X-InstructBLIP: A Framework for Aligning Image, 3D, Audio, Video to LLMs and its Emergent Cross-modal Reasoning

Artemis Panagopoulou[1⋆], Le Xue[2,⋆⋆], Ning Yu[2,⋆⋆], Junnan Li[2], Dongxu Li[2], Shafiq Joty[2], Ran Xu[2], Silvio Savarese[2], Caiming Xiong[2], and Juan Carlos Niebles[2]

[1] University of Pennsylvania
artemisp@seas.upenn.edu
[2] Salesforce AI Research

Abstract. Recent research has achieved significant advancements in visual reasoning tasks through learning image-to-language projections and leveraging the impressive reasoning abilities of Large Language Models (LLMs). This paper introduces an efficient and effective framework that integrates multiple modalities (images, 3D, audio and video) to a frozen LLM and demonstrates an emergent ability for cross-modal reasoning (2+ modality inputs). Our approach explores two distinct projection mechanisms: Q-Formers and Linear Projections (LPs). Through extensive experimentation across all four modalities on 16 benchmarks, we explore both methods and assess their adaptability in integrated and separate cross-modal reasoning. The Q-Former projection demonstrates superior performance in single modality scenarios and adaptability in joint versus discriminative reasoning involving two or more modalities. However, it exhibits lower generalization capabilities than linear projection in contexts where task-modality data are limited. To enable this framework, we devise a scalable pipeline that automatically generates high-quality, instruction-tuning datasets from readily available captioning data across different modalities, and contribute 24K QA data for audio and 250K QA data for 3D. To facilitate further research in cross-modal reasoning, we introduce the DisCRn (**Dis**criminative **C**ross-modal **R**eason**n**ing (DisCRn)) benchmark comprising 9K audio-video QA samples and 28K image-3D QA samples that require the model to reason discriminatively across disparate input modalities. Code and data is available at https://github.com/salesforce/LAVIS/tree/main/projects/xinstructblip.

Keywords: multimodal · x-modal alignment · cross-modal reasoning

1 Introduction

Humans inherently process information from multiple sensory modalities to interpret their surroundings and make decisions based on a comprehensive view of

⋆ Work done while interning at Salesforce Research, ⋆⋆Equal mentorship contribution.

X-InstructBLIP：图像、3D、音频、视频对齐LLMs及其紧急跨模态推理的框架

Artemis Panagopoulou、Le Xue、Ning Yu、Junnan Li、Dongxu Li、Shafiq Joty、Ran Xu、Silvio Savarese、Caim Xiong 和 Juan Carlos Niebles

[1] 宾夕法尼亚大学
artemisp@seas.upenn.edu
[2] Salesforce AI 研究

抽象。最近的研究通过学习图像到语言的投影和利用大型语言模型 （）LLMs 令人印象深刻的推理能力，在视觉推理任务方面取得了重大进展。本文介绍了一个高效且有效的框架，该框架将多种模态（图像、3D、音频和视频）集成到冻结LLM中，并展示了跨模态推理的新兴能力（2+ 模态输入）。我们的方法探索了两种不同的投影机制：Q-Formers 和 Linear Projections （LPs）。通过在 16 个基准上对所有四种模态进行广泛实验，我们探索了这两种方法，并评估了它们在综合和单独的跨模态推理中的适应性。Q-Former 投影在单模态情景中表现出卓越的性能，并且在涉及两个或多个模态的联合推理与判别推理中的适应性。然而，在任务模态数据有限的上下文中，它表现出的泛化能力低于线性投影。为了实现这个框架，我们设计了一个可扩展的管道，从不同模态的现成字幕数据中自动生成高质量的指令调整数据集，并为音频贡献 24K QA 数据，为 3D 贡献 250K QA 数据。为了促进跨模态推理的进一步研究，我们引入了 DisCRn （判别性跨模态推理 （DisCRn）） 基准，包括 9K 音频视频 QA 样本和 28K 图像-3D QA 样本，要求模型对不同的输入模态进行判别推理。代码和数据可在
https://github.com/salesforce/LAVIS/tree/main/projects/xinstructblip 获取。

关键词： 多模态 ·X 模态对齐 ·跨模态推理

1 介绍

人类天生就处理来自多种感官模式的信息，以解释周围的环境，并根据

* 在 Salesforce Research 实习期间完成的工作，平等的导师贡献。

their environment. However, Multimodal Large Language Models (MLLMs) are primarily concentrated on visual tasks, often overlooking the rich diversity of other common modalities like Audio, Video, and 3D, and failing to tap into the potential of comprehensively understanding multiple modalities (>2) in unison, which is crucial for advanced tasks such as cross-modal reasoning[3].

The incorporation of various modalities beyond images into LLMs is still an area ripe for exploration, particularly regarding effective integration frameworks. A significant challenge lies on the lack of instruction-tuning datasets for other modalities like Audio, 3D, and Video, especially for data that involve two or more modalities simultaneously, making joint modality training a plausible but resource intensive approach to enable cross-modal reasoning.

In response to the above challenges, we introduce X-InstructBLIP, an extendable framework - illustrated in Figure 1 and further analyzed in Section 3 - designed to align various modalities (image, 3D, audio, video) to LLMs, achieving single-modal reasoning tasks for each modality and enabling cross-modal reasoning across *three or more modalities*. To facilitate this exploration and given the scarcity of unary instruction-tuning data for a spectrum of modalities other than the image modality, we introduce a simple yet potent approach in Section 4.1: a three-stage-query data augmentation technique to leverage *open-source* LLMs to extract instruction-tuning data from captioning datasets.

Our framework explores two state-of-the-art projection mechanisms on frozen LLMs - a prerequisite for maintaining separate modality training - instruction-aware Q-Formers [19] and linear projections [59]. Through an expansive evaluation on 13 benchmarks across 4 modalities we find that Q-Formers tend to exhibit higher performance on single modality tasks and versatility in distinguishing when to reason in a joint or discriminative manner in the presence of 2+ extra-linguistic modalities. Figure 2 shows illustrative results, highlighting the capabilities of our framework. To quantify and challenge this emergent ability we introduce `DisCRn` in Section 4.2, an automatically curated **Dis**criminatory **C**ross-modal **R**easo**n**ing challenge dataset requiring models to distinguish between diverse combinations of modalities, such as audio-video and 3D-image.
Our contributions are summarized as follows:
(i) We introduce an extendable framework that aligns Image, 3D, Audio, and Video to LLMs, and we benchmark its emergent cross-modal reasoning capability across two projection mechanisms. This framework does not need specific pre-training tailored to each modality. To the best of our knowledge, this is the first attempt to demonstrate that discriminative cross-modal reasoning emerges naturally through individual modality alignment to LLMs.
(ii) We introduce an automatic approach for crafting instruction-tuning datasets for a variety of modalities, leveraging only readily available captioning data and open-source language models. Contributing $\sim$ 250k samples for 3D QA data and $\sim$ 24k samples for Audio QA data.

[3] *Cross-modal reasoning* is the ability to integrate and discriminate information from multiple modalities over text, in contrast to "multimodal reasoning," traditionally reserved for vision-language tasks.

他们的环境。然而，多模态大型语言模型 （MLLM） 主要集中在视觉任务上，往往忽视了音频、视频和 3D 等其他常见模态的丰富多样性，未能挖掘出全面理解多种模态（>2） 的潜力，这对于跨模态推理等高级任务至关重要。

将图像以外的各种模式纳入其中LLMs仍然是一个值得探索的成熟领域，尤其是在有效的集成框架方面。一个重大挑战在于缺乏音频、3D 和视频等其他模态的指令调整数据集，特别是对于同时涉及两个或多个模态的数据，这使得联合模态训练成为一种合理但资源密集型的方法，可以实现跨模态推理。

为了应对上述挑战，我们引入了 X-InstructBLIP，这是一个可扩展的框架 - 如图 1 所示，并在第 3 节中进一步分析，旨在将各种模态（图像、3D、音频、视频）对齐 LLMs，实现每种模态的单模态推理任务，并支持跨三种或更多模态的跨模态推理。为了促进这种探索，并考虑到图像模态以外的一系列模态的一元指令调整数据的稀缺性，我们在第 4.1 节中引入了一种简单而有效的方法：一种三阶段查询数据增强技术，利用开源LLMs从字幕数据集中提取指令调整数据。

我们的框架探索了两种最先进的冻结投射机制 LLMs - 保持独立模态训练的先决条件 - 教学感知 Q-Formers [19] 和线性投影 [59]。通过对 4 种模态的 13 个基准的广泛评估，我们发现 Q-Formers 往往在单一模态任务上表现出更高的表现，并且在存在 2+ 语言外模态的情况下，在区分何时以联合或判别方式进行推理方面具有多功能性。图 2 显示了说明性结果，突出了我们框架的功能。为了量化和挑战这种新兴能力，我们在第 4.2 节中引入了 DisCRn，这是一个自动策划的判别性跨模态推理挑战数据集，要求模型区分不同的模态组合，例如音频-视频和 3D 图像。我们的贡献总结如下：

（i） 我们引入了一个可扩展的框架，将图像、3D、音频和视频与 LLMs对齐，并在两种投影机制中对其新兴的跨模态推理能力进行了基准测试。该框架不需要为每种模式量身定制的特定预训练。据我们所知，这是第一次尝试证明判别性跨模态推理是通过个体模态对齐自然产生的LLMs。

（ii） 我们引入了一种自动方法，用于为各种模态制作指令调整数据集，仅利用现成的字幕数据和开源语言模型。为 3D QA 数据提供 ~ 250k 样本，为音频 QA 数据提供 ~ 24k 样本。

[3] 跨模态推理是指在文本上整合和区分来自多种模态的信息的能力，这与传统上为视觉语言任务保留的“多模态推理”形成鲜明对比。

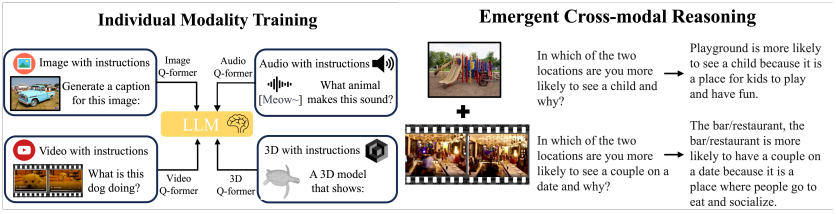


Fig. 1: Despite utilizing distinct pre-trained encoders for each modality and independently aligning them to language through individual instruction aware Q-Formers, X-InstructBLIP demonstrates emergent abilities in cross-modal comprehension.

(iii) We collect `DisCRn`, the first dataset designed for evaluating instruction-based cross-modal discriminative reasoning. Which includes ∼36k examples across various modalities such as video, audio, 3D, and images.

2 Related Work

Vision Language Models: Recent years have seen a surge in models capable of executing a spectrum of vision-language tasks, leading to the creation of Multimodal Language Models (MLMs). These models align the static vision and language representations through various techniques, such as unified pretraining [18, 39, 47, 83, 90, 92–94, 108], vision-to-language alignment through textual feature extraction [31, 56, 80, 96, 109], vision-encoder optimization [86], and linear [21, 45, 59], transformer-based [12, 67, 73], or auto-encoder based projections [58, 111]. More relevant to this work are approaches that learn intermediate vision-informed language token representations either interleaved in LLM layers such as in Flamingo [3] and LLaMA adapter [114] or only to the input layer such as in the BLIP series [19, 50, 51] which employ Q-Former based projections, LLAVA [59], and MiniGPT4 [116] which employ linear projections.
Cross-Modal Language Models: Projection-based approaches, initially focused on images, have recently broadened to encompass audio [20, 28, 44, 85], video [4, 63, 66, 107], and 3D projections [32, 37, 104] into pre-trained large language models (LLMs). This expansion has seen the advent of unified pretraining frameworks such as mPLUG2 [102] and OnePeace [91], as well as projection-based methods for enhancing frozen LLMs, like VideoLLaMA [113] and X-LLM [11], which aim to jointly train audio and video processors. Notably, X-LLM focuses on this integration primarily during the latter stages of training. In a similar vein, ChatBridge [115] adopts a training approach akin to X-LLM but utilizes a perceiver-based projection [41]. Audio-Visual LLM [81] follows a similar training paradigm of a final joint finetuning stage, but instead of maintaining a frozen LLM, it updates it using LoRA [38]. Our method is set appart by maintaining independent finetuning throughout and a frozen LLM avoiding the instability in training due to disparately aligned modality projections. Another line of work, including ImageBind-LLM [35], PandaGPT [82] and PointLLM [32]

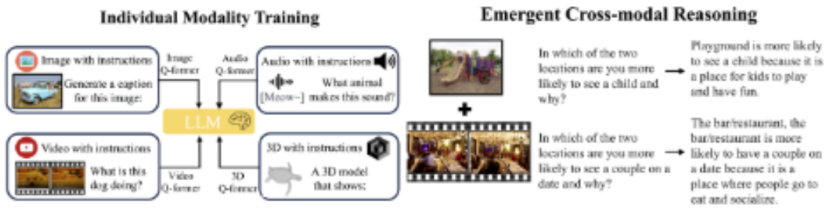


图 1：尽管为每种模态使用不同的预训练编码器，并通过单独的教学感知 Q-Formers 将它们独立地与语言对齐，但 X-InstructBLIP 在跨模态理解中表现出了涌现的能力。

（iii） 我们收集了 DisCRn，这是第一个旨在评估基于指令的跨模态判别推理的数据集。其中包括各种模式（如视频、音频、3D 和图像）的 ~36k 示例。

2 相关工作

视觉语言模型：近年来，能够执行一系列视觉语言任务的模型激增，导致了多模态语言模型 （MLM） 的创建。这些模型通过各种技术对齐静态视觉和语言表示，例如统一预训练 [18， 39， 47， 83， 90， 92–94， 108]、通过文本特征提取实现视觉到语言的对齐 [31， 56， 80， 96， 109]、视觉编码器优化 [86] 和线性 [21， 45， 59]、基于 transformer 的 [12， 67， 73] 或基于自动编码器的投影 [58,111]。与这项工作更相关的是学习中间视觉信息语言标记表示的方法，这些表示要么交错在层中LLM，如 Flamingo [3]和LLaMA适配器[114]，要么只到输入层，如采用基于Q-Former的投影的 BLIP系列[19,50,51]、LLAVA [59]和MiniGPT4 [116]，它们采用线性投影。

跨模态语言模型：基于投影的方法最初专注于图像，最近已经扩展到包括音频 [20， 28， 44， 85]、视频 [4， 63， 66， 107] 和 3D 投影 [32， 37， 104] 到预训练的大型语言模型 （）LLMs 中。这种扩展见证了统一的预训练框架的出现，如 mPLUG2 [102] 和 OnePeace [91]，以及基于投影的增强冻结LLMs的方法，如 VideoLLaMA [113] 和 XLLM [11]，旨在联合训练音频和视频处理器。值得注意的是，XLLM 主要在训练的后期阶段专注于这种集成。与此类似，ChatBridge [115] 采用了类似于 X-LLM 的训练方法，但采用了基于感知器的投影 [41]。Audio-Visual LLM [81] 遵循类似的最终关节微调阶段的训练范式，但它不是保持冻结LLM，而是使用 LoRA [38] 对其进行更新。我们的方法是通过在整个过程中保持独立的微调和冻结LLM来设置的，避免了由于不同对齐的模态投影而导致的训练不稳定。另一条工作路线，包括 ImageBind [LLM35]、PandaGPT [82] 和 PointLLM [32]

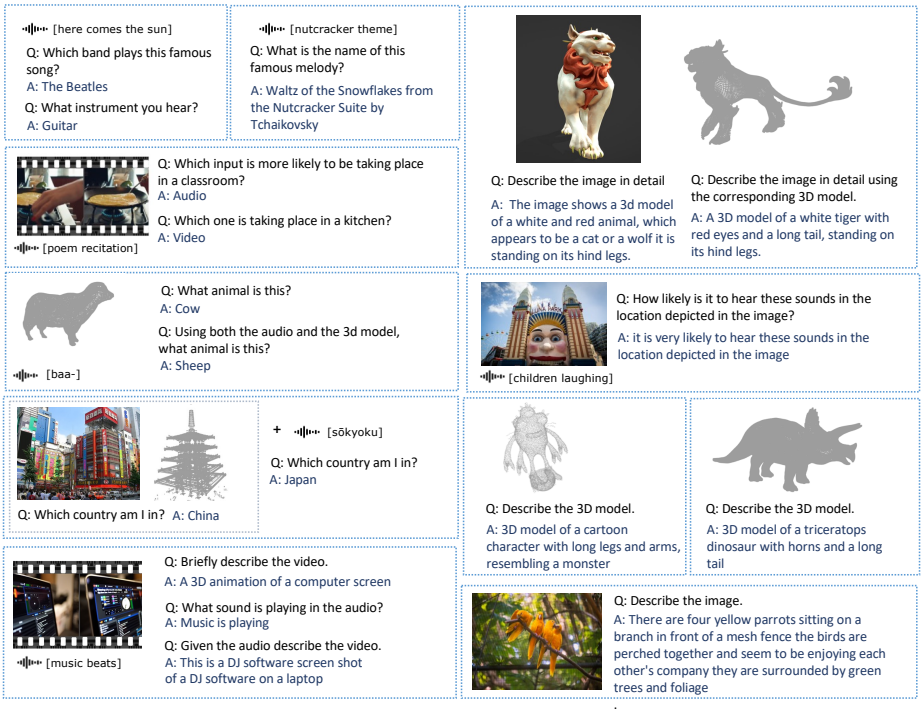


Fig. 2: Qualitative Examples: X-InstructBLIP framework effectively handles both unimodal and cross-modal tasks without training on joint data.

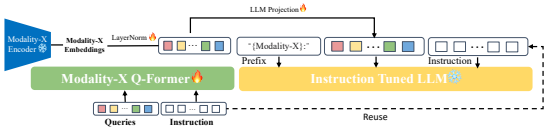
leverages unified representations such as ImageBind [27] to only implicitly align additional modalities to LLMs by only training on image-text pairs. Contemporary works such as AnyMAL [70] and OneLLM [34] have pushed the boundaries further by extending the application of projection-based approaches to additional modalities, such as 3D. Unlike other models that keep the LLM frozen, both opt to unfreeze the LLM during training. OneLLM adopts a router-based mixture of experts strategy to learn the mapping between different modalities. In contrast, AnyMAL focuses on jointly learning a LLaVA-style Projection layer for each modality during a portion of the training process.

Multimodal Multi-Input Language Tasks: The advancements in single input vision-language tasks have paved the way for the development of tasks necessitating models to concurrently reason about multiple non-linguistic inputs, such as engaging in spatial reasoning across multiple images [5], deliberating over a series of slides [84], responding to queries necessitating cross-modal reasoning across images and tables [54], or executing a range of instruction-based tasks involving multiple image inputs [54]. Despite their complexity, these tasks operate mostly within the realms of image-text modalities. Even though cross-modal tasks exist, predominantly requiring models to reason over joint audio

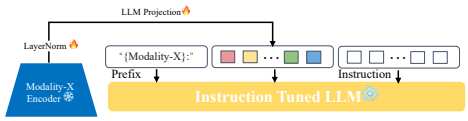
图 2：定性示例：X-InstructBLIP 框架可以有效地处理单模态和跨模态任务，而无需对关节数据进行训练。

利用 ImageBind [27] 等统一表示，仅LLMs通过对图像-文本对进行训练来隐式对齐其他模态。AnyMAL [70] 和 OneLLM [34] 等当代作品通过将基于投影的方法的应用扩展到其他模态（如 3D）而进一步突破了界限。与其他保持LLM冻结状态的模型不同，两者都选择LLM在训练期间解冻 。OneLLM 采用基于路由器的专家混合策略来学习不同模态之间的映射。相比之下，AnyMAL 专注于在训练过程的一部分中为每种模态共同学习 LLaVA 风格的投影层。

多模态多输入语言任务：单输入视觉语言任务的进步为需要模型同时推理多个非语言输入的任务的发展铺平了道路，例如跨多个图像进行空间推理 [5]，对一系列幻灯片进行审议 [84]，响应需要跨图像和表格进行跨模态推理的查询 [54]， 或执行一系列涉及多个图像输入的基于指令的任务 [54]。尽管它们很复杂，但这些任务主要在图像文本模态的领域内运行。尽管存在跨模态任务，但主要要求模型对联合音频进行推理



(a) X-Instruct Projection



(b) X-LLaVA-style Projection

Fig. 3: Projection types explored in the X-InstructBLIP Framework. (a) is an an instruction aware Q-Former projection [19] and (b) is a linear projection [59].

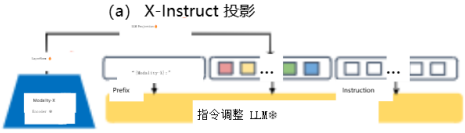
and video [2, 49], there is a gap in the evaluation of models' generative capabilities in reasoning about cross-modal inputs *contrastively*. While models are often optimized on contrastive objectives [15, 16, 36, 42, 50, 52, 53, 77], even in cross-modal settings [27, 33, 72], their evaluation is confined to classification tasks or utilizing the contrastively learned representations for downstream tasks. To address this gap, we introduce `DisCRn`, a task requiring contrastive reasoning across cross-modal inputs in an open generation setting, evaluating a model's ability to translate features of various modalities from its internal representations to its generative output distribution.

3 Method

Framework Overview: Figure 1 depicts an overview of the framework's setup which extends instruction finetuning for image alignment [19, 59] to an arbitrary number of modalities through independent fine-tuning of modality-specific projections to a frozen LLM, further broken down in Algorithm 1. X-InstructBLIP's alignment framework involves the following steps: **(1)** For each modality, collect an instruction tuning dataset suite $(x, y) \in \mathbb{D}_M$ s.t. $x = (x_M, x_T)$ is a tuple of a modality input and text, and y is the expected text output. **(2)** Let Enc_M be a modality encoder to R^{d_M} and Enc_T be a mapping from text to the LLM's embedding space R^{d_L}. Optimize a single separate projection module $f_\theta^M : R^{d_M} \rightarrow R^{kd_L}$ for each modality M on $\mathbb{D}_M$ while maintaining the parameters of the LLM frozen, where k is the number of LLM input tokens corresponding to the non-linguistic input. For sequential data, such as video and audio, we extract $N \times k$ query tokens; each frame is encoded and processed separately by the projection module. **(3)** The model is optimized under a causal language modeling objective [60]: $\min_\theta \mathcal{L}_{\text{CE}}\Big(\text{LLM}(x_{\text{LLM}}), \text{h}(y)\Big)$ where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss, θ the Q-Former parameters, y is the target sequence, $\text{LLM}(x_{\text{LLM}})$ is the LLM's prediction.

X-Instruct Projection: Figure 3a highlights all components associated with learning instruction aware Q-Former projections [19] for multiple modalities.

（a） X-Instruct 投影



（b） X-LLaVA 型投影

图 3： 在 X-InstructBLIP 框架中探索的投影类型。（a） 是指令感知的 Q-Former 投影 [19]，（b） 是线性投影 [59]。

和视频 [2， 49]，在评估模型在推理跨模态输入方面的生成能力存在差异。虽然模型通常在对比物镜上进行优化 [15， 16， 36， 42， 50， 52， 53， 77]，即使在跨模态设置 [27， 33， 72] 中，它们的评估也仅限于分类任务或将对比学习的表示用于下游任务。为了解决这一差距，我们引入了 DisCRn，这是一项需要在开放生成环境中对跨模态输入进行对比推理的任务，评估模型将各种模态的特征从其内部表示转换为其生成输出分布的能力。

3 方法

框架概述：图 1 描述了框架设置的概述，该框架通过将特定于模态的投影独立微调到冻结 LLM[19,59] 扩展到任意数量的模态，在算法 1 中进一步细分。X-InstructBLIP 的对齐框架涉及以下步骤： （1） 对于每种模态，收集一个指令调优数据集套件 （x， y） ∈ Ds.t. x = （x， x） 是模态输入和文本的元组，y 是预期的文本输出。（2） 让 Encbe 成为 Rand Encbe 的模态编码器，从文本到 LLM的嵌入空间 R 的映射。为 D 上的每个模态 M 优化一个单独的投影模块 f： R→ R，同时保持冻结的LLM参数，其中 k 是对应于非语言输入的LLM输入标记的数量。对于视频、音频等顺序数据，我们提取 N × k 个查询 Token;每一帧都由 Projection Module 单独编码和处理。 （3） 在因果语言建模目标 [60] 下对模型进行优化：minL

其中 Lis 交叉熵损失，θ Q-Former 参数，y 是目标序列，LLM（x） 是 LLM的预测。

X-Instruct 投影：图 3a 突出显示了与多种模式的学习教学感知 Q-Former 投影 [19] 相关的所有组成部分。

Algorithm 1 X-InstructBLIP Optimization Framework

Require: Set of modalities $\mathbb{M}$, each associated with a set of datasets $\mathbb{D}_M$, and set of templates $\mathbb{I} = \{I_{M_t} : M \in \mathbb{M}, t \in \mathbb{T}\}$ for each task $t \in \mathbb{T}$

for each modality M in $\mathbb{M}$ **do**
  Initialize modality-specific pre-trained encoder Enc_M
  Initialize LLM encoder Enc_T (tokenize and embed text)
  Initialize projection $f_\theta^M : R^{d_M} \rightarrow R^{kd_L}$
  for each step in number of iterations **do**
    Sample (x, y) from $\cup\mathbb{D}_M$
    Sample i_M from I_{M_t} where t is the task mapping x to y
    $z_M \leftarrow \text{Enc}_M(x)$ ▷ Encode input to embedding space
    $w_M \leftarrow f_\theta^M(z_M)$ ▷ Transform encoded input to LLM embedding space
    $x_{\text{LLM}} \leftarrow w_M \| \text{Enc}_T(i_M) \| \text{Enc}_T(x_T)$
    Prediction $\leftarrow \text{LLM}(x_{\text{LLM}})$ ▷ Get LLM's prediction
    Loss $\leftarrow \mathcal{L}_{\text{CE}}(\text{Prediction}, \text{Enc}_T(y))$ ▷ Calculate cross-entropy loss
    $\theta \leftarrow \theta - \alpha \nabla_\theta \text{Loss}$ ▷ Update projection parameters

Given a modality M encoding $z_M = \text{Enc}_M(x_M)$ and task instruction $i_M \in \mathbb{I}_{M_t}$, the Q-Former module transforms a set of k learnable embeddings $\mathbf{Q}_M = \{\mathbf{q}_{M_1} \dots \mathbf{q}_{M_K}\}$ termed *input query tokens* into instruction-aware representations of $\mathbf{Q}'_M = \text{QF}_M(\mathbf{Q}_M, z_M, i_M)$. The Q-Former module consists of two transformer submodules that share the same self-attention layers: one submodule interacts with the output of the modality encoder Enc_M and the other is a $\text{BERT}_{\text{base}}$ text transformer that serves as both an encoder and decoder. Each Q-Former is initialized with the pre-trained weights from BLIP-2 [50], without the cross-attention layers due to a dimension mismatch between the image encoder in BLIP-2 and the other modality encoders. The modality embedding z_M interacts with the instruction text i_M and input query tokens $\mathbf{Q}_M$ via cross-attention layers inserted every other transformer block, yielding the *output query tokens* $\mathbf{Q}'_M$ which are linearly projected to the frozen LLM's space through a learnable projection layer LP_M specific to each modality. Let pf_M the modality prefix, x the example text input, and y the target phrase. With $\|$ denoting concatenation, the *LLM input tokens* are: $x_{\text{LLM}} = \text{Enc}_T(\text{pf}_M) \,\|\, \text{LP}_M(\mathbf{Q}'_M) \,\|\, \text{Enc}_T(i_M) \,\|\, \text{Enc}_T(x_T))$.

X-LLaVA-style Projection: We implement an adaptation of LLaVA's architecture [59] to cater multiple modalities, similarly to the instruction aware Q-Former. Figure 3b depicts the architecture of this simple projection which linearly transforms the outputs of the modality encoder directly to the input embedding space of the LLM. Formally, the model consists of a single linear projection layer $\text{LP}_M : R^{d_M} \rightarrow R^{kd_{\text{LLM}}}$, where d_{LLM} is the LLM's embedding dimension. To compare the two projection types we match the number of trainable parameters for each modality, and maintain the training set-up.

算法 1 X-InstructBLIP 优化框架

要求：模态集 M，每个模态与一组数据集 D 相关联，以及模板集 I = {I： M ∈ M， t ∈ T} 对于每个任务 t ∈ T 1：对于每个模态 M do 2：初始化特定于模态的预训练编码器 Enc 3：初始化LLM编码器 Enc（标记化和嵌入文本） 4：初始化投影 f： R→ R 5：对于迭代次数的每个步骤，执行 6：从 ∪D 7 中采样 （x， y）： 样本 ifrom I其中 t 是任务映射到 y 8：z← Enc（x） ▷ 将输入编码为嵌入空间 9：w← f （z） ▷ 将编码输入转换为LLM嵌入空间 10：x← w‖Enc（i）‖Enc（x） 11：预测 ← LLM（x） ▷ 获取 LLM的预测 12：损失 ← L（预测， Enc（y）） ▷ 计算交叉熵损失 13：θ ← θ – α∇损失 ▷ 更新投影参数

给定编码 z= Enc（x） 的模态 M 和任务指令 i∈ I，Q-Former 模块将一组 k 个可学习的嵌入 Q= {q. . . q} 称为输入查询标记转换为 Q= QF（Q， z， i） 的指令感知表示。Q-Former 模块由两个 transformer 子模块组成，它们共享相同的自注意力层：一个子模块与模态编码器 Encand 的输出交互，另一个是 BERT 文本转换器，既用作编码器又用作解码器。每个 Q-Former 都使用 BLIP-2 [50] 中的预训练权重进行初始化，由于 BLIP-2 中的图像编码器与其他模态编码器之间的尺寸不匹配，因此没有交叉注意力层。模态嵌入与指令文本 i 和输入查询标记 Qvia 交叉注意力层交互，每隔一个转换器块插入一次，产生输出查询标记 Q，这些标记通过特定于每种模态的可学习投影层 LP 线性投影到冻结LLM的空间。让 pf 为模态前缀，x 为示例文本输入，y 为目标短语。‖表示串联时，LLM输入标记为：x= Enc（pf） ‖ LP（Q） ‖ Enc（i） ‖ Enc（x）） 。

X-LLaVA风格的投影：我们对LLaVA的架构[59]进行了改编，以适应多种模式，类似于指令感知的Q-Former。图 3b 描述了这种简单投影的架构，它将模态编码器的输出直接线性转换为 LLM.从形式上讲，该模型由单个线性投影层 LP 组成： R→ R，其中 dis 的 LLM嵌入维度。为了比较这两种投影类型，我们匹配每种模态的可训练参数数量，并保持训练设置。

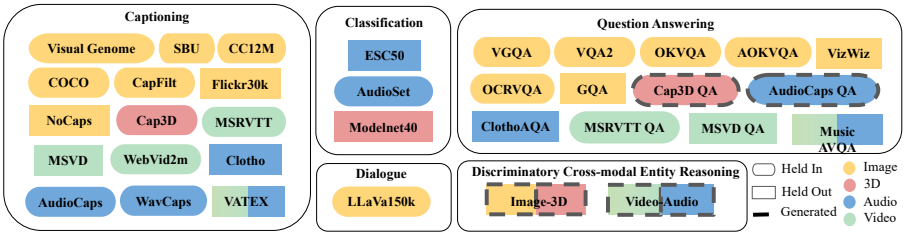


Fig. 4: Instruction Tuning and Evaluation Datasets: Oval-enclosed and square datasets are tuning and out-of-domain evaluation datasets respectively. Dashed outline is used for automatically derived datasets as described in Section 4.1.

4 Datasets

X-InstructBLIP is optimized and evaluated on a collection of pre-existing and automaticaly generated datasets succintly presented in Figure 4, discussed in Section 4.1, with more details available in the supplementary material. Section 5.2 introduces the Discriminatory Cross-modal Reasoning challenge dataset `DisCRn`[4] used to evaluate the emergent abilities of X-InstructBLIP (Section 4.2).

4.1 Fine-tuning Datasets

Existing Datasets: Figure 4 illustrates the datasets utilized for both instruction tuning and evaluation. A detailed breakdown of the dataset statistics and formats can be found in the supplementary material. For each dataset in $\mathbb{D}_M$, the collection of held-in datasets specific to modality M, a modified sampling strategy from [19] is adopted accommodating a broader range of modalities. The sampling probability for any given dataset $D_{M_d} \in \mathbb{D}_M$ is $\frac{\sqrt{|D_{M_d}|}}{\sum_{d \in [1 \ldots |\mathbb{D}_M|]} \sqrt{|D_{M_d}|}}$, with minimal adjustments as justified in the supplementary matierial.

Instruction Data Augmentation: Extracting instruction-aware representations necessitates diverse instruction-related tasks across all modalities. Notably, datasets for 3D and audio modalities are marjorly caption-centric. To address this, we leverage the open-source large language model `google/flan-t5-xxl` [97] to automatically generate question-answer pairs for the 3D and audio modalities from their corresponding captions. The process begins by prompting the model with captions to generate potential answers. These answers are then used to prompt the model to generate candidate questions. If the model's response to a question, using the caption as context aligns closely with the initial answer, the example is added to our dataset, yielding ∼250k 3D examples from

[4] The term *discriminative reasoning*, adapted from [105], refers to the ability to distinguish between sets of inputs, as opposed to *joint reasoning*, the synthesis of information from aligned sources.

图 4： 指令调优和评估数据集： 椭圆封闭和方形数据集分别是调优和域外评估数据集。虚线轮廓用于自动派生的数据集，如 Section 4.1 中所述。

4 个数据集

X-InstructBLIP 在图 4 中简洁地呈现的预先存在和自动生成的数据集集合上进行了优化和评估，第 4.1 节中讨论了，补充材料中提供了更多详细信息。第 5.2 节介绍了判别性跨模态推理挑战数据集 DisCRn 用于评估 X-InstructBLIP 的涌现能力（第 4.2 节）。

4.1 微调数据集

现有数据集：图 4 说明了用于指令调整和评估的数据集。数据集统计数据和格式的详细分类可以在补充材料中找到。对于 D 中的每个数据集，特定于模态 M 的保留数据集的集合，采用了 [19] 中修改后的采样策略，以适应更广泛的模态。任何给定数据集 D∈ Dis 的采样概率

$$p = \frac{\sqrt{|D_M|}}{\sum_{d \in [1...|D|]} \sqrt{|D_M|}}$$ 跟

根据补充材料中合理的最小调整。指令数据增强：提取指令感知表示需要跨所有模态的各种与指令相关的任务。值得注意的是，3D 和音频模态的数据集以字幕为中心。为了解决这个问题，我们利用开源大型语言模型 google/flan-t5-xxl [97] 从相应的字幕中自动生成 3D 和音频模态的问答对。该过程首先提示带有标题的模型以生成可能的答案。然后，这些答案用于提示模型生成候选问题。如果模型对问题的回答，使用标题作为上下文与初始答案紧密一致，则该示例将添加到我们的数据集中，从

[4] 术语判别推理，改编自 [105]，是指区分输入集的能力，而不是联合推理，即来自一致来源的信息的综合。

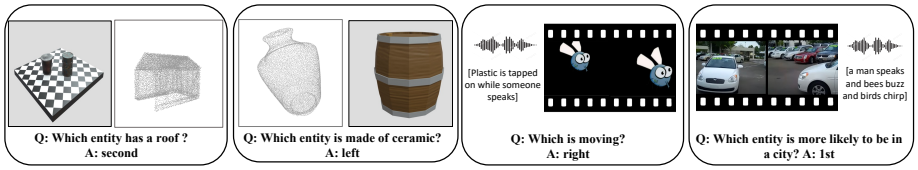


Fig. 5: `DisCRn`. Given two distinct modality inputs, select which one fits the query.

Cap3D [64] [5] and ∼24k audio examples from AudioCaps [43]. Details about the data generation and distribution are provided in the supplement.

4.2 Discriminative Cross-modal Reasoning

X-InstructBLIP offers a distinct emergent ability: reasoning across different modalities, despite individual modality training. This highlights the model's versatility and potential scalability across numerous modalities. To study this cross-modal reasoning capability, we present a **Dis**criminatory **C**ross-modal **R**easo**n**ing (`DisCRn`) challenge dataset. As shown in Figure 5 the task requires the model to discern between the properties of two entities across modalities by selecting which one satisfies a queried property. This task mandates the model to not only discriminate the inherent characteristics of the involved modalities but also to consider their relative positioning in the input. This strategic imposition serves to diminish reliance on simplistic text-matching heuristics, order bias, or potential deceptive correlations between modalities.

To generate the dataset, we prompt `google/flan-t5-xxl` in a Chain-of-Thought [98] manner to generate a set of properties for each dataset instance. Each instance is then paired with a random entity from the dataset to form a (question, answer, explanation) triplet using three examples to leverage in-context-learning [7]. A pivotal step in this creation process is a round-trip-consistency check: an example is integrated into the final dataset only when the model's predictions on the generated question, given the captions, exhibits a Levenshtein distance above 0.9 to the example answer. This refined dataset encompasses 8,802 audio-video samples sourced from the AudioCaps validation set, and 29,072 image-point cloud instances from a reserved subset of 5k point clouds from Cap3D [64]. Each instance in the dataset is coupled with two representations corresponding to the captions: (audio, video) from AudioCaps and (point cloud, images) from Cap3D. Given that the arrangement of the data can be altered, this allows for maintaining a balanced set of answers, not only in terms of the position of the answers, but also the answer modality. Human performance on the task stands at 90% indicating its high quality. More details found in the supplementary material.

[5] A subset of 5k point clouds is held-out from Cap3D for the construction of `DisCRn` (Section 4.2). This exclusion is maintained both in captioning and QA.

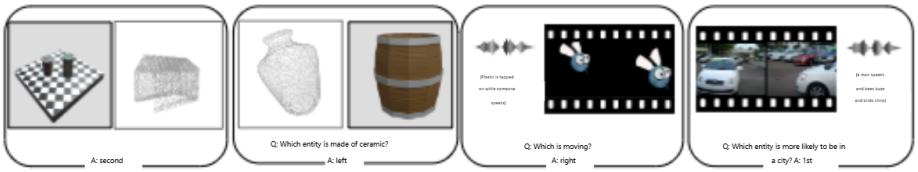


图 5： DisCