IntentVCNet：弥合空间-时间差距的意图导向可控视频字幕生成

Tianheng Qiu*
University of Science and Technology of China
Hefei, China
Hefei Institutes of Physical Science,Chinese Academy of Sciences
Hefei, China
thqiu.cs@mail.ustc.edu.cn

Jingchun Gao*
University of Science and Technology of China
Hefei, China
gaojc0714@mail.ustc.edu.cn

Jingyu Li†
Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
Hefei, China
State Key Lab. for Novel Software Technology, Nanjing University
Nanjing, China
jingyuli@iai.ustc.edu.cn

Huiyi Leong
University of Chicago
Chicago, America
Joyce.yong@uchicago.edu

Xuan Huang
Hefei Institutes of Physical Science,Chinese Academy of Sciences
Hefei, China
huangxuan@iim.ac.cn

Xi Wang
National University of Defense Technology
Changsha, China
wx_23ndt@nudt.edu.cn

Xiaocheng Zhang
Harbin Institute of Technology
Harbin, China
22s136029@stu.hit.edu.cn

Kele Xu
National University of Defense Technology
Changsha, China
kele.xu@ieee.org

Lan Zhang
University of Science and Technology of China
Hefei, China
zhanglan@ustc.edu.cn

Abstract

面向意图的受控视频字幕生成旨在基于定制的用户意图为视频中特定目标生成有针对性的描述。目前的大型视觉语言模型（LVLMs）已获得了强大的指令跟随和视觉理解能力。虽然LVLMs分别在空间和时间理解方面表现出较高的能力，但它们无法在时间序列中直接响应指令进行细粒度的空间控制。这种显著的时空差距使得在视频中实现细粒度的意图导向控制变得复杂。为此，我们提出了一种新的IntentVCNet，将LVLMs固有的时间和空间理解知识统一起来，从提示和模型的角度弥合时空差距。具体而言，我们首先提出了一种提示组合策略，旨在使LLM能够建模表征用户意图和视频序列之间的隐式关系。然后，我们提出了一种参数高效的框适配器，增强了全局视觉上下文中的对象语义信息，使视觉标记事先具有关于用户意图的信息。最终实验证明，这两种策略的结合可以进一步增强LVLM建模视频序列中的空间细节的能力，并促进LVLMs准确生成受控的意图导向字幕。我们提出的方法在多个开源LVLM中取得了最先进的结果，并在IntentVC挑战中获得了亚军。我们的代码可以在https://github.com/thqiu0419/IntentVCNet上获取。

*Equal contribution. †Corresponding author.

MM '25, October 27-October 31, 2025, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Natural language generation**.

Keywords

Intention-Oriented Controllable Video Captioning, Spatial Representation, Large Video-Language Model, Ensemble Learning

1 引言

视频描述生成旨在自动生成给定视频的描述，由于其在增强跨越空间和时间维度的视觉理解方面的潜力，吸引了大量关注。如图1所示，传统的视频描述生成优先考虑描述的准确性和通用性，更加侧重于视频的整体理解，并难以关注用户感兴趣的对象，这使得传统的视频描述生成在个性化、高度可访问的场景中表现不佳。因此，引入旨在意图导向的可控描述生成具有重要意义，这使得生成与意图导向对象一致的定制描述成为可能，并促进更个性化的人机交互体验。

面向意图的可控视频字幕生成需要在动态视频流中跟踪目标对象，这对理解每个静态帧中的区域对象及其相应的时间动作提出了挑战。最近的研究见证了其在大视觉语言模型（LVLM）方面的重要发展。LVLMs [13, 20, 55] 将大型语言模型 [9, 45] (LLM) 的知识扩展到视觉领域，展示了在各种图像级任务（包括图像字幕生成）上的显著性能。后续工作深入探讨了空间和时间维度的更细致理解。在空间维度上，研究 [5, 25, 42, 50, 54] 将显式位置信息集成到LVLMs中以支持区域任务，例如视觉指向。它们设计了各种位置参考方法以增强细粒度区域理解。在时间维度上，[1, 39, 51] 采用视频指令微调使模型适应视频格式并有效建模时间关系，在视频字幕生成方面表现优异。鉴于LVLMs的有限上下文长度，它们还探索在帧序列中压缩冗余视觉标记。

