arXiv:2602.01538v1 [cs.CV] 2 Feb 2026

让虚拟形象进行交互：迈向文本驱动的人-物体交互，实现可控的会话虚拟形象

Youliang Zhang[1,2] † Zhengguang Zhou[2] † Zhentao Yu[2] Ziyao Huang [2]Teng Hu[2]
Sen Liang[2] Guozhen Zhang[2] Ziqiao Peng[2] Shunkai Li[2]Yi Chen[2] Zixiang Zhou[2]
Yuan Zhou[2] § Qinglin Lu[2] ‡ Xiu Li[1]‡

[1]清华大学。 [2]腾讯 HY。



图 1. **输入文本、音频和参考图像后，我们的方法可以生成能够说话、表演动作并与物体交互的人物视频。接地意味着所有动作和物体交互都发生在参考图像所提供的环境中。**

摘要

生成会话虚拟形象是视频生成中的一项基础任务。尽管现有方法能够生成仅包含简单人体动作的全身会话虚拟形象，但将这一任务扩展至具身人-物交互（*GHOI*）仍是一项开放性挑战，这要求虚拟形象与周围物体执行与文本对齐的交互。该挑战源于 *GHOI* 生成对环境感知的需求，以及控制质量困境。为此，我们提出了一种新颖的双流框架 ***InteractAvatar***，将基于场景的人物与物体交互中的感知和规划与视频合成解耦。借助检测来增强环境感知，我们引入感知与交互模块（*PIM*），用于生成与文本对齐的交互动作。

此外，我们提出音频交互感知生成模块（*AIM*），用于合成执行物体交互的生动会话虚拟形象。通过专门设计的动作到视频对齐器，*PIM* 和 *AIM* 共享相似的网络结构，并支持动作与逼真视频的并行协同生成，从而有效缓解控制质量困境。最后，我们建立了一个用于评估 *GHOI* 视频生成的基准测试 ***GroundedInter***。大量实验和对比结果证明了我们的方法在为会话虚拟形象生成基于场景的人物与物体交互方面的有效性。项目主页： *https://interactavatar.github.io/*。

1. 引言

以人为中心的视频生成已取得显著进展，在照片级逼真的说话头像合成等任务上尤为突出，主要聚焦于精准的唇形同步和面部表情生成

† 共同贡献 ‡ 通讯作者 § 项目负责人 * 在腾讯 HY 实习期间完成的工作。

成 [3, 22, 27, 30, 35]。然而，要构建生动且实用的数字人，我们必须超越面部建模，赋予他们执行复杂动作并与周围环境进行有意义互动的能力。

我们提出**具身人-物交互**，用于说话头像，旨在通过环境感知和文本驱动的人-物交互生成能力，增强现有数字人。与当前的说话头像和人-物交互生成任务不同，GHOI具有以下特点：1）环境感知，能够感知周围环境并在其中执行合理的动作；2）与初始帧保持一致，而非生成新场景；3）遵循文本指令，无需提供明确的姿态或物体轨迹来进行控制；4）不需要额外的物体条件，可依据文本操作场景中的物体。

尽管近期取得了进展，现有范式仍受根本性架构局限的制约，无法实现这种细致且具备场景感知能力的交互：***(1)*音频驱动方法**，尽管已有多项开创性工作探索简单的全身动画 [3, 7, 19, 31], ，通常仍是学习从声学特征到像素空间的直接映射，缺乏对物体和环境的显式建模，因此难以控制复杂的人-物交互；***(2)*姿态驱动方法** [5, 26, 28, 33]提供了显式控制，但将规划负担转嫁给用户。这些方法要求输入预先定义的骨架序列，而这类序列不仅难以获取且成本高昂，还常常与参考图像的具体情境不匹配；***(3)* 主体一致性方法** [2, 13, 20]擅长保持主体身份并实现连贯融合，但缺乏具身交互机制。它们旨在根据文本合成新视频，而非在交互环境中执行情境相关的指令。因此，如何让说话头像稳定地执行文本驱动物体交互，仍是一项重大且尚未解决的挑战。其核心困难有两方面。

(1)* 场景-动作锚定**：GHOI 不仅要求生成包含人-物体交互的视频，还要求交互发生在特定环境中，并涉及指定物体。这要求模型理解空间布局和语义内容，并在这一视觉语境中锚定文本指令。(2)*控制与质量困境**：由于 GHOI 的复杂性，现有方法往往面临可控性与视觉质量之间的权衡。它们要么生成保真度高、却难以遵循指令的视频，要么以牺牲视频保真度为代价，获得尚可的响应。

为应对这些挑战，我们提出 InteractAvatar，一种新颖的双流扩散Transformer（DiT）框架，旨在为会话虚拟形象生成基于场景的人物与物体交互。该框架将感知与规划从视频合成中显式解耦。为清晰起见，我们将**运动**定义为渲染后的骨架姿态以及物体框。我们引入感知与交互模块（PIM）来解决落地挑战，该模块依据对参考图像的感知生成文本对齐的运动序列。PIM 在检测与运动生成任务上联合训练，以确保生成具有场景感知且与文本对齐的运动。在以音频和运动潜在特征为条件的情况下，我们提出音频-交互感知生成模块（AIM）来合成最终视频。PIM 与 AIM 共享相似的网络架构，并实现运动与视频的并行协同生成。借助专门设计的运动到视频（M2V）对齐器，我们这一对称却解耦的框架有效缓解了复杂 GHOI 视频生成中的控制质量困境，为会话虚拟形象实现生动且合理的基于场景的人物与物体交互。

为了严格评估我们的方法，尤其是其对基于场景的人-物交互的建模能力，我们构建了 GroundedInter 基准测试。GroundedInter 包含 600 个不同的测试案例，每个案例均包含一张参考图像、一段描述交互的结构化文本描述，以及相应的语音音频。在该基准测试上开展的大量实验表明，我们提出的方法在生成合理、可控且高质量的基于场景的人-物交互视频方面，显著优于现有方法。

我们的主要贡献总结如下：

- 我们提出了**InteractAvatar**，一种新颖的双流 DiT 框架，通过明确将感知规划与视频合成解耦，实现面向说话头像的文本驱动的落地式人-物交互生成。
- 我们引入了**感知与交互模块(PIM)**，通过检测和运动生成任务进行训练，以生成具备场景感知能力且与文本对齐的运动。此外，我们提出了**音频交互感知生成模块**(AIM)，并配备 M2V 对齐器，以生成具有可控性和高视觉质量的视频。
- 除核心 GHOI 生成外，我们的框架还提供**统一且灵活的多模态控制**，可接受文本、音频和运动的任意组合为输入。
- 我们建立了**GroundedInter**，这是一个用于细粒度评估 GHOI 的基准测试。实验表明，在生成具有可控性和高质量、且包含基于场景的人物与物体交互的说话头像方面，我们的方法显著优于现有方法。

