2023。

Weixi Feng、JiachenLi、Michael Saxon、Tsu-jui Fu、Wenhu Chen 和 William YangWang。TC-Bench：评测文本到视频和图像到视频生成中的时间组合性。*arXiv* 预印本 *arXiv:2406.08656*，2024年。

Yoav HaCohen、Nisan Chiprut、Benny Brazowski、Daniel Shalem、Dudu Moshe、Eitan Richardson、Eran Levin、Guy Shiran、Nir Zabari、Ori Gordon 等。LTX-Video：实时视频潜在扩散。*arXiv* 预印本 *arXiv:2501.00103*，2024年。

Hui Han、Siyuan Li、Jiaqi Chen、Yiwen Yuan、Yuling Wu、Chak Tou Leong、Hanwen Du、Junchen Fu、Youhua Li、Jie Zhang 等。《Video-bench：人类对齐视频生成基准》。*arXiv* 预印本 *arXiv:2504.04907*，2025。

Yaosi Hu、Chong Luo 和 Zhenzhong Chen。《可控文本到图像到视频生成基准》。《style id='1'>IEEE多媒体汇刊，26：1706–1719，2024年。DOI：10.1109/TMM.2023. 3284989。

Haoyang Huang、Guoqing Ma、Nan Duan、Xing Chen、Changyi Wan、Ranchen Ming、Tianyu Wang、Bo Wang、Zhiying Lu、Aojie Li 等。《Step-video-ti2v 技术报告：最先进的文本驱动图像到视频生成模型》。*arXiv* 预印本 *arXiv:2503.11251*，2025年。

Ziqi Huang、Yinan He、Jiashuo Yu、Fan Zhang、Chenyang Si、Yuming Jiang、Yuanhan Zhang、Tianxing Wu、Qingyang Jin、Nattapol Chanpaisit 等。Vbench：面向视频生成模型的综合基准套件。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 21807–21818 页，2024年。

Pengliang Ji、Chuyang Xiao、Huilin Tai 和 Mingxiao Huo。T2vbench：对文本到视频生成的时间动态进行基准测试。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 5325–5335 页，2024年。

Jianying。即梦 AI。*https://jimeng.jianying.com/ai-tool/home*，2024年。

Xuan Ju、Yiming Gao、Zhaoyang Zhang、Ziyang Yuan、Xintao Wang、Ailing Zeng、Yu Xiong、Qiang Xu 和 Ying Shan。Miradata：包含长时长视频和结构化描述的大规模视频数据集。神经信息处理系统进展，37:48955–48970，2024年。

Weijie Kong、Qi Tian、Zijian Zhang、Rox Min、Zuozhuo Dai、Jin Zhou、Jiangfeng Xiong、Xin Li、Bo Wu、Jianwei Zhang 等。混元视频：大型视频生成模型的系统化框架。*arXiv* 预印本 *arXiv:2412.03603*，2024年。

Terry K Koo 和 Mae Y Li。信度研究中选择和报告组内相关系数的指南。《脊椎按摩医学杂志》，15(2):155–163，2016年。

快手。可灵 AI。*https://klingai.kuaishou.com/*，2024年。

Chengxuan Li、Di Huang、Zeyu Lu、Yang Xiao、Qingqi Pei 和 Lei Bai。长视频生成综述：挑战、方法与前景。*arXiv* 预印本 *arXiv:2403.16407*，2024年。

Mingxiang Liao、Qixiang Ye、Wangmeng Zuo、Fang Wan、Tianyu Wang、Yuzhong Zhao、Jingdong Wang、Xinyu Zhang 等。文生视频模型评估：动力学视角。神经信息处理系统进展，37:109790–109816，2024年。

Zhiqiu Lin、Deepak Pathak、Baiqi Li、Jiayao Li、Xide Xia、Graham Neubig、Pengchuan Zhang 和 Deva Ramanan。利用图像到文本生成评估文本到视觉内容生成。载于欧洲计算机视觉会议，第 366–384 页。施普林格，2024年。

Xinran Ling、Chen Zhu、Meiqi Wu、Hangyu Li、Xiaokun Feng、Cundian Yang、Aiming Hao、Jiashu Zhu、Jiahong Wu 和 Xiangxiang Chu。VMBench：用于感知对齐的视频运动生成的基准测试。*arXiv* 预印本 *arXiv:2503.10076*，2025年。

Yaofang Liu、Xiaodong Cun、Xuebo Liu、Xintao Wang、Yong Zhang、Haoxin Chen、Yang Liu、Tieyong Zeng、Raymond Chan 和 Ying Shan。EvalCrafter：大型视频生成模型的基准测试与评估。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 22139–22149 页，2024年。

LumaLabs。Dream Machine。*https://lumalabs.ai/dream-machine/*，2024年。

Fanqing Meng、Jiaqi Liao、Xinyu Tan、Wenqi Shao、Quanfeng Lu、Kaipeng Zhang、Yu Cheng、Dianqi Li、Yu Qiao 和 Ping Luo。迈向世界模拟器：为视频生成打造基于物理常识的基准测试。*arXiv*预印本 *arXiv:2410.05363*，2024年。

Haomiao Ni、Changhao Shi、Kai Li、Sharon X Huang 和 Martin Renqiang Min。基于潜在流扩散模型的条件图像到视频生成。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18444–18455页，2023。

OpenAI。作为世界模拟器的视频生成模型。2024年。 *https://openai.com/research/video-generation-models-as-world-simulators*，2024年。

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 200 k 上训练商业级视频生成模型。*arXiv* 预印本 *arXiv:2503.09642*，2025年。

PikaLabs。Pika。*https://pika.art/*，2024年。

Runway。Gen-3。*https://runwayml.com/*，2024年。

ShengShu-AI。Vidu。*https://www.vidu.studio/*，2024年。

Kaiyue Sun、Kaiyi Huang、Xian Liu、Yue Wu、Zihan Xu、Zhenguo Li 和 Xihui Liu。T2v-compbench：用于组合式文本到视频生成的综合基准测试。*arXiv* 预印本 *arXiv:2407.14505*，2024年a。

Rui Sun、Yumin Zhang、Tejal Shah、Jiahao Sun、Shuoying Zhang、Wenqi Li、Haoran Duan、Bo Wei 和 Rajiv Ranjan。从 Sora 中我们能看到什么：文本到视频生成综述。*arXiv* 预印本 *arXiv:2405.10674*，2024年b。

CreateAI 团队。Ruyi-mini-7b。https://github.com/IamCreateAI/Ruyi-Models，2024年。

Pauli Virtanen、Ralf Gommers、Travis E. Oliphant、Matt Haberland、Tyler Reddy、David Cournapeau、Evgeni Burovski、Pearu Peterson、Warren Weckesser、Jonathan Bright、Stéfan J. van der Walt、Matthew Brett、Joshua Wilson、K. Jarrod Millman、Nikolay Mayorov、Andrew R. J. Nelson、Eric Jones、Robert Kern、Eric Larson、C J Carey、 ̇Ilhan Polat、Yu Feng、Eric W. Moore、Jake VanderPlas、Denis Laxalde、Josef Perktold、Robert Cimrman、Ian Henriksen、E. A. Quintero、Charles R. Harris、Anne M. Archibald、Antônio H. Ribeiro、Fabian Pedregosa、Paul van Mulbregt 和 SciPy 1.0 贡献者。SciPy 1.0：Python 科学计算的基础算法。《自然·方法》，17：261–272，2020年。doi：10.1038/s41592-019-0686-2。

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, JianxiaoYang, Jianyuan Zeng 等。Wan：开放且先进的大规模视频生成模型。*arXiv* 预印本 *arXiv:2503.20314*，2025年。

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge 等。Qwen2-VL：增强视觉语言模型在任意分辨率下对世界的感知。*arXiv* 预印本 *arXiv:2409.12191*，2024a。

Yiping Wang、Xuehai He、Kuan Wang、Luyao Ma、Jianwei Yang、Shuohang Wang、Simon Shaolei Du 和 Yelong Shen。你的世界模拟器是出色的故事讲述者吗？面向未来长视频生成的连续事件基准。*arXiv* 预印本 *arXiv:2412.16211*，2024b。

Zhuoyi Yang、Jiayan Teng、Wendi Zheng、Ming Ding、Shiyu Huang、Jiazheng Xu、Yuanming Yang、Wenyi Hong、Xiaohan Zhang、Guanyu Feng 等。Cogvideox：采用专家Transformer的文本到视频扩散模型。*arXiv*预印本 *arXiv:2408.06072*，2024年。

Shenghai Yuan、Jinfa Huang、Yongqi Xu、Yaoyang Liu、Shaofeng Zhang、Yujun Shi、Rui-Jie Zhu、Xinhua Cheng、Jiebo Luo 和 Li Yuan。Chronomagic-bench：用于文本到延时视频生成变形评估的基准测试。神经信息处理系统进展，37：21236–21270，2024年。

Lvmin Zhang 和 Maneesh Agrawala。为视频生成的下一帧预测模型打包输入帧上下文。*Arxiv*，2025年。

Rui Zhang、Yaosen Chen、Yuegen Liu、Wei Wang、Xuming Wen 和 Hongxia Wang。《TVG：一种无需训练的扩散模型转场视频生成方法》。*arXiv* 预印本 *arXiv:2408.13413*，2024年。

视频生成任务	条件	输入长度	运动信息	场景	信息复杂度
T2V	文本提示	仅文本	是（隐式）	Free	High
I2V	初始图像	1 帧	No	Same	中等
视频扩展	初始视频片段	∼24-48 帧	Yes	Same	中等
帧插值	两个相邻帧	2帧	No	Same	小
FLF2V	首帧/末帧	2帧	No	Same	中等
VC	**两个视频片段**	**∼48-96帧**	**Yes**	**不同**	**High**