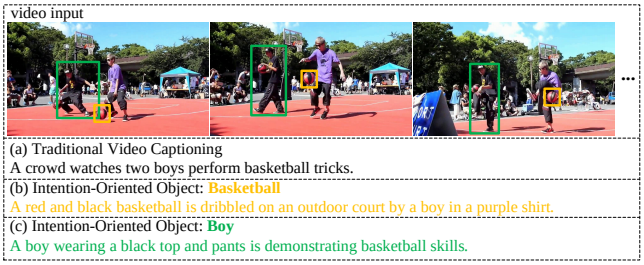

Figure 1: 传统视频字幕任务与意图导向可控视频字幕任务之间的比较。(a) 传统视频字幕提供了内容的总体概述，但缺乏解决特定用户需求的具体性。(b，c) 意图导向可控视频字幕强调用户感兴趣的对象，同时考虑上下文信息，从而生成更详细和有针对性的字幕。

尽管 LVLMs 在空间理解和视频字幕生成方面分别表现出良好的前景，但在跨帧序列跟踪细粒度对象时仍存在时空差距。这一限制妨碍了 LVLMs 在意图导向的可控视频字幕生成中的细粒度可控性。这个问题的产生是因为目前的 LVLMs 是通过在简单的视频级指令数据集上的预训练来获得时间建模能力，而通过在静态图像上的预训练来发展空间理解能力。在静态空间理解和动态时间建模之间存在需要弥合的时空差距。CAT-V [33] 将 LVLM 与其他在对象识别和时间分析方面的专家整合，以促进以对象为中心的字幕生成。然而，CAT-V 是一个无需训练的框架，因此，其性能受到各种专家模块有效性的限制。此外，在 CAT-V 中，LVLM 仅作为一个基本的字幕生成器，未能解决时空差距。因此，目前的 LVLMs 在理解与特定对象相关的更细粒度的时间变化方面仍然面临困难。

为了解决时空差距问题，我们提出了 IntentVCNet，即一种时空增强的多模态协作框架。我们通过改进提示学习技术和模型架构，大大提高了 LVLMs 的细粒度空间理解能力。一方面，我们通过提示组合来增强 LLM 中细粒度对象的空间建模，而不是仅使用单一位置表示 [5, 22, 25] 。另一方面，我们在视觉编码器中开发了一种全局-局部交互模块，以有效提取区域增强的视觉特征。此外，我们进行参数高效的视频指令调整，以保留固有的视觉-语言知识，并提高 LVLM 对视频中意图导向对象动态变化的理解能力。最终，我们通过协作投票机制整合这些模型的结果以提高整体性能。具体来说，对于提示组合，我们将语言指令中的数字坐标序列和视频中的视觉提示进行融合，这增强了从视觉和语言域的细粒度对象定位能力，并获取各种异构模型。对象的数字坐标在指令中按每帧进行规范化。对于视觉提示，意图导向对象在每帧中用红色框突出显示。在模型层面上，我们采用稳健的 InternVL3 [57] 和 InternVideo2.5 [40] 作为我们的基础模型。InternVL3 有助于处理高分辨率视频，从而确保每帧视觉信息的完整保留。相反，InternVideo2.5 实施高效的视觉语义压缩以减少冗余标记，从而增强其对较长视频理解的适应性。为了增强意图导向对象与帧图像之间的空间互动，我们提出了一种盒子适配器，该适配器结合了全局-局部互动模块。这些模块有助于将对象语义融合到帧的全局特征中。最后，为了实现协同结果，我们基于多个异构模型生成的描述的文本相似性实施协作投票过程。我们的贡献总结如下：

我们提出了一种提示组合方法，该方法融合了指令和视频数据中有效的位置指代，提升了大型语言模型的空间建模能力，以识别意图导向的对象。

我们提出了一种参数高效的边框适配器，以增强意图导向对象和帧图像之间的空间交互，从而获取区域增强的视觉特征。

我们在 IntentVC 基准上进行广泛的实验，并在测试集上获得 225.19 % CIDEr 分数，取得了杰出的表现，在与 ACM MM '25 联合举办的 IntentVC 挑战赛中排名第二。

2 相关工作

视频字幕生成（VC）取得了显著的进步，从早期复杂的神经网络架构发展到大型视觉语言模型。这些工作利用了编码器-解码器框架，其中视觉编码器（CNNs/ViTs）提取视觉特征，文本解码器（RNNs/Transformers）生成字幕。早期的努力使用了注意力机制 [7, 14, 19] 、图网络 [41, 52] 和强化学习 [21, 24, 38] 。随着预训练技术的兴起，后续工作 [31, 34, 44, 49] 遵循“预训练-微调”范式。预训练模型可以通过微调来适应各种下游任务，包括视频字幕生成。最近，LVLMs 快速发展。许多工作也探索了在视频理解中使用 LVLMs，以获得能够执行多种任务的多功能模型。他们持续优化时空交互 [4, 8, 18, 21] 和训练策略 [27, 43, 51] ，以增强基础 LVLMs 的时间建模能力。InternVL [40, 57] 和 QwenVL [1] 代表了视频理解领域，特别是视频字幕生成中的尖端模型。

随着人机交互系统能力的提升，对既具描述性又能针对特定用户意图的字幕的需求也在增长。这一演变催生了可控视频字幕生成。可控信号主要可以分为两类：结构控制和内容控制。前者调节生成句子的语法结构 [32, 36] ，而后者限制内容，包括对象 [47, 56] 、关系 [3] 和情感方面 [28–30, 48] 。对于面向对象的控制，OVC-Net [56] 提出了一个时间图以强调特定对象。Elysium [37] 和 GroundingGPT [17] 构建了对象级指令数据集，并在基础任务上取得了良好性能。然而，由于训练数据的稀缺和模型适应性不足带来的时空差距，尚无法充分利用它们进行面向对象的可控视频字幕。

2.1 LVLM 中的空间理解

为了通过 LVLMs 增强对视觉世界的空间理解，现有文献中提出了各种位置表示方法。Kosmos [25] 首次通过使用专门的位置标记来表示区域，引入了一种统一的位置表示方法。Shikra [5] 进一步简化了早期的方法，直接使用数字坐标进行表示。GPT4RoI [54] 从特征角度增加了交互中对象级区域特征的重要性。Ferret [50] 整合了先前的表示方法，并引入了一种混合空间表示方法，结合了三元组，包括区域名称、数字坐标和区域特征来定义一个区域。一个区域由一个四维坐标系定义，由左上角和右下角点表示。先前描述的方法将位置表示纳入语言指令中。然而，在 LVLMs 的当前范式下，这种方法消耗了可用上下文长度的很大一部分，可能导致窗口溢出和模型性能下降。此外，[42] 显示 LVLMs 中的视觉编码器对视觉标记特别敏感。因此，这些特殊标记 [46] 也可以作为视觉提示，而不会增加上下文中的位置标记长度。

3 方法

我们提出的模型如图 2 所示。从提示的角度来看，我们首先设计了一种提示组合方法，其中将语言指令中的数值坐标和视频中的视觉提示相结合，从而增强 LLM 的细粒度对象定位，并获取各种异构模型。在视觉提示方面，感兴趣的目标对象在每一帧中被显著地用红色方框突出显示。从模型的角度来看，为了增强目标对象与帧图像之间的空间交互，我们提出了一种通过交叉注意模块结合的方框适配器。这些模块能够将对象