2. 相关工作

2.1. 音频驱动人体动画。

以音频和人物参考图像为输入，音频驱动的人体动画旨在生成唇部运动与音频良好对齐的视频。早期研究主要集中于头肩动画，SadTalker [41]采用三维运动系数来提升真实感，

Hallo [35] 采用分层扩散方法提升唇形同步的准确性，EchoMimic [4]结合音频与面部标志点以增强稳定性，Loopy [15] 利用时序模块生成自然运动，无需手动模板。随后，研究逐步将这一能力拓展至半身和全身数字人合成，并对身体动作进行建模。Hallo3 [7] 率先将预训练 DiT 模型用于肖像动画，Huanyuanvideo-Avatar [3]实现了动态、可控情绪的多角色对话视频生成。FantasyTalking [31]提出了针对全局运动和唇部同步的双层优化方案，而 OmniHuman-1 [19] 通过混合运动条件扩充训练数据，以生成更逼真的全身视频。然而，这些方法缺乏生成更复杂的、以地面为基础的人-物体交互的能力。我们的工作将音频驱动动画拓展至融入人-物体交互，从而增强生成的虚拟形象的真实感和实用性。

2.2. 人-物交互视频生成

近年来，人-物交互（HOI）视频生成日益受到关注。一些研究致力于编辑现有素材。MIMO [21]和 AnimateAnyone2 [12] 会替换动态场景中的人物，而HOI-Swap [37] 和ReHoLD [9] 则会替换手持物体。然而，这些方法的主要局限在于无法从零开始将静态参考图像制作成动画。另一些研究则通过复杂的条件输入来实现细粒度控制。AnchorCrafter [36] 利用姿态和深度图实现交互感知的人体运动；ManiVideo [23] 使用人物和物体的精确三维模型进行精准控制；HunyuanVideo-HOMA [14] 则探索了借助新颖外观注入方法实现弱 HOI 条件。尽管如此，要获取这些复杂条件并使之与任意参考图像对齐，往往并不实际，这限制了它们在数字人应用中的使用。相比之下，我们的方法仅以文本为条件，即可生成逼真的人-物体交互视频，从而无需源视频或其他复杂的控制信号。

近年来，主体一致性视频生成也展现出合成人-物交互视频的能力。Phantom [20]通过将参考图像潜变量与视频潜变量拼接，利用基于 DiT 的模型固有的一致性。同期工作如 InterActHuman [34] 和 HunyuanCustom [13] 将主体一致性视频生成与音频驱动动画相结合，而 HuMo [2]提出了一种渐进式多模态训练范式，以实现灵活控制。然而，GHOI 与主体一致性视频生成之间的一个关键区别在于，这些方法并不保留由参考图像的环境所定义的世界。相反，它们通常会完全生成一个基于文本和主体外观的新场景，而不是将交互置于所提供的环境之中。

3. 方法

如图所示2，我们提出了一种双流扩散Transformer（DiT），用于生成文本控制的基于场景的人物与物体交互。我们的架构将这一复杂任务分解为两个子问题：感知规划和视频渲染，分别由感知与交互模块（PIM）和音频交互感知生成模块（AIM）负责。我们将**运动**定义为渲染出的人体骨架姿态和物体框。我们设计的一个关键之处在于并行共同生成结构运动序列及相应的视频帧。PIM 充当**规划中枢**，专注于高层结构，生成感知场景且与文本对齐的运动。AIM 则充当**渲染引擎**，合成高保真视频并实现精准的音频唇形同步。AIM 并非通过顺序式流水线处理，而是在整个生成过程中由 M2V 对齐器提供引导；该对齐器采用逐层残差注入机制，确保生成视频始终与 PIM 同步演进的结构运动保持一致。

两个模块均采用统一的流匹配范式进行训练。预训练的变分自编码器 (VAE) [29]将所有视觉输入（视频、运动、参考图像）编码到潜在空间中，而 T5 [25] 编码器则对文本指令进行嵌入。模型 u_θ 经过优化，以预测向量场 v_t，其目标函数如下：

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,c}\left[\|v_t(z_t) - u_\theta(z_t,t,c)\|_2^2\right], \qquad (1)$$

其中，z_0 是来自 VAE 的干净潜在表示，z_t 是其在时间步 t 处加噪后的版本，$v_t(z_t)$是向量场的真实值，而 c表示所有条件信息。

3.1. 感知与交互模块

PIM 负责解析参考图像 I_{ref} 中的环境上下文，并在文本提示 T的引导下生成时空上合理的交互运动。我们将交互运动 m 定义为一种复合视觉表征，由一系列人体骨骼关键点 $P = \{p_t\}_{t=0}^{N-1}$和物体边界框轨迹 $O = \{o_t\}_{t=0}^{N-1}$组成，其中 N 为总帧数。该运动作为一种中间表征，在特征层面注入视频生成模块 AIM，从而实现对人-物交互动态的解耦控制。

交互运动生成。 为精确控制生成过程，我们明确将运动生成划分为两种不同的任务类型：纯动作生成，仅涉及人体姿态序列 P；以及 HOI生成，同时包含姿态序列 P 和物体

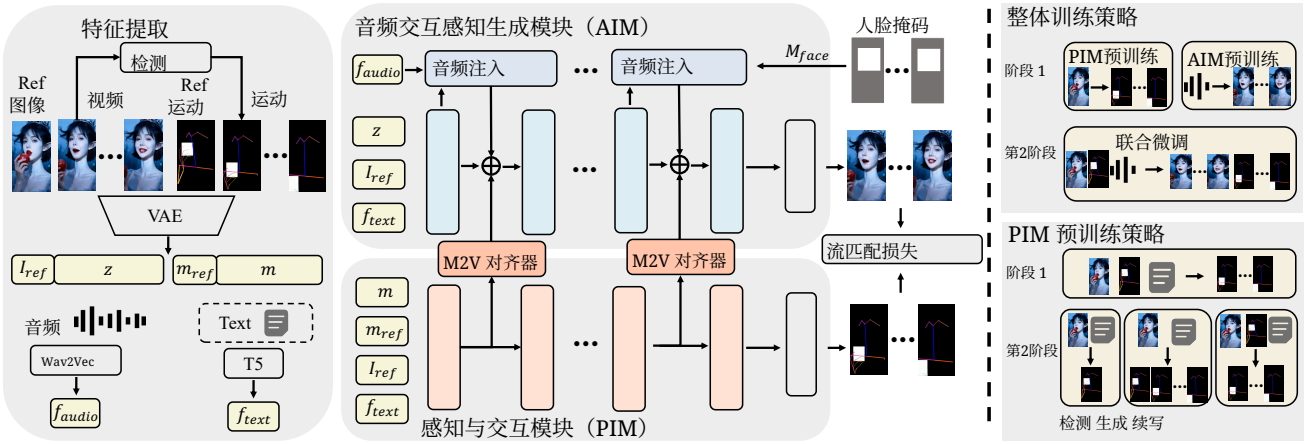


图2。**概　览我们的双流 InteractAvatar 框架及多模态条件训练策略**

轨迹 O。我们将与任务相关的信息注入模型，方法是将与任务类型 $c_{task} \in$ {ACTION、HOI} 对应的任务嵌入向量 f_{task} 与文本提示嵌入 f_{text}进行拼接。这为模型的交叉注意力层构建了统一的条件向量：