表5：视频衔接任务与其他视频生成任务的区别。

Yingjie Zhu、Xuefeng Bai、Kehai Chen、Yang Xiang、Jun Yu 和 Min Zhang。《大型视觉语言模型在基础视觉图理解与推理方面的基准测试与改进》。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第30678–30701页，2025年。

大型语言模型（LLM）的使用

根据关于使用大型语言模型（LLM）的政策，我们在此披露，准备本手稿期间曾将LLM用作辅助工具。LLM（具体来说是DeepSeek）仅用于提升叙述文本的清晰度、流畅度和整体可读性。

LLM的具体作用仅限于：

- 文本润色与完善：改写和重述句子，以提升语法准确性和文体流畅度。
- 提升连贯性：协助确保段落和章节之间过渡顺畅。

必须强调的是，大语言模型并未参与本工作的核心智力内容。所有研究构思、方法设计、数据分析、结果解释、科学结论以及手稿初稿的撰写，均完全由人类作者完成。

B 方法论与视频衔接任务说明

B.1 视频衔接任务与其他视频生成任务的区别

与文本生成视频、图像生成视频、视频扩展和帧插值等其他视频生成任务相比，本文提出的视频衔接任务要求高得多。例如，文本生成视频仅依赖文本提示，并通过隐式推断运动，允许模型生成自由场景视频；而图像生成视频以单帧为条件，不包含运动信息，这两种任务的信息要求都相对较低。视频扩展则取用来自同一场景的一段短连续视频片段（24–48 帧），侧重于延长时间跨度；帧插值和 FLF2V 仅处理两个相邻帧或边界帧，不需要明确的运动线索，且始终处于固定场景中，因此复杂度较低至中等。

相比之下，视频衔接以两个完整的视频片段作为输入，每个片段约含 48–96 帧，且都包含丰富的时序运动模式，并分别源自不同场景。这种设置要求模型不仅要理解并保留每个片段中的运动动态，还要弥合语义鸿沟、协调场景差异并生成连贯的过渡。这些要求使视频衔接成为一项高度复杂且信息密集的任务，难度远高于以往的视频生成设置。视频转换任务与其他视频生成任务的区别如表 5 所示。

B.2 评估指标详情

本节将详细介绍基准测试中的各项指标及其实现方法。

(1) 主体一致性 Q_S：该指标用于评估视频中的主要主体（例如人物、动物或物体）在整个序列中是否保持身份、外观和结构一致。它确保同一主体在整个视频中始终连贯一致。在生成视频中，应保留输入视频起始帧和结束帧中的主体，同时确保中间过渡部分也保持一致。我们采用 Dino 模型来实现主体一致性检测。由于 Dino 模型未经训练，无法对主体进行分类并忽略其类内差异，因此它对视频中主体差异的变化尤为敏感，适合用于检测主体变化。计算公式如下：

$$Q_S = \frac{1}{N-2}\sum_{t=2}^{N-1}\frac{1}{3}\left(D\left\langle I_1, I_t\right\rangle + D\left\langle I_{t-1}, I_t\right\rangle + D\left\langle I_t, I_N\right\rangle\right), \tag{5}$$

其中，N 表示视频帧数，I_t 表示第 $t-th$ 帧，$D\langle\cdot,\ \cdot\rangle$ 表示两帧之间 Dino 特征余弦相似度的点积。通常，我们通过计算每一帧与视频首帧、前一帧和末帧的平均余弦相似度，来检测当前帧是否保持主体一致性。最后，我们对所有检测帧取平均值，作为最终的主体一致性得分。

(2) 背景一致性 Q_B：除主体一致性外，视频背景也需要保持一致。该指标用于衡量背景随时间变化的稳定性，确保视频中没有不自然的偏移或抖动。具体而言，我们会计算视频每一帧的 CLIP 特征图，然后仿照主体一致性的计算方式，计算每一帧与视频第一帧、前一帧和最后一帧之间 CLIP 特征的平均余弦相似度，以检测背景的一致性。最后，我们将所有检测帧的结果取平均值，作为最终的背景一致性分数。计算公式如下：

$$Q_B = \frac{1}{N-2}\sum_{t=2}^{N-1}\frac{1}{3}\left(C\left\langle I_1, I_t\right\rangle + C\left\langle I_{t-1}, I_t\right\rangle + C\left\langle I_t, I_N\right\rangle\right), \tag{6}$$