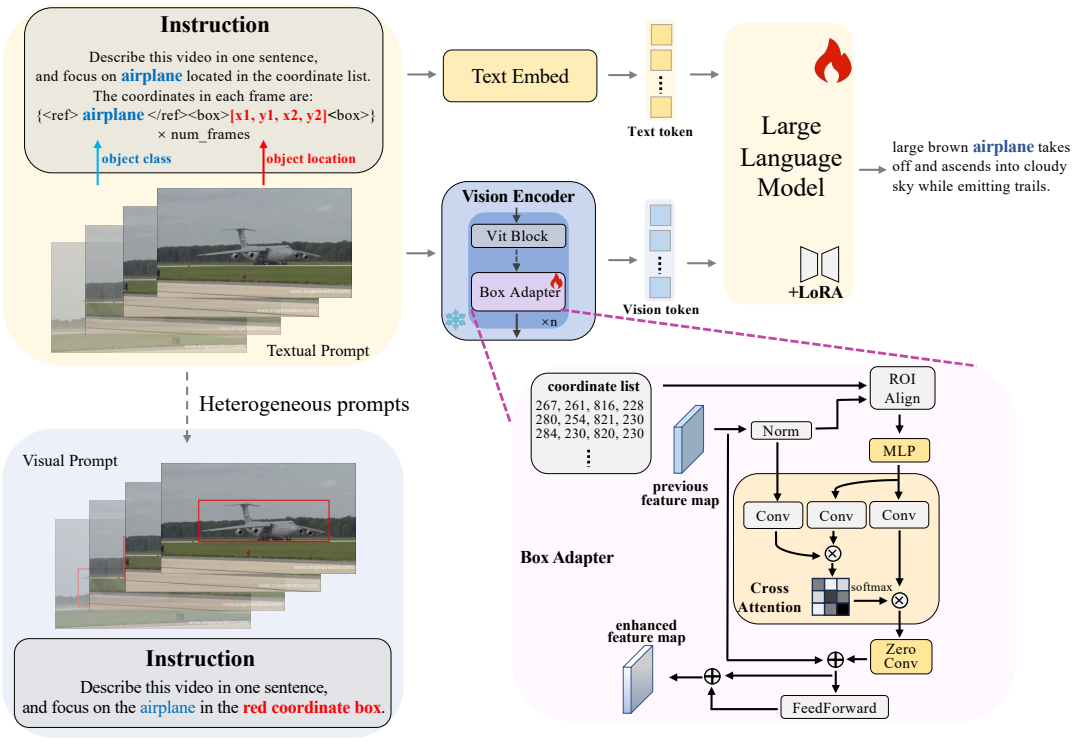

Figure 2: 我们的意图导向视频字幕框架概述。1）我们首先设计了一个提示组合，其中合并了语言坐标和视觉提示。2）在视觉编码器中，我们插入了框适配器，以通过全局-局部互动增强区域的视觉信息。3）最后，原始的视觉编码器被冻结，仅优化轻量级框适配器。此外，LLM 使用 LoRA [12] 进行训练。

语义整合到帧的全局特征中。最后，我们引入了一种多模型协作策略，旨在整合不同长度视频的各种模型。

3.1 预训练的大型视觉-语言模型

大型视觉语言模型是基于大语言模型开发的，并通过广泛的视频指令数据持续预训练，展示了出色的视频理解和指令跟随能力。在本文中，我们利用 InternVL3 [57] 和 InternVideo2.5 [40] 来分析不同长度的视频。

InternVL3 由三个模块组成：一个视觉编码器、一个多模态连接器和一个 LLM。输入的视频帧最初被分割成图像块。随后，采用固定分辨率的视觉编码器提取它们的视觉特征，从而支持动态高分辨率，以最大限度地保留视觉信息。多模态连接器由一个 MLP 层和像素解组操作构成，将视觉内容投射到 LLM 的表示空间，并简化视觉嵌入。这些视觉特征随后被定位在嵌入式语言指令的指定槽位中，共同形成 LLM 的上下文嵌入。

InternVideo2.5。在 InternVL 基础模型的基础上，InternVideo2.5 通过对长视频数据的后训练进行改进。InternVideo2.5 还实现了基于视觉特征语义相似性的层级视觉标记压缩，使模型能够在有限的上下文长度中纳入更多的视频帧，从而实现长距离的视频建模。此外，在训练策略方面，InternVideo2.5 采用直接偏好优化来增强高密度视觉任务。

3.2 提示组合

以往的研究 [5, 25, 50, 54] 在指令中采用了各种位置参考方法，以帮助模型理解特定区域。在本文中，我们提出了一种分别在用户指令和视觉输入中使用提示组合的方法。通过设计组合位置提示，LLM 获得了细粒度的空间建模能力，使其能够扩展到各种异构模型。具体而言，提示组合在指令中包含了数值坐标和视觉提示。

① 指令中的数值坐标。LVLMs 通过用户指令提供可控性，这些指令包含用户的意图，使其对于意图导向的视频描述非常重要。本文中，我们的可控元素是特定的对象，而视频数据中的对象不断移动和变化。因此，简单的文本指令无法充分作为意图导向对象的参考。我们将数值坐标的方法从静态图像的空间理解扩展到动态图像。具体来说，我们将每帧中感兴趣对象区域的坐标映射到各自帧的文本格式中。这些坐标表示为四维向量，具体指示左上角和右下角位置的水平和垂直坐标，表示为 $[x_1, y_1, x_2, y_2]$ 。为了标准化不同大小，这些值被规范化到 0 到 1000 的范围内，以及生成的用户指令。

② 视觉提示。[42, 46] 已表明，LVLMs 的视觉编码器对特定的显著视觉标记特别敏感。因此，后续研究尝试通过在图像中加入视觉标记来突出显示意图参考区域。这些标记作为视觉提示，也可以有效地扩展到视频数据。我们将意图导向对象的坐标可视化到相应的视频帧上。如图 2 所示，红色矩形区域表示我们对这些坐标的可视化结果。需要注意的是，与原始坐标大小相比，我们稍微扩大了边界框的范围，以尽量减少红色框内目标对象的过度遮挡。

当前的 LVLMs 旨在增强空间理解，但在与特定区域的交互方面表现不足。它们通过预训练获取了广泛的多模态知识，这些知识嵌入在它们的参数中。因此，直接改变模型结构以改善细粒度的区域交互可能会危及内在知识。之前的工作引入了

参数高效微调（PEFT）方法，例如前缀微调 [15] 、适配器微调 [11] 和 LoRA [12] ，这些方法固定了原始 LVLMs 并插入了少量可训练的新参数，从而在保存预训练模型所获得知识的同时，促进了模型的微调。受这些 PEFT 方法的启发，我们提出了盒子适配器，该适配器被集成到原始的 LVLMs 中，以增强与意图导向对象的更深入交互。

具体而言，如图 2 所示，给定第 i 帧的视觉特征图 $V_f = \{v_{fi} \in \mathbb{R}^{d\times h\times w}\}_{i=1}^{N_v}$ ，一个框适配器首先通过兴趣区域（RoI）对齐提取意图导向对象的区域特征，可以表示为：

$$R = \text{RoI\_Align}\left(\text{LN}\left(V_f\right), bbox\right), \tag{1}$$

，其中 $bbox$ 是意图导向对象的数值坐标，$R \in \mathbb{R}^{N_v\times d\times h'\times w'}$ 表示其区域特征。然后，我们通过交叉注意模块执行全局-局部交互。完整的视觉特征图 V_f 用作查询嵌入，而区域特征则作为键-值嵌入。这种设计将区域视觉信息注入整体视觉特征，从而建立全局和局部视觉元素之间的空间关联。正式地，给定区域特征 R 和视觉特征图 V_f ，公式为：

$$\begin{aligned} \tilde{V}_f &= V_f + \mathbb{Z}\left(\text{MHA}\left(Conv_Q(V_f), Conv_K(R), Conv_V(R)\right)\right), \\ V_{fr} &= \tilde{V}_f + \text{FFN}\left(\text{LN}\left(\tilde{V}_f\right)\right), \end{aligned} \tag{2}$$

，其中 MHA 、LN 和 FFN 分别表示多头注意力、层归一化和前馈网络。$Conv_Q$ 、$Conv_K$ 、$Conv_V$ 是 1*1 卷积，它们负责进行查询、键和值的投影。$\mathbb{Z}$ 表示零卷积，受到 [53] 的启发，我们引入了权重和偏置初始化为 0 的零卷积，以防止初始训练带来的不稳定性。最终的 $V_{fr} \in \mathbb{R}^{N_v\times d\times h\times w}$ 是区域增强的视觉特征图。

为了促进全局和局部视觉信息之间的深度交互，我们将框适配器融入 LVLMs 的视觉编码器中。更深层的视觉特征本质上包含更多高级语义信息，因此我们将框适配器插入到视觉编码器的几个更深层中。InternVL 系列模型使用 Vision Transformer [10] (ViT) 作为它们的视觉编码器。因此，我们将框适配器放置在 ViT 层后面，逐步增强视觉特征的局部对象信息。全局-局部深度融合结果如下：

$$\begin{aligned} V_f^{(\tilde{l})} &= \text{ViT\_Layer}\left(V_f^{(l)}\right), \\ V_f^{(l+1)} &= V_{fr}^{(l)} = \text{Box\_Adapter}\left(V_f^{(\tilde{l})}\right). \end{aligned} \tag{3}$$

表示为 $V_f^{(l)} \in \mathbb{R}^{N_v\times d\times h\times w}$ 的视觉特征图被输入到 ViT 的第 l 层。因此，最终的区域增强视觉特征从模型视角有效减轻了时空差距。

在视频指令微调之后，我们从基础模型 InternVL3 和 InternVideo2.5 中获取了异构模型。受 [16] 的启发，我们开发了一种协作投票机制来整合多个模型的描述结果。具体来说，我们计算由多个模型生成的描述之间的文本相似度。可以通过各种方法获得相似度分数，包括句子级文本嵌入的余弦相似性以及单词或字符级的匹配分数。我们选择平均相似度分数最高的句子作为最终描述。较高的平均相似度表明多个模型已达成共识，暗示该句子最准确地反映了输入视频。我们使用 IntentVC 挑战赛官方提供的数据集，该数据集基于 LaSoT 数据集标注。数据集共包含 70 种不同的类别作为特定的用户意图，每个类别包含 20 个不同对象的视频。更具体地说，每个视频的帧率设置为 1，每个视频帧对其对应的对象在标准的 COCO 格式中具有唯一的视觉定位标注，如 [x,y,w,h]。当对象移出场景时，其对应的定位框被设置为 [0,0,0,0,0]。训练集、公共测试集和私有测试集按 14:3:3 的比例划分，其中训练集中的每个视频都有五个精细手动标注的标题。

我们所有的实验都是在 Pytorch 2.1.1 和 CUDA 12.1 环境中使用 4 个 NVIDIA H100 80G GPU 进行的。在训练过程中，我们冻结视觉提取器，然后使用 rank=128 的 lora 策略训练 LLM。对于每个消融实验，我们使用批量大小为 16 的 AdamW 优化器（$\beta_1 = 0.9$ 、$\beta_2 = 0.999$ 和 $weight\_decay = 0.05$ ）。学习率的初始值为 2×10^{-5} ，并通过余弦退火调度更新。训练图像大小被强制设定为 448×448 像素。对于数据增强，我们仅在时间维度上使用随机采样策略，训练过程中随机采样 32-48 帧，推理时固定使用 48 帧。

3.3 评估指标。

遵循 IntentVC 挑战，我们将使用四个最常用的度量标准来评估视频字幕生成：BLEU@4 [23] 、METEOR [2] 、CIDEr [35] 和 ROUGE-L [26] 。

为了验证效果，我们将所提方法与一些先进的 LVLM 方法在 IntentVC 公共测试集上进行了比较。定性比较结果如表 **??** 所示。我们选择了四个最新的开源 LVLM，包括 VAST [6] 、Qwen2.5-VL [1] 、InternVideo2.5 [40] 、InternVL3 [57] ，然后在 IntentVC 数据集上对它们进行了微调以进行公平比较。如表 **??** 所示，我们提出的方法在 CIDEr、METEOR、BLEU@4、ROUGE-L 上取得了最佳结果，这证明了我们提出的方法和策略的有效性。尽管 InternVideo 是一个专注于视频领域的生成性大型模型，但我们提出的方法仍在 CIDEr 上超过它 37.71，所有其他指标也有显著提升。

在本节中，我们提供了深入的分析，并展示了我们提出的每个组件的有效性。表 1 显示了使用 InternVL3 作为基线的每个组件的消融实验。为了简洁和易于理解，我们仅展示公共测试集的指标，私有测试集也表现出大致相似的趋势。

Table 1: 消融实验。TP、VP、BA 分别代表文本提示、视觉提示和框适配器。所有实验均使用长度为 5 的束搜索策略进行推理，其余实验设置与 Sec. ?? 相同。

TP	VP	BA	BLEU@4	METEOR	CIDEr	ROUGE-L
			40.56	56.97	196.2	58.01
✓			43.45	58.54	211.45	59.02
	✓		43.22	58.88	210.76	58.89
✓	✓		42.17	<u>59.84</u>	214.45	58.43
		✓	42.19	57.73	204.71	58.02
✓		✓	44.98	60.67	223.01	60.7
✓	✓	✓	<u>43.72</u>	59.29	<u>217.17</u>	<u>59.08</u>

提示组合。如表 1 所示，这两种不同模态的提示可以为基线提供相当大的性能提升，这表明合理的提示可以显著提高模型对用户意图的关注，并能够有效引导大型语言模型生成符合意图的文本。然而，结合视觉和文本提示并没有带来预期中的大幅提升，模型在 CIDEr 上仅表现出小幅提升（211.45 → 214.45）。我们认为这是因为任一提示都足以提高模型关注目标的能力，而将它们结合使用会导致冗余，从而引发过拟合。因此，我们将视觉和文本提示分开作为异质模型参与最终的集成，而不是在单一模型中同时使用它们。

Box adapter。引入 box adapter 后，模型理解用户意图的能力进一步提高。具体来说，与使用文本提示的模型相比，CIDEr

Table 2: 关于框适配器位置的实验。“嵌入层”表示视觉模型后的嵌入部分，其余部分表示在最后 XMATHX_n 层中加入框适配器。

Settings	BLEU@4	METEOR	CIDEr	ROUGE-L
baseline	43.45	58.54	211.45	59.02
+embed layer	43.79	59.64	217.74	59.96
+last 3 layers	43.79	59.31	219.54	59.62
+last 5 layers	44.98	60.67	223.01	60.7
+last 8 layers	42.14	58.32	206.94	58.2
+last 9 layers	42.22	57.87	205.93	58.48

Table 3: 集成实验。我们简单地让 InternVL 处理较短的视频，InternVideo 处理较长的视频，最后将结果连接在一起。

Settings	BLEU@4	METEOR	CIDEr	ROUGE-L
InternVL3 [57]	43.45	58.54	211.45	59.02
InternVideo2.5 [40]	42.77	61.37	215.62	59.00
fusion	44.28	61.01	221.0	59.96

的性能从 211.45 提高到 223.01，这证明了 box adapter 在控制视频字幕生成方面的有效性。此外，由于 box adapter 可以动态地集成到视觉提取器中，表格 2 展示了在不同层级添加 box adapter 的效果对比。从实验结果来看，在过多层中加入 box adapter 不仅会使网络变得臃肿，还会因过拟合而影响准确性。权衡利弊，我们选择将 box adapter 集成到视觉模型的最后五层中，CIDEr 可以达到最高的 223.01。

融合的必要性。我们选择了视频领域的两种主流 LVLMs 作为基准，其中 InternVL 适合处理短视频，而 InternVideo 由于使用了标记压缩策略能够处理较长的视频。为了验证融合的必要性，我们手动截断每个视频，帧数小于 74 的视频使用 InternVL 处理，反之则使用 InternVideo 处理，实验结果如表 3 所示。从结果可以清楚地看出，根据模型的舒适区进行简单融合也能有效提高模型的准确性，这促使我们最终使用投票策略来融合更多的模型。

在本文中，我们提出了 IntentVCNet，这是一种新颖的面向意图的可控视频字幕框架，旨在解决现有大型视觉语言模型的基本时空差距。我们的方法通过在静态空间理解和动态时间建模之间架起桥梁，解决了生成用户可控的、面向意图的字幕的核心挑战。首先，我们引入了一种提示组合策略，将语言指令中的数值坐标与视频数据中的视觉提示融合，从而实现跨视觉和语言领域的细粒度目标定位。其次，我们开发了一种参数高效的框适配器，通过全局-局部特征融合增强面向意图的对象与帧图像之间的空间交互性。我们的方法可以生成以特定对象为中心的、面向意图的字幕，同时保持上下文连贯性，这代表了可控视频理解的重大进步。未来的工作将探索将我们的方法扩展到多对象意图控制，并研究更复杂的长篇视频内容时间建模策略。

4 致谢

本工作部分由中国博士后科学基金（2025M771515）和安徽省博士后科研项目基金（2025C1166）资助。本文的计算工作得到中国科学技术大学网络信息中心和智慧校园项目的技术支持。我们对他们的支持表示诚挚的感谢。

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[3] Qianwen Cao, Heyan Huang, and Boran Wang. 2025. From Skeleton to Flesh: Aggregated Relational Transformer Towards Controllable Video Captioning with Two-Step Decoding. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*. 61–70.

[4] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292* (2023).

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195* (2023).

[6] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems* 36 (2023), 72842–72866.

[7] Tseng-Hung Chen, Kuo-Hao Zeng, Wan-Ting Hsu, and Min Sun. 2017. Video captioning via sentence augmentation and spatio-temporal attention. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*. Springer, 269–286.

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476* (2024).

[9] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2, 3 (2023), 6.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[14] Xuelong Li, Bin Zhao, Xiaoqiang Lu, et al. 2017. MAM-RNN: Multi-level attention model based RNN for video captioning.. In *IJCAI*, Vol. 2017. 2208–2214.

[15] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4582–4597.

[16] Yiming Li and Zhao Zhang. 2024. The First Place Solution of WSDM Cup 2024: Leveraging Large Language Models for Conversational Multi-Doc QA. arXiv:2402.18385 [cs.CL]

[17] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071* (2024).

[18] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* (2023).

[19] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17949–17958.

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[21] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093* (2023).

[22] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*. Springer, 417–435.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[24] Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. *arXiv preprint arXiv:1708.02300* (2017).

[25] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[26] Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

[27] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2023. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720* (2023).

[28] Peipei Song, Dan Guo, Jun Cheng, and Meng Wang. 2023. Contextual Attention Network for Emotional Video Captioning. *IEEE Transactions on Multimedia* 25 (2023), 1858–1867.

[29] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, and Meng Wang. 2024. Emotional Video Captioning With Vision-Based Emotion Interpretation Network. *IEEE Transactions on Image Processing* 33 (2024), 1122–1135.

[30] Peipei Song, Dan Guo, Xun Yang, Shengeng Tang, Erkun Yang, and Meng Wang. 2023. Emotion-Prior Awareness Network for Emotional Video Captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 589–600.

[31] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7464–7473.

[32] Jiahui Sun, Peipei Song, Jing Zhang, and Dan Guo. 2024. Syntax-Controllable Video Captioning with Tree-Structural Syntax Augmentation. In *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*. 1–7.

[33] Yunlong Tang, Jing Bi, Chao Huang, Susan Liang, Daiki Shimada, Hang Hua, Yunzhong Xiao, Yizhi Song, Pinxin Liu, Mingqian Feng, et al. 2025. Caption Anything in Video: Fine-grained Object-centric Captioning via Spatiotemporal Multimodal Prompting. *arXiv preprint arXiv:2504.05541* (2025).

[34] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[36] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. 2019. Controllable video captioning with pos sequence guidance based on gated fusion network. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2641–2650.

[37] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. 2024. Elysium: Exploring object-level perception in videos via mllm. In *European Conference on Computer Vision*. Springer, 166–185.

[38] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4213–4222.

[39] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*. Springer, 396–416.

[40] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. 2025. InternVideo2. 5: Empowering Video MLLMs with Long and Rich Context Modeling. *arXiv preprint arXiv:2501.12386* (2025).

[41] Xinlong Xiao, Yuejie Zhang, Rui Feng, Tao Zhang, Shang Gao, and Weiguo Fan. 2020. Video captioning with temporal and region graph convolution network. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[42] Jinheng Xie, Songhe Deng, Bing Li, Haozhe Liu, Yawen Huang, Yefeng Zheng, Jurgen Schmidhuber, Bernard Ghanem, Linlin Shen, and Mike Zheng Shou. 2024. Tune-an-ellipse: Clip has potential to find what you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13723–13732.

[43] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841* (2024).

[44] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430* (2022).

[45] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[46] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441* (2023).

[47] Linli Yao, Yuanmeng Zhang, Ziheng Wang, Xinglin Hou, Tiezheng Ge, Yuning Jiang, Xu Sun, and Qin Jin. 2024. Edit As You Wish: Video Caption Editing with Multi-grained User Control. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1924–1933.

[48] Cheng Ye, Weidong Chen, Jingyu Li, Lei Zhang, and Zhendong Mao. 2024. Dual-path collaborative generation network for emotional video captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 496–505.

[49] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15405–15416.

[50] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704* (2023).

[51] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106* (2025).

[52] Junchao Zhang and Yuxin Peng. 2019. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8327–8336.

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

[54] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2025. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European conference on computer vision*. Springer, 52–70.

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).

[56] Fangyi Zhu, Jenq-Neng Hwang, Zhanyu Ma, Guang Chen, and Jun Guo. 2020. Ovc-net: Object-oriented video captioning with temporal graph and detail enhancement. *arXiv preprint arXiv:2003.03715* (2020).

[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479* (2025).