$$f_{text} = \text{Concat}(f_{text}, f_{task}). \tag{2}$$

参考图像 I_{ref} 在时间维度上前置于第一帧运动帧 m_{ref}，模型将 m_{ref} 作为首帧，生成整体运动序列 $\{m_t\}_{t=1}^{N-1}$，并以 I_{ref}为条件。为确保参考图像能够提供有效的全局引导，同时避免与运动序列的位置编码冲突，我们为其旋转位置嵌入（RoPE）采用了定制映射。对于来自参考图像的给定图块 z_I，其原始三维位置索引为（l，i，j），其位置将重新映射为：

$$\text{RoPE}_{z_I}(l, i, j) = \text{RoPE}(-1, i + w, j + h). \tag{3}$$

该策略将参考图像置于虚拟时间步 -1，作为一个特殊的环境帧，使其能够在不破坏视频帧自身时间完整性的情况下，为整个序列生成提供条件。鉴于运动编码的是结构而非细粒度纹理，因此我们将运动较短边的长度限制为 256 像素，在保留关键结构信息的同时，大幅降低计算开销。

环境感知训练。 仅以参考图像作为条件，无法让模型理解复杂的场景布局和物体几何结构。为此，我们设计了一套环境感知训练课程，鼓励模型更深入地利用参考图像。训练过程中，我们在两种任务模式之间动态交替：**(1) 条件续写。** 在此模式下，模型接收参考图像 I_{ref}、首帧运动 m_{ref}和文本提示 T 作为条件，其目标是生成后续运动序列 $\{m_t\}_{t=1}^{N-1}$。**(2) 感知即生成。** 在此模式下，首帧运动 m_{ref} 会从输入中屏蔽。模型需要仅以 I_{ref} 和 T为条件，生成包括首帧 m_{ref}在内的完整序列。为进一步强化对场景感知的关注，我们以一定概率将目标视频长度设为 1。这样一来，生成任务便简化为一个类检测目标：模型只需生成单帧，其真实标注为运动 m_{ref}。

这一类似检测的任务也采用相同的流匹配目标进行优化，损失计算仅限于第一帧。该统一框架能够无缝训练 PIM，使其利用一组网络参数处理两个看似迥异、实则密切相关的任务：解析场景结构，以及根据感知生成时间运动。这一策略免去了使用异构损失函数及其相关超参数的需要，避免它们可能引发的训练不稳定。

3.2. 音频交互感知视频生成

音频注入。 要生成逼真的虚拟形象，与输入音频精确实现唇形同步至关重要。我们采用预训练的 Wav2Vec 模型 [1] 作为音频特征提取器。为捕捉语音的协同发音和时间动态，而非依赖孤立音素，我们从每个时间步周围的上下文窗口中提取特征。对于第 i 个视频帧，其对应的音频特征$f_{\text{audio},\ i}$ 通过编码和聚合计算得出：

$$f_{\text{audio},i} = \text{Aggregate}\left(\text{Wav2Vec}(\mathbf{a}_{[c_i - w : c_i + w]})\right), \tag{4}$$

其中$\mathbf{a}$ 表示原始音频，c_i 表示与帧 i 对齐的中心采样点，w 表示上下文窗口的半宽度，而 Aggregate(·) 是一种特征聚合操作，例如拼接，用于实现视频与音频的时间对齐。

为实现帧级音视对齐，提取的音频特征序列 $\{f_{\text{audio},i}\}$ 通过交叉注意力注入 AIM。为加快收敛并明确引导这一条件机制，尤其是在训练早期，我们引入空间人脸掩码 $\mathbf{M}_{\text{face}}$。该掩码在空间上对音频注入模块的输出进行加权，将其影响集中于面部区域，

以实现稳定训练：

$$\mathbf{h}'_{v,l} = \mathbf{h}_{v,l} + \text{CrossAttention}(\mathbf{h}_{v,l}, f_{\text{audio}}) \odot \mathbf{M}_{\text{face}}, \quad (5)$$

其中，AIM 第 l 层的视觉特征表示为$\mathbf{h}_{v,l}$，$\odot$表示逐元素乘法。该策略有效地将音频信号的影响引导至目标面部区域，从而带来更稳定的学习过程。

运动到视频对齐器。利用 PIM 与 AIM 的同构架构，我们引入了用于运动控制的 M2V 对齐器，并采用逐层残差注入机制。具体而言，我们将特征残差（即 PIM 中每个 DiT 块输出之间的差异）注入 AIM 中对应的层。对于 PIM 中给定的第 $l > 0$ 层，其残差 $\mathbf{h}\ \Delta_{m,l}$ 计算如下：

$$\Delta\mathbf{h}_{m,l} = \mathbf{h}^{\text{out}}_{m,l} - \mathbf{h}^{\text{out}}_{m,l-1}. \quad (6)$$

对于基础层（$l = 0$），我们直接使用$\mathbf{h}^{\text{out}}_{m,0}$。

为协调低分辨率运动与高分辨率视频之间的差异，我们首先对残差特征$\mathbf{h}\ \Delta_{m,l}$ 进行上采样，使其与视频特征$\mathbf{h}_{v,l}$ 的空间尺寸相匹配，采用简单的双线性插值。随后，通过一个零初始化线性层进行投影。AIM 块的最终输出计算如下：

$$\mathbf{h}^{\text{out}}_{v,l} = \text{Block}_v(\mathbf{h}^{\text{in}}_{v,l} + \text{Linear}_{\text{zero}}(\text{Interp}(\Delta\mathbf{h}_{m,l}))). \quad (7)$$

这种残差注入策略显著降低了训练过程中的损失波动。此外，线性层的零初始化使模型能够逐步学习运动引导，从而确保训练早期的稳定性。这有效防止了视觉伪影（例如运动骨骼线条的重影）泄漏到最终生成视频中。

生成与驱动控制的统一训练。一个真正通用的数字人系统不仅应能根据文本提示生成动作，还应能忠实遵循明确的驱动信号，例如骨架序列、交互动作或物体轨迹。为此，我们在两个互补任务上共同训练模型：联合生成和外部驱动。在外部驱动任务中，模型接收干净的驱动信号（例如骨架序列），并将其直接输入 PIM。在此，PIM 实际上充当特征编码器，这一设计通过将扩散时间步 t 设为 0 优雅地实现。驱动信号还会被提取为逐层特征残差 $\{\Delta\mathbf{h}_{m,\ l}\}$，随后注入 AIM，以引导生成过程。为平衡这两种能力，我们将联合生成任务与外部驱动任务的训练数据比例设为 4:1。这种联合训练策略使模型在单一统一框架内，既能作为富有创造力的文本引导运动视频生成器，也能作为精准、可控的动画生成器。

3.3. 多模态条件训练策略

我们设计了一个三阶段训练课程，逐步整合多模态条件，确保学习稳定，并避免不同模态之间产生冲突。

面向场景感知规划的 PIM 预训练。初始阶段专门用于独立训练 PIM。其目标是建立核心能力，即从参考图像解析静态场景，并基于文本生成连贯的运动计划。训练课程融合了多项任务：（1）用于构建场景感知的静态检测任务；（2）续写任务，即从给定的首个运动进行运动延续 m_{ref}，以及（3）一项关键的混合生成任务，该任务要求仅根据图像 I_{ref} 和文本生成运动序列。我们发现，这种促使感知与生成协同的混合任务方法优于纯生成流程。这些任务的训练策略如图 2。

面向视听同步的AIM预训练。第二阶段使AIM具备音频驱动的唇形同步能力，同时保持其底层图像到视频（I2V）质量。至关重要的是，我们发现训练顺序极为关键。模型应先使用音频进行训练，再接触交互运动。我们将此归因于条件信号的不同性质：音频是一种通过交叉注意力注入的柔和、局部且异构的模态，而运动则是通过残差注入引入的刚性、全局且同质的结构约束。仅在较弱的局部音频条件稳定之后再引入占主导地位的结构动作条件，可有效防止音频信号在训练优化过程中被抑制。

端到端联合微调。 在最后阶段，对预训练的PIM与AIM进行联合微调，并将PIM的输出以逐层残差的形式注入AIM。训练策略采用任务与条件的概率混合，以统一模型适应文本、音频与运动输入任意组合的能力。具体而言，30%的样本以音频为条件以保持唇形同步能力，15%以真实运动为条件以获得驱动能力，60%用于联合视频-运动生成。保留相当比例的纯图像到视频样本作为关键正则化项，可提升身份保持、视频动态与泛化能力。

4. 实验结果与分析

4.1. 实验细节

我们的 PIM 和 AIM 均基于 wan2.2- 5B [29] 初始化。PIM 使用 3–10 秒的骨架序列和物体轨迹片段（短边为 256px）训练 30,000 步，而 AIM 使用 3–6 秒的视频片段（短边为 704px）训练 5,000 步。在 PIM 和 AIM 的初始预训练阶段，学习率固定

为 1e-5。随后进入为期 4,000 步的联合训练阶段，学习率降至 2e-6。

4.2. 基准构建

为了对人-物体交互生成进行严格评估，我们构建了一个基准测试，名为 GroundInter，其中图像由即梦4.0生成。GroundInter 基准包含 400 张图像，描绘了涉及 1–3 个物体的实际或潜在人-物体交互，涵盖 100 种不同的常见物体。每张图像都附有丰富标注，包括物体类型、1–3 条相应的动作描述（详细说明交互类型、物体及时间片段），以及使用 CosyVoice 合成的匹配对话脚本 [8], ，共形成 600 个测试用例。此外，我们还提供物体分割真值掩码 [17], 检测结果 [40], 和 DW-Pose 关键点 [38]。我们提出了一套聚焦于交互质量的指标：(1) 视觉语言模型问答，我们围绕三个类别（物体、人物、交互）设计了 30 个问题，视觉语言模型 [32] 根据提供的参考图像和视频，为每个问题评定 1 分或 0 分。(2) 语义一致性（CLIP$_{re}$），即原始文本提示与视觉语言模型为生成视频生成的说明文字之间的 CLIP分数 [24] ；(3) 手部质量（HQ），通过手部动态分数与手部拉普拉斯锐度分数的乘积进行评估 [18]；(4) 物体质量（OQ），通过物体动态分数 [42] 与物体DINO一致性的乘积进行评估 [40]；(5) 像素级交互（PI），用于验证 DINO物体框与 DW-Pose关键点之间是否存在接触。我们还采用标准指标对评估进行补充：使用 DINO 衡量人物外观的参考一致性（DINO^{subject}）和参考图像外观的参考一致性（DINO^{ref}）；使用同步置信度（Sync^{conf}）衡量视听一致性 [6], ；使用 CLIP-B-T（CLIP$_{text}$）衡量文本-视频对齐；此外，还采用 VBench 的时间一致性（Temp-C）和动态程度（DD） [39]。更多基准测试详情请参阅补充材料。

4.3. 与最先进方法的比较

由于目前尚无专门针对 GHOI 任务设计的方法，我们将本方法与来自多个相邻领域的最先进（SOTA）方法进行比较。对于音频驱动方法，我们选取混元数字人 [3],Wan-S2V [11], OmniAvatar [10], 和 Fantasy-Talking [31]进行比较。对于主体一致性生成，我们与 HuMo [2]进行对比。对于姿态驱动动画，我们以 UniAnimate-DiT [33]作为基准。VACE [16] 也用于比较。如表 1所示，**音频驱动方法**在参考一致性指标上表现出优势，但在人-物交互方面表现不佳，尤其是在手部质量（HQ）和物体质量（OQ）得分上。其原因在于，这些方法主要关注面部动态，因此只能生成简单、低动态的动作，并且基本避免了交互行为；它们较低的动态度（DD）得分也证实了这一点。这使得它们能够生成简单、低动态的动作，并且基本避免交互行为；其较低的动态度（DD）得分也证实了这一点。虽然这种低动态性有助于保持背景和主体的一致性，但缺乏有意义的动作显然反映在较差的交互得分上。相比之下，我们的 TA2V 变体在交互指标上取得了显著领先，与强基线 Wan-S2V 相比，HQ 提升约 180%，OQ 提升约 111%。重要的是，在 DINO^{subject} 和 DINO^{ref}方面，本方法仍保持了相当的参考图像保真度。在唇形同步性能方面，在 HOI 场景中，我们的方法甚至优于所有音频驱动方法。对于**姿态驱动方法**，UniAnimate-DiT 以我们的 PIM 生成的运动序列作为输入。即使在这种高质量引导下，我们的 TAM2V 变体仍在所有指标上胜过它，尤其在 OQ 得分上取得了显著优势。这种差异源于姿态驱动方法没有考虑物体形态，从而导致交互错位和不自然的物体变形。其固有限制还包括难以获取并将姿态序列与参考图像对齐，以及说话能力有限。对于**主体一致性方法**，HuMo 使用从参考图像中借用的主体生成一段全新视频，而不是延续原始场景，因此在参考一致性方面得分较低。此外，我们的方法在交互质量方面显著超过 HuMo。这是因为主体一致性模型生成的交互往往较为刻板，无法实现从非交互状态到交互状态的平滑过渡。值得注意的是，我们的所有变体共用一个单一模型，仅采用不同的推理模式，凸显了其多功能性和实用价值。

4.4. 定性评估

在图 3中，我们针对具身人-物交互任务，对比了我们的方法与相关领域最先进方法的定性结果。VACE 能够响应文本，但在交互过程中会产生严重伪影，无法保持物体完整性，也不能准确执行交互指令。此外，它还缺少关键的音频驱动能力。Wan-S2V 成功保留了参考图像所指定的交互环境。然而，其动态性主要局限于音频驱动的面部区域，因此难以响应更广泛的交互指令。HuMo 擅长生成细粒度的交互运动细节和面部动态。然而，它无法维持原始交互环境，而是将参考图像中的主体作为外观模板，合成一个完全新的视频，这不适用于 GHOI 任务。相比之下，我们的方法能够准确遵循文本指令，使人物在参考图像环境中与物体进行稳定且合理的交互

表1. **GroundedInter 上的定量比较结果**。使用不同的输入信号时，我们的单一模型可以通过多种模式进行推理。TA2V 表示音频驱动，TAM2V 表示音频与动作驱动，T2MV 则表示使用自生成的运动。

Task	方法	人-物体交互					参考一致性		文本-视频对齐		音频和视频	
		视觉语言模型问答 ↑	HQ ↑	OQ ↑	DD ↑	PI ↑	$DINO_{subject}$ ↑	$DINO_{ref}$ ↑	$CLIP_{text}$ ↑	$CLIP_{re}$ ↑	Temp-C ↑	同步$_{conf}$ ↑
音频驱动	Wan-S2V [11]	24.65	0.336	0.063	0.095	0.619	0.670	**0.870**	0.281	0.885	**0.955**	5.43
	混元数字人 [3]	26.88	0.745	0.112	0.321	0.666	0.661	0.841	0.285	0.873	0.934	5.98
	Fantasy-Talking [31]	25.35	0.416	0.098	0.127	0.593	**0.673**	0.855	0.276	0.889	0.941	5.79
	OminiAvatar [10]	26.76	0.732	0.118	0.305	0.732	**0.639**	0.847	0.282	0.885	0.940	5.95
	我们的方法（TA2V）	**27.32**	**0.931**	**0.133**	**0.765**	**0.803**	0.671	0.857	**0.286**	**0.899**	0.945	**6.04**
姿态驱动	UniAnimate-DiT [33]	24.65	0.862	0.082	0.441	0.669	0.621	0.782	0.277	0.875	0.938	–
	我们的方法（TAM2V）	**28.47**	**0.957**	**0.141**	**0.773**	**0.832**	**0.648**	**0.842**	**0.287**	**0.902**	**0.940**	5.73
文本驱动	VACE [16]	26.74	0.908	0.118	0.719	0.705	0.629	0.817	0.285	0.893	0.940	–
	HuMo [2]	24.12	0.910	0.101	0.380	0.491	0.566	0.726	0.281	0.880	0.935	5.15
	我们的方法 (T2MV)	29.05	**0.975**	**0.150**	**0.792**	**0.852**	0.637	0.835	0.289	**0.904**	0.932	–
	我们的方法 (TA2MV)	**29.07**	0.973	0.147	0.783	0.850	**0.641**	**0.839**	**0.290**	**0.904**	**0.942**	**5.92**



图3. **与最先进方法的定性比较**。混元数字人和Fantasy-Talking的结果请参见补充材料。

。这得益于将感知和规划与视频合成进行显式解耦。PIM 负责规划合理的场景感知交互运动，而 AIM 则负责合成交互中的照片级真实细节。

4.5. 消融实验

环境感知训练的效果。我们进行消融实验，以评估不同的配置

表 2. **环境感知训练的效果**。

Rope	Det.	Det.+Cont.	Last	视觉语言模型问答 ↑	$CLIP_{re}$ ↑	HQ ↑	OQ ↑	PI ↑
				26.36	0.895	0.826	0.118	0.711
			✓	26.71	0.897	0.818	0.115	0.710
✓				27.15	0.902	0.845	0.123	0.730
✓	✓			27.87	0.902	0.917	0.132	0.751
✓	✓	✓		**28.89**	**0.904**	**0.925**	**0.144**	**0.780**

我们的环境感知训练过程，结果汇总于表 2中。默认情况下，参考图像会置于视频序列的开头。 **Last** 表示

表 3. **PIM（上）和多模态训练策略（下）的消融实验。**

方法	视觉语言模型问答 ↑	HQ ↑	OQ ↑	DD ↑	PI ↑	DINO$_{subject}$ ↑	DINO$_{ref}$ ↑	CLIP$_{text}$ ↑	CLIP$_{re}$ ↑	Temp-C ↑	同步$_{conf}$ ↑
不含人物一致性	26.19	0.711	0.104	0.587	0.685	0.638	0.840	0.285	0.892	0.936	5.41
级联	28.19	0.876	0.137	0.730	0.746	0.642	0.841	0.289	0.901	0.940	5.27
最后一层	28.31	0.880	0.141	0.752	0.769	0.643	**0.849**	0.289	0.901	0.938	5.32
加法	28.46	0.903	0.134	0.751	0.777	0.630	0.846	0.286	0.887	0.935	5.14
Ours	**28.89**	**0.925**	**0.144**	**0.761**	**0.780**	**0.646**	0.846	**0.290**	**0.904**	**0.942**	**5.43**
AM	28.23	0.912	0.133	0.745	0.770	0.639	0.838	0.289	0.900	0.933	4.23
$M \rightarrow AM$	28.84	0.920	0.141	**0.764**	0.778	0.636	0.837	**0.290**	0.903	0.934	3.98
$A \rightarrow AM$	28.02	0.906	0.134	0.738	0.755	0.640	0.842	0.287	0.897	0.940	**5.49**
$AI \rightarrow IAM$ (我们的方法)	**28.89**	**0.925**	**0.144**	0.761	**0.780**	**0.646**	**0.846**	**0.290**	**0.904**	**0.942**	5.43



图 4. **PIM效应的可视化。**

参考图像会附加在最后一帧之后。**RoPE**spec- 表示是否对参考图像应用修改后的旋转位置编码（RoPE）。**Det.** 表示在 PIM预训练期间使用检测数据，该过程会基于参考图像生成单个运动帧。**Det.+Cont.** 表示在没有参考运动的情况下生成完整的运动序列。我们的研究结果表明，与将参考图像附加在序列末尾相比，将其置于开头能带来更优的性能，这归因于参考图像与运动帧之间的结构相似性。然而，要发挥这一优势，必须修改 RoPE，以减轻参考图像对生成帧的过度影响。此外，我们发现，检测数据对于建立环境感知能力至关重要，而 Det.+Cont. 数据则是连接感知与生成的重要桥梁。这两类数据均有助于 PIM 的训练。

PIM 的效果。我们的 PIM 与 AIM 采用同源架构，因此其特征表示具有内在对齐性，我们借此实现有效的信息融合。如表 3（上）所示，我们对注入交互运动特征进行了消融实验。w/o PIM 指将单个 AIM 训练为由文本、图像和音频驱动的视频生成器，其文本对齐和人-物交互质量相对较差。这导致模型缺乏环境感知能力，并加剧了 GHOI 中的控制质量困境。采用 PIM 后，我们发现，统一的联合生成显著优于级联流程；在级联流程中，单独训练的（冻结的）PIM 会生成运动序列来驱动 AIM，

表 4. **运动表示消融实验。**

方法	视觉语言模型问答 ↑	高质量 ↑	整体质量 ↑	动态度 ↑	PI ↑	DINO$_{sub}$ ↑	同步$_{conf}$ ↑
二维 (x,y) 表示	26.07	0.705	0.094	0.698	0.655	0.623	4.72
二维 (x,y) 级联	25.58	0.884	0.121	0.659	0.672	0.625	5.04
Ours	**28.89**	**0.925**	**0.144**	**0.761**	**0.780**	**0.646**	**5.43**

级联方法造成了信息瓶颈。它依赖显式运动条件，因此过于受限，无法捕捉物体形状变化，进而导致伪影并降低生成手部的可信度。相比之下，联合训练能够实现更丰富的信息流。在联合训练范式下，仅将 PIM 的最后一层特征注入所有 AIM 块层，其效果与级联模型相当，这表明全局应用运动不如自适应的逐层引导有效。我们采用的逐层残差注入机制通过零初始化线性层进行增强，其效果也优于简单的逐层加法，证实了我们设计的优势。

如图 4所示，我们展示了 PIM 的效果。使用冻结的PIM时，级联变体可视为姿态驱动模型，而不使用 PIM 则对应于 TIA2V模型。姿态驱动变体能够生成文本对齐交互视频，但容易出现异常物体变形。TIA2V 变体通常难以理解语义，导致无法准确执行所请求的交互；相比之下，我们的方法能够成功生成合理且与文本对齐的交互视频。

运动表示。在表4 （上）中，我们比较

表5. **真实场景对比。**

方法	视觉语言模型问答 ↑	HQ ↑	OQ ↑	PI ↑	DINO_{sub} ↑	CLIP_{text} ↑	同步$_{conf}$ ↑
Wan-S2V	24.97	0.851	0.115	0.740	0.646	0.285	5.36
HY-Avatar	25.29	0.817	0.108	0.724	0.644	0.285	5.45
HuMo	26.23	0.840	0.116	0.781	0.635	0.288	5.33
Ours	**28.49**	**0.910**	**0.135**	**0.794**	**0.665**	**0.289**	**5.51**

我们的RGB运动表征与更简单的二维坐标表示。坐标运动生成模型缺乏先验和检测能力（生成运动需要初始框和关节位置），且泛化较差。它与视频之间巨大的模态差距阻碍了对齐（大多数驱动方法使用RGB），而级联的“先生成(x,y)，再渲染RGB以供驱动”明显不及联合生成。RGB运动利用视频生成先验，在无需大规模数据集的情况下提升了文本到运动准确率与泛化能力。共享RGB空间简化了视频-运动对齐并降低了训练复杂度，从而为GHOI生成带来更好的结果。

真实场景评估。 为了更好地与最先进方法进行比较，我们从真实场景中收集了50个案例，以构建一个额外的测试集。在表格中给出了定量实验结果。5 结果表明，我们的方法在真实场景中表现良好。

多模态训练策略。 本节探讨AIM训练期间多模态条件排序的最优策略。如表格 3 （底部） $A \rightarrow AM$ 所示，该策略首先在音频上进行训练，随后在音频和运动上进行联合训练。 AM 表示从一开始就同时训练两种模态，而不进行任何AIM音频预训练。 $A \rightarrow AM$ 策略在音频指标上的表现显著优于从一开始进行联合训练（AM）或先对运动模态进行预训练（$M \rightarrow AM$。这一结果源于条件信号的不对称性。与Omnihuman的结论一致， [19], 音频相较于交互运动是较弱的信号，需要专门的预训练阶段才能实现有效学习；否则，其影响会被占主导地位的运动信号所掩盖。此外，我们发现，使用文本-图像到视频数据（不含配对音频）丰富音频驱动预训练阶段，能够带来显著收益。这可确保模型保留通用的运动动态先验，为模型在最终联合训练阶段学习复杂的交互运动提供更坚实的基础。

用户研究。 我们针对每种方法的 50 个生成样本，对 20 名参与者开展了用户研究。给定由不同方法生成的两段视频，参与者需选择更好的一段。结果见图5。

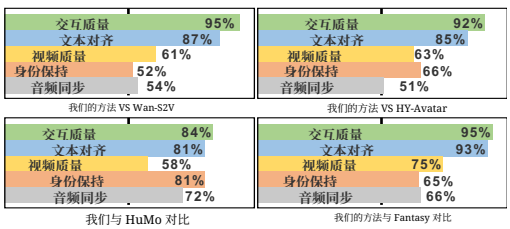


图 5。 **用户研究结果。**

5. 结论

本文提出了一种双流说话虚拟人视频生成框架，能够生成基于场景的人物与物体交互。我们的方法将感知和规划与视频合成解耦，从而生成合理、可控且高质量的 GHOI 视频。PIM 经过精心设计，用于感知参考图像的上下文并规划文本对齐的交互动作，而 AIM 则借助 M2V 对齐器生成生动的说话与交互虚拟形象。我们提出了一个基准测试，用于评估 GHOI 生成。全面的实验展示了我们模型的性能，并凸显了各模块的贡献与影响。我们的工作为该领域的未来研究提供了有价值的见解。我们工作的主要局限在于，它只能处理单人场景，无法生成涉及多人的人-物交互。

参考文献

[1] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：用于语音表示自监督学习的框架。《神经信息处理系统进展》，33：12449–12460，2020年。4[2] Liyang Chen、Tianxiang Ma、Jiawei Liu、Bingchuan Li、Zhuowei Chen、Lijie Liu、Xu He、Gen Li、Qian He 和 Zhiyong Wu。HuMo：通过协同多模态条件控制实现以人为中心的视频生成，2025年。2、3、6、7[3] Yi Chen、Sen Liang、Zixiang Zhou、Ziyao Huang、Yifeng Ma、Junshu Tang、Qin Lin、Yuan Zhou 和 Qinglin Lu。混元视频数字人：面向多角色的高保真音频驱动人体动画。arXiv预印本 arXiv:2505.20156，2025年。2、3、6、7[4] Zhiyuan Chen、Jiajiong Cao、Zhiquan Chen、Yuming Li 和 Chenguang Ma。EchoMimic：通过可编辑关键点条件实现栩栩如生的音频驱动人像动画。载于《AAAI人工智能会议论文集》，第2403–2410页，2025年。3[5] Gang Cheng、Xin Gao、Li Hu、Siqi Hu、Mingyang Huang、Chaonan Ji、Ju Li、Dechao Meng、Jinwei Qi、Penchong Qiao 等。Wan-Animate：通过整体复刻实现统一的角色动画与替换。arXiv预印本 arXiv:2509.14055，2025年。2[6] J. S. Chung 和 A. Zisserman。时光之外：野外环境中的自动唇形同步。载于多视角唇读研讨会，ACCV，2016年。6[7] Jiahao Cui、Hui Li、Yun Zhan、Hanlin Shang、Kaihui Cheng、Yuqi Ma、Shan Mu、Hang Zhou、Jingdong Wang 和 Siyu Zhu。Hallo3：借助视频扩散Transformer实现高度动态且逼真的人像图像动画。载于《计算机视觉与模式识别会议论文集》，第21086–21095页，2025年。2、3[8]Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等。CosyVoice 2：利用大语言模型实现可扩展的流式语音合成。arXiv预印本 arXiv:2412.10117，2024年。6

[9] Yingying Fan、Quanwei Yang、Kaisiyuan Wang、Hang Zhou、Yingying Li、Haocheng Feng、Errui Ding、Yu Wu 和 Jingdong Wang。Re-hold：通过自适应布局引导扩散模型实现视频中的手-物交互重演。载于《计算机视觉与模式识别会议论文集》，第17550–17560页，2025年。3[10] Qijun Gan、Ruizi Yang、Jianke Zhu、Shaofei Xue 和 Steven Hoi。OmniAvatar：通过自适应身体动画实现高效的音频驱动虚拟形象视频生成。arXiv预印本 arXiv:2506.18866，2025年。6、7[11] Xin Gao、Li Hu、Siqi Hu、Mingyang Huang、Chaonan Ji、Dechao Meng、Jinwei Qi、Penchong Qiao、Zhen Shen、Yafei Song 等。Wan-S2V：音频驱动的电影级视频生成。arXiv预印本 arXiv:2508.18621，2025年。6、7[12] Li Hu、Guangyuan Wang、Zhen Shen、Xin Gao、Dechao Meng、Lian Zhuo、Peng Zhang、Bang Zhang 和 Liefeng Bo。Animate Anyone 2：结合环境可供性的高保真角色图像动画。arXiv预印本 arXiv:2502.06145，2025年。3[13] Teng Hu、Zhentao Yu、Zhengguang Zhou、Sen Liang、Yuan Zhou、Qin Lin 和 Qinglin Lu。HunyuanCustom：用于定制化视频生成的多模态驱动架构，2025年。2、3[14] Ziyao Huang、Zixiang Zhou、Juan Cao、Yifeng Ma、Yi Chen、Zejing Rao、Zhiyong Xu、Hong mei Wang、Qin Lin、Yuan Zhou 等。HunyuanVideo-HOMA：多模态驱动人体动画中的通用人-物交互。arXiv预印本 arXiv:2506.08797，2025年。3[15] Ji anwen Jiang、Chao Liang、Jiaqi Yang、Gaojie Lin、Tianyun Zhong 和 Yanbo Zheng。Loopy：借助长期运动依赖驾驭音频驱动的肖像虚拟形象。arXiv预印本 arXiv:2409.02634，2024年。3[16] Zeyinzi Jiang、Zhen Han、Chaojie Mao、Jingfeng Zhang、Yulin Pan 和 Yu Liu。VACE：一站式视频创作与编辑。arXiv预印本 arXiv:2503.07598，2025年。6、7[17] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C. Berg、Wan-Yen Lo、Piotr Doll´ar 和 Ross Girshick。Segment Anything。arXiv:2304.02643，2023年。6[18] Gaojie Lin、Jianwen Jiang、Chao Liang、Tianyun Zhong、Jiaqi Yang 和 Yanbo Zheng。Cyberhost：利用区域码本注意力驾驭音频驱动的虚拟形象扩散模型。arXiv预印本 arXiv:2409.01876，2024年。6[19] Gaojie Lin、Jianwen Jiang、Jiaqi Yang、Zerong Zheng 和 Chao Liang。OmniHuman-1：重新思考单阶段条件人体动画模型的规模化。arXiv预印本 arXiv:2502.01061，2025年。2、3、9[20] Lijie Liu、Tianxiang Ma、Bingchuan Li、Zhuowei Chen、Jiawei Liu、Gen Li、Siyu Zhou、Qian He 和 Xinglong Wu。Phantom：通过跨模态对齐实现主体一致性视频生成。arXiv预印本 arXiv:2502.11079，2025年。2、3[21] Yifang Men、Yuan Yao、Miaomiao Cui 和 Liefeng Bo。MIMO：通过空间分解建模实现可控角色视频合成。载于《计算机视觉与模式识别（CVPR）》，2025年IEEE会议，2025年。3[22] Rang Meng、Xingyu Zhang、Yuming Li 和 Chenguang Ma。Echomimicv2：迈向引人注目、简化且半身的人体动画。载于《计算机视觉与模式识别会议论文集》，第5489–5498页，2025年。2 – 。[23] Youxin Pang、Ruizhi Shao、Jiajun Zhang、Hanzhang Tu、Yun Liu、Boyao Zhou、Hongwen Zhang 和 Yebin Liu。Manivideo：通过灵巧且具泛化能力的抓握生成手物体操作视频。载于《计算机视觉与模式识别会议论文集》，第12209–12219页，2025年。3[24] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于《国际机器学习会议》，第8748–8763页。PmLR，2021年。6[25] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。利用统一的文本到文本转换器探索迁移学习的极限。《机器学习研究杂志》，21(140):1–67，2020年。3[26] Shuai Tan、Biao Gong、Xiang Wang、Shiwei Zhang、Dandan Zheng、Ruobing Zheng、Kecheng Zheng、Jingdong Chen 和 Ming Yang。Animate-x：具有增强运动表示的通用角色图像动画。arXiv预印本 arXiv:2410.10306，2024年。2[27] Linrui Tian、Qi Wang、Bang Zhang 和 Liefeng Bo。Emo：在弱条件下利用音频转视频扩散模型生成富有表现力的肖像视频，使肖像栩栩如生。载于《欧洲计算机视觉会议》，第244–260页。施普林格，2024年。2[28] Shuyuan Tu、Zhen Xing、Xintong Han、Zhi-Qi Cheng、Qi Dai、Chong Luo 和 Zuxuan Wu。Stableanimator：高质量且保持身份一致性的人体图像动画。载于《计算机视觉与模式识别会议论文集》，第21096–21106页，2025年。2[29] Team Wan、Ang Wang、Baole Ai、Bin Wen、Chaojie Mao、Chen-Wei Xie、Di Chen、Feiwu Yu、Haiming Zhao、Jianxiao Yang、Jianyuan Zeng、Jiayu Wang、Jingfeng Zhang、Jingren Zhou、Jinkai Wang、Jixuan Chen、Kai Zhu、Kang Zhao、Keyu Yan、Lianghua Huang、Mengyang Feng、Ningyi Zhang、Pandeng Li、Pingyu Wu、Ruihang Chu、Ruili Feng、Shiwei Zhang、Siyang Sun、Tao Fang、Tianxing Wang、Tianyi Gui、Tingyu Weng、Tong Shen、Wei Lin、Wei Wang、Wei Wang、Wenmeng Zhou、Wente Wang、Wenting Shen、Wenyuan Yu、Xianzhong Shi、Xiaoming Huang、Xin Xu、Yan Kou、Yangyu Lv、Yifei Li、Yijing Liu、Yiming Wang、Yingya Zhang、Yitong Huang、Yong Li、You Wu、Yu Liu、Yulin Pan、Yun Zheng、Yuntao Hong、Yupeng Shi、Yutong Feng、Zeyinzi Jiang、Zhen Han、Zhi-Fan Wu 和 Ziyu Liu。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。3，5[30] Cong Wang、Kuan Tian、Jun Zhang、Yonghang Guan、Feng Luo、Fei Shen、Zhiwei Jiang、Qing Gu、Xiao Han 和 Wei Yang。V-express：用于肖像视频生成渐进式训练的条件丢弃。arXiv预印本 arXiv:2406.02511，2024年。2

[31] Mengchao Wang、Qiang Wang、Fan Jiang、Yaqi Fan、Yunpeng Zhang、Yonggang Qi、Kun Zhao 和 Mu Xu。Fantastytalking：通过连贯运动合成实现逼真的说话肖像生成。arXiv预印本 arXiv:2504.04842，2025年。2，3，6，7[32] Weiyun Wang、Zhangwei Gao、Lixin Gu、Hengjun Pu、Long Cui、Xingguang Wei、Zhaoyang Liu、Linglin Jing、Shenglong Ye、Jie Shao 等。InternVL3.5：在多功能性、推理和效率方面推进开源多模态模型。arXiv预印本 arXiv:2508.18265，2025年。6[33]Xiang Wang、Shiwei Zhang、Longxiang Tang、Yingya Zhang、Changxin Gao、Yuehuan Wang 和 Nong Sang。UniAnimate-DiT：利用大规模视频扩散Transformer实现人物图像动画。arXiv预印本 arXiv:2504.11289，2025年。2，6，7[34] Zhenzhi Wang、Jiaqi Yang、Jianwen Jiang、Chao Liang、Gaojie Lin、Zerong Zheng、Ceyuan Yang 和 Dahua Lin。InteractHuman：具有布局对齐音频条件的多概念人物动画。arXiv预印本 arXiv:2506 09984，2025年。3。 。[35] Mingwang Xu、Hui Li、Qingkun Su、Hanlin Shang、Liwei Zhang、Ce Liu、Jingdong Wang、Yao Yao 和 Siyu Zhu。Hallo：用于人像图像动画的分层音频驱动视觉合成。arXiv预印本 arXiv:2406.08801，2024年。2，3[36] Ziyi Xu、Ziyao Huang、Juan Cao、Yong Zhang、Xiaodong Cun、Qing Shuai、Yuchen Wang、Linchao Bao、Jintao Li 和 Fan Tang。AnchorCrafter：通过人-物交互视频生成，让网络主播为你的产品代言。arXiv预印本 arXiv:2411.17383，2024年。3[37] Zihui Sherry Xue、Romy Luo、Changan Chen 和 Kristen Grauman。HOI-Swap：利用手物体交互感知在视频中交换物体。《神经信息处理系统进展》，37:77132–77164，2024年。3[38] Zhendong Yang、Ailing Zeng、Chun Yuan 和 Yu Li。采用两阶段蒸馏的高效全身姿态估计。载于《IEEE/CVF国际计算机视觉会议论文集》，第4210–4220页，2023年。6[39] Fan Zhang、Shulin Tian、Ziqi Huang、Yu Qiao 和 Ziwei Liu。Evaluation Agent：面向视觉生成模型的高效且可提示评估框架。arXiv预印本 arXiv:2412.09645，2024年。6[40] Hao Zhang、Feng Li、Shilong Liu、Lei Zhang、Hang Su、Jun Zhu、Lionel M. Ni 和 Heung-Yeung Shum。DINO：采用改进去噪锚框的DETR，实现端到端目标检测，2022年。6[41] Wenxuan Zhang、Xiaodong Cun、Xuan Wang、Yong Zhang、Xi Shen、Yu Guo、Ying Shan 和 Fei Wang。SadTalker：学习逼真的三维运动系数，用于风格化音频驱动单图像说话人脸动画。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8652–8661页，2023年。2[42] Youliang Zhang、Zhaoyang Li、Duomin Wang、Jiahe Zhang、Deyu Zhou、Zixin Yin、Xili Dai、Gang Yu 和 Xiu Li。SpeakerVid-5M：用于音视频双人交互式人物生成的大规模高质量数据集。arXiv预印本 arXiv:2507.09862，2025年。6