其中，$C\langle\cdot,\cdot\rangle$ 表示两帧之间 CLIP 特征余弦值的点积。

(3) 闪烁严重程度 Q_F：在生成视频中，连续帧之间的亮度、颜色、纹理或内容常常会发生快速且不规则的变化，从而产生视觉不稳定感。这种现象在视频生成中尤为常见，并会严重影响观看体验。常见的闪烁包括亮度闪烁和颜色闪烁。在我们的指标中，我们首先将图像划分为互不重叠的区域，然后根据图像的 YUV 空间（关注图像亮度变化）和 HSV 空间（关注图像颜色变化）判断相邻帧之间是否存在闪烁。我们使用以下公式计算两帧之间闪烁区域所占的比例：

$$r_t = \frac{\sum_{P_t \in S_t} I\left(\left(\mathcal{L}\left\langle P_t, P_{t+1}\right\rangle + \mathcal{C}\left\langle P_t, P_{t+1}\right\rangle\right)/2 > \eta\right)}{|S_t|}. \tag{7}$$

其中，S_t 表示将 $t-th$ 帧划分后得到的互不重叠区域集合，P_t 表示划分出的区域。$I(\cdot)$ 是一个指示函数。如果输入值大于阈值η，则取值为 1，否则取值为 0。它主要用于根据某一区域的变化程度判断该区域是否为闪烁区域。$\mathcal{L}\langle\cdot,\ \cdot\rangle$ 表示 YUV 空间中两个相邻帧的同一区域的亮度变化，$\mathcal{C}\langle\cdot,\ \cdot\rangle$ 表示 HSV 空间中的颜色变化。如果两个变化值的平均值大于阈值，则认为该区域正在闪烁。通过比例定义，可以得到闪烁区域所占的比例。最后，我们可以计算视频的平均闪烁比例 $\bar{r}_t$，以表示视频的闪烁严重程度。需要注意的是，该指标是负向指标，在可视化时需要进行正向处理。

(4) 美学评分 Q_A：它用于量化图像或视频的视觉美感，广泛应用于图像生成、视频合成、摄影评估及其他领域。它衡量的是

内容的美学质量，例如构图、颜色和主题吸引力，这些内容通过算法或人工评分进行评估。我们使用预训练的图像美学质量预测器LAION，对每一帧进行美学评分（0-10），计算平均美学评分，最后将分数线性归一化至0-1。

(5) 成像质量 Q_I：它主要考量生成视频中的帧级生成质量，重点关注帧图像中的失真、模糊和噪声等问题。我们使用预训练的 MUSIQ 模型作为图像质量预测器，对每一帧进行评分。最后，计算整个视频序列的平均分，并除以 100，将其归一化到 $0-1$ 范围内。

(6) 像素一致性 C_P：该指标用于评估生成视频起始和结束片段的像素级对齐能力。具体而言，对于每一对位置对应的原始帧和生成帧 ˆ(I_t, I_t)，我们使用 SSIM 作为指标，衡量两帧之间的像素级相似性。SSIM 综合考量亮度、对比度和结构相似性等像素级特征，能够很好地评估起始视频和结束视频所保持的像素一致性。最后，我们取平均值作为衡量视频像素一致性的指标。计算公式如下：

$$C_P = \frac{1}{N_s + N_e} \sum_{t=1}^{N_s+N_e} SSIM(I_t, \hat{I}_t), \tag{8}$$

其中 N_s 和 N_e 分别表示起始视频和结束视频的帧数。

(7) 光流误差 C_{OF}：光流描述连续视频帧之间逐像素的运动，表示每个像素随时间在二维平面上的移动方式。该指标对于评估生成视频与原始视频之间的运动一致性至关重要，可确保动态元素（例如物体运动、相机运动）自然对齐。具体而言，对于位置对应的每一对原始帧和生成帧（I_t， I_t），我们计算其光流 $OF()$，并计算平均光流误差。值得注意的是，光流误差的取值范围不在 $0-1$范围内，但由于我们的光流窗口大小为 16，因此很容易得知光流误差的最大值为 32（基于曼哈顿距离）。我们将其除以 32，将其归一化到 $0-1$。

$$C_{OF} = \frac{1}{N_s + N_e - 1} \sum_{i=1}^{N_s+N_e-1} \left| OF(I_t) - OF(\hat{I}_t) \right|, \tag{9}$$

(8) 视频连接距离 T_{CD}：该指标用于衡量中间生成部分与起始视频和结束视频之间的连接距离。具体而言，我们首先使用 DTW 方法，将生成视频的帧与原始视频的起始和结束片段对齐。然后，对于起始（结束）视频中一对对应的帧（I， I_G），我们随机选取中间生成部分的帧I_M，并分别计算 $SSIM\langle I,\ I_M\rangle$ 和 $SSIM\langle I_G,\ I_M\rangle$。我们希望这两个值之间的差异较小。不过，考虑到从 I_M 到 I 和 I_G 的帧间距离可能不同，我们使用 $SSIM/d$ 作为衡量标准，其中 d 是两帧之间的帧间距离。公式如下：

$$T_{CD} = \left| \frac{SSIM\langle I, I_M\rangle}{d_{I_M \to I}} - \frac{SSIM\langle I_G, I_M\rangle}{d_{I_M \to I_G}} \right|. \tag{10}$$

我们分别从起始视频和结束视频中选取 K 组对应帧，并从中间生成部分选取 Z 帧进行计算，最后计算平均值，以此作为该指标的表示。

(9) 局部感知一致性 T_{LP}：为了严格评估生成视频中过渡的自然度，我们提出了一种基于深度特征提取和 LPIPS 的感知相似度指标。该方法着重检测过渡区域（例如场景切换或插值片段）中相邻帧之间突兀的感知不连续性，这对于评估时间连贯性至关重要。具体而言，我们首先使用 VGG模型逐帧提取特征，然后计算每两帧之间特征的绝对差值，并取平均值作为这两帧的感知误差。最后，对所有帧取平均值，得到视频的感知误差。由于该误差是一个负向指标，且其取值介于 0 和 1 之间，因此我们用 1 减去该

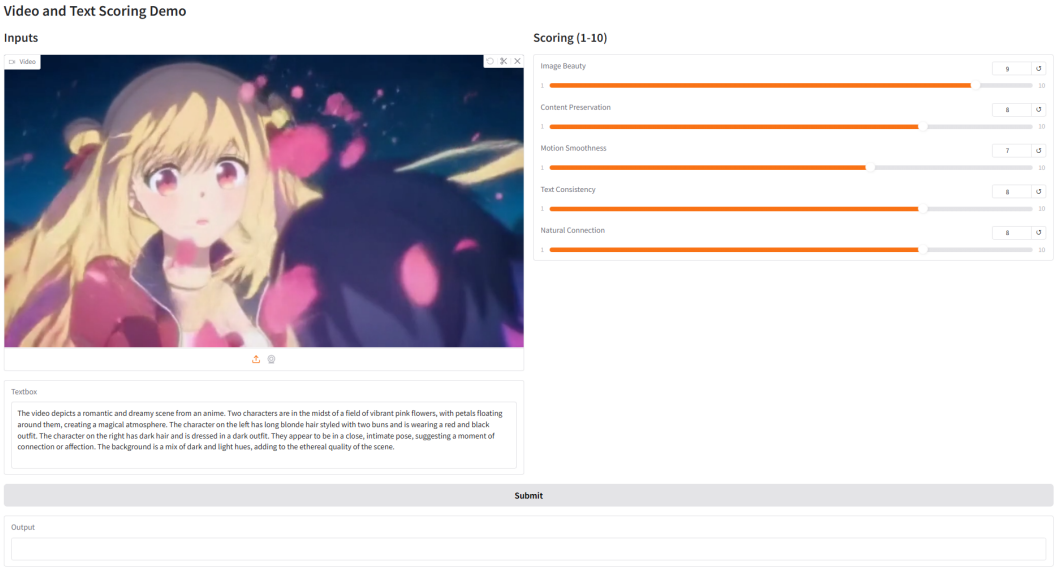


图5：主观人工评估的说明文字和截图。

指标	两个片段	三个片段	四个片段	五个片段
主体一致性	0.844	0.836	0.828	0.835
背景一致性	0.915	0.909	0.913	0.914
闪烁严重程度（↓）	0.278	0.262	0.290	0.255
美学评分	0.538	0.542	0.554	0.545
成像质量	0.625	0.630	0.623	0.629
像素一致性	0.914	0.908	0.917	0.916
光流误差（↓）	0.095	0.083	0.102	0.073
连接距离（↓）	0.143	0.149	0.155	0.142
局部感知一致性	0.515	0.524	0.518	0.516

表6：多片段连接的扩展。

值，作为局部感知一致性指标。感知误差公式如下 ：

$$\varepsilon = Mean(|\mathcal{V}(I_t) - \mathcal{V}(I_{t-1})|). \tag{11}$$

其中，$\mathcal{V}(\cdot)$表示使用 VGG模型 提取的时空特征，而$Mean()$表示根据特征维度计算的均值。

B.3 主观人工评估的构建

我们邀请了30名志愿者，请他们从视频质量、起止一致性和过渡平滑度三个方面进行评估。视频质量采用10分制评分（1分表示“极差”，10分表示“完美”）。人类主观评分通过主观评估方法计算得出。主观人工评估的说明文字和截图如图5所示。

C 实验扩展

C.1 扩展至多片段视频连接

VC-Bench 的设计天然支持更复杂的场景，包括多片段衔接和循环视频生成。为说明这一点，我们在一个小型多片段测试集上评估了 Wan2.1，该测试集包含由 2、3、4 和 5 个片段组成的衔接。实验结果如表6所示。

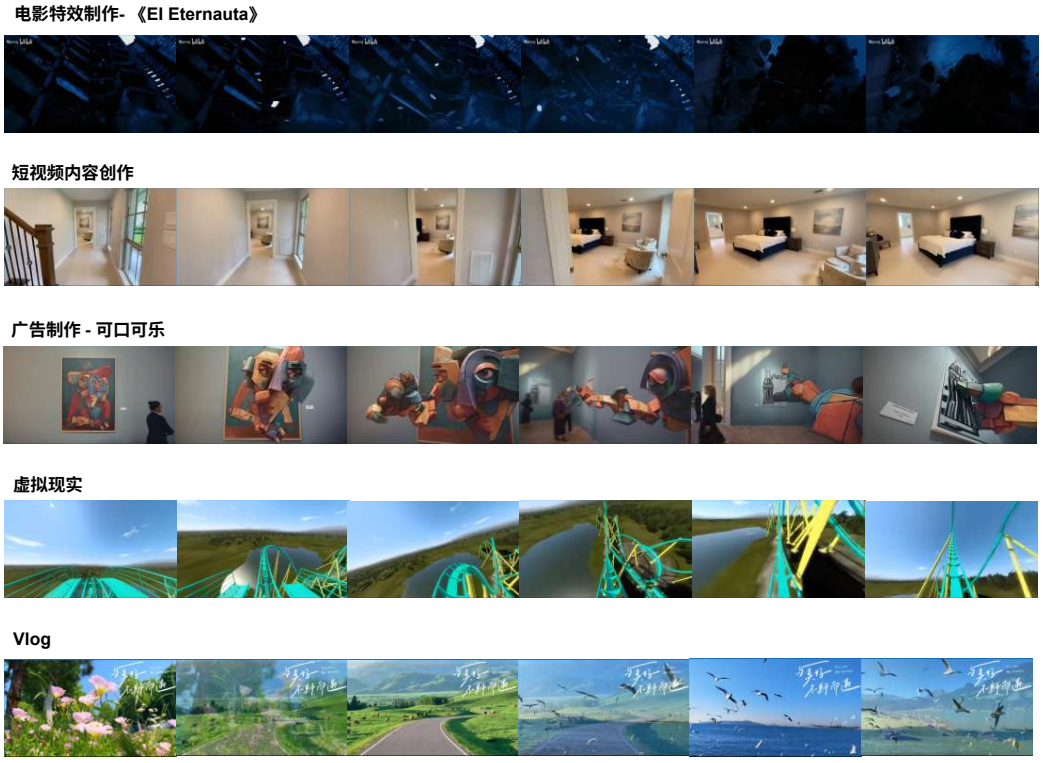


图 6：视频连接的应用

C.2 视频连接的应用

如图6所示，视频连接任务在创意和工业领域有着广泛的适用性。在电影制作和视觉效果领域，视频衔接可以生成过渡镜头或实现平滑的场景切换，从而减少昂贵补拍和后期制作的需求。对于 TikTok 或 YouTube Shorts 等短视频平台，视频衔接使创作者能够将零散片段连接成连贯叙事，提升流畅度和观众参与度。广告宣传活动也能受益于产品展示片段与叙事片段之间的无缝过渡。除传统媒体外，视频衔接还支持虚拟现实和增强现实中的沉浸式体验；在这些场景中，连续性对于保持真实感和用户沉浸感至关重要。即使是在旅行视频博客等个人媒体中，视频衔接也能让业余创作者将不连续的风景镜头连接成流畅且引人入胜的故事。这些应用凸显了视频衔接作为一种具备连续性感知能力的视频生成多功能工具的价值。

D VC-Bench 的分析与稳健性

D.1 敏感性分析

由于所有开源模型生成的视频时长固定为 5 秒，延长视频时长自然会导致分数下降，因此这一点是可以预料的。我们希望客观评估社区模型的最佳性能。相反，我们使用三种输出分辨率： 640×320、 960×432 和 1024×576，在相同的模型上评估其对视频分辨率的敏感性。实验结果如表 7 所示。

在所有指标中，只有**美学评分** 对视频分辨率表现出统计学上显著的敏感性（F=6.18，p=0.020）。其他所有指标均无显著差异（$p > 0.17$），这表明我们的评估框架具有很强的稳健性。

指标	F值	p值
主体一致性	0.146	0.866
背景一致性	0.777	0.488
闪烁严重程度（↓）	0.081	0.923
美学评分	**6.179**	**0.020**
成像质量	2.126	0.175
像素一致性	0.005	0.995
光流误差（↓）	1.235	0.336
连接距离（↓）	0.169	0.847
局部感知一致性	0.005	0.995

表 7：指标对分辨率的敏感性分析。

案例类型	VCD 值	定性示例	现象描述
更好	0.011	一只小狮子跟在一只河马后面，身处水中	视频连贯且稳定，物体运动平滑且无闪烁。
中等	0.025	场景聚焦于一位海滩上的人物	正面和背面人物视图之间的切换很突兀，但主体结构保持稳定。
Poor	0.052	前景中有一只卡通熊猫位于一家街边商店前	时间上不连贯，存在明显的物体变形和内容在过渡帧中跳变。

表8：视频连接距离的定性示例。

D.2 指标可解释性表示

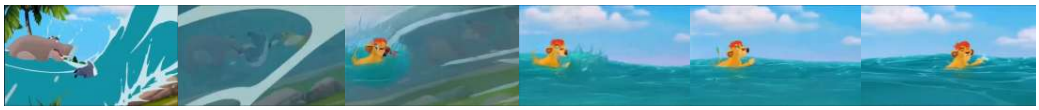
在本节中，我们选取具体案例来展示部分指标的可解释性，包括一致性指标（主体一致性/背景一致性）和转场平滑度指标（连接距离）。具体示例见表8、图7和图8；可以看出，我们提供的指标能够有效反映人类的主观评估。

D.3 案例研究

我们展示了不同模型在视频转场任务中生成的样本，涵盖多个类别，每个类别都包含正反例。案例图中标注了生成结果存在的问题，包括角色不一致、物体变形、时间不连续、闪烁等。详情见图9。

视频连接距离（↓）

良好 生成 与 VCD=0.011



在一个色彩鲜明的动画场景中，一头犀牛在清澈的蓝色水中游动，身后跟着一头红发、活泼好动的狮子。狮子享受着海洋，背景是晴朗的天空和蓬松的云朵。

中等 生成，VCD=0.025



视频以俯视图开始，展现海浪拍打海洋中的岩石。随后，画面转到一处海滩，一名男子赤裸上身，穿着黑色短裤，正在用药球锻炼，展现出力量与敏捷。

糟糕的 生成，VCD=0.052



在混乱的城市环境中，一个卡通浣熊角色惊慌失措地穿过碎片逃跑。背景中有一辆卡车，以及玻璃破碎、易拉罐散落的受损店面。场景充满紧张与混乱，突显浣熊在一片废墟中仓皇逃离的情景。

图7：视频连接距离的定性示例。

主体与背景一致性良好 生成



A c一对情侣坐在舒适的木制A字形小木屋里，举起酒杯互相祝酒。温暖的灯光和节日装饰环绕着他们，营造出亲密而rom静谧的氛围。 (Wan-2.1 1.3B，主体一致性 **0.982**，背景一致性 **0.969**)



视频展现了宁静的场景：白色雏菊在微风中轻轻摇曳。柔和的焦点突显了娇嫩的花瓣和鲜艳的黄色花心，衬托在一片模糊的绿色背景前，营造出宁静自然的氛围。 (Wan-2.1 1.3B，主体一致性 **0.941**，背景一致性 **0.959**)



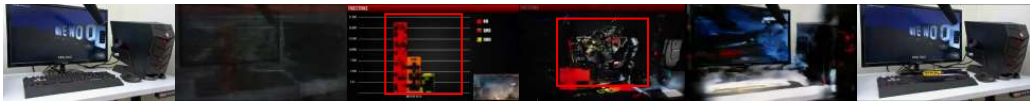
视频展现了黄昏时分宁静的海岸景观：一条狭长的陆地上分布着田野和蜿蜒的道路，平静的海水环绕其侧。天空染上柔和的粉色与橙色，接近地平线时逐渐转为深蓝色。整个场景散发着宁静与自然之美。(Wan-2.1 14B，主体一致性 **0.958**，背景一致性 **0.918**)

主体与背景一致性

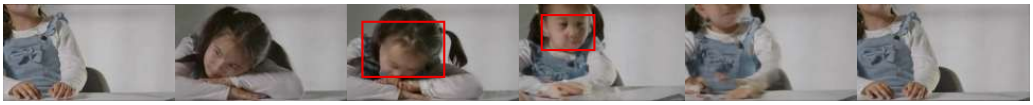
糟糕 生成



在健身房里，一男一女一边在跑步机上行走，一边击掌。视频展示了一份锻炼计划表，列出了从星期一到星期日的安排，包括力量训练、休息和跑步。 (Wan-2.1 1.3B，主体一致性 **0.731**，背景一致性 **0.863**)



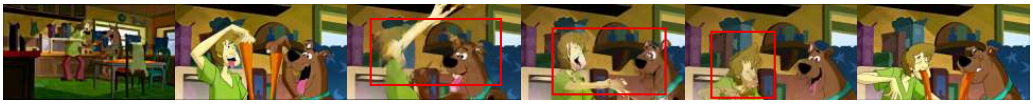
视频展示了一套游戏设备，其中有一张 MSI GTX 1070 显卡，游戏中显示着“好莱坞”。内容还包括性能指标，以及一张对比高清、四倍高清和超高清分辨率的柱状图，突出了这张显卡的性能。(CogVideoX 2B，主体一致性 **0.761**，背景一致性 **0.815**)



一个扎着双马尾、穿着蓝色连衣裙和白衬衫的小女孩坐在桌旁。她向前倾身，把头靠在手臂上，看起来若有所思或有些疲倦。她神情沉思，低头看着桌子。(CogVideoX 5B，主体一致性 **0.837**，背景一致性**0.855**)



一位女士在公园里推着一个小女孩荡秋千。女孩荡得很高，开心地笑着，享受着荡秋千的乐趣。女士面带微笑，确保女孩在来回摆动时的安全。场景以绿草、树木和停放的汽车为背景，捕捉到了一个欢乐的玩耍时刻，rl's ay.
(CogVideoX 5B，主体一致性 **0.812**，背景一致性 **0.867**)



在一个卡通场景中，两个人物——一个人类和一只狗——手持大胡萝卜。人类将胡萝卜举过头顶，狗则模仿这一动作。两者看起来都很兴奋，咧嘴微笑、眉毛扬起，背景是色彩缤纷的室内环境。 (Ruyi，主体一致性 **0.728**，背景一致性 **0.905**)

图 8：一致性指标的定性示例。

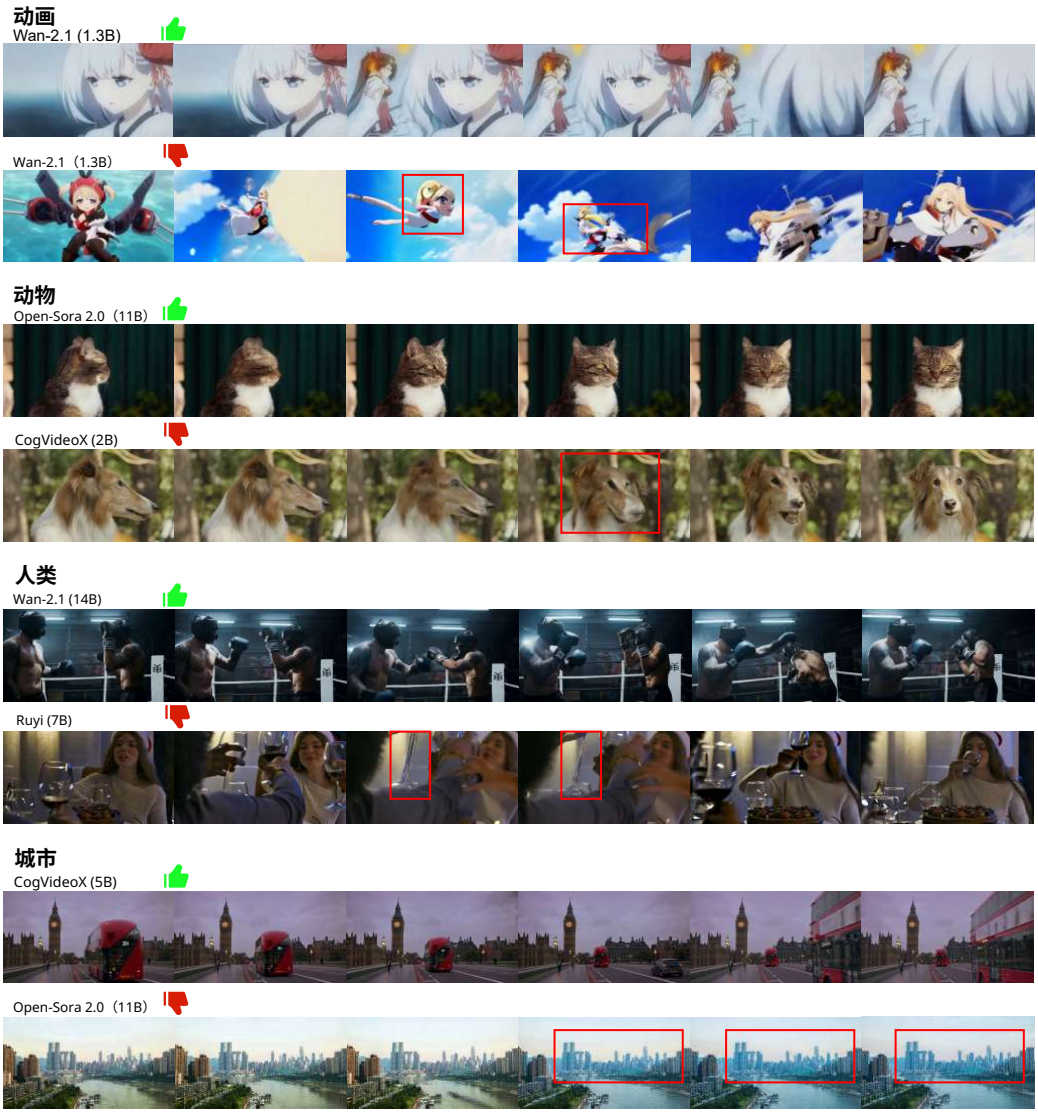

动画
Wan-2.1 (1.3B)
Wan-2.1（1.3B）
动物
Open-Sora 2.0（11B）
CogVideoX (2B)
人类
Wan-2.1 (14B)
Ruyi (7B)
城市
CogVideoX (5B)
Open-Sora 2.0（11B）

图9：案例研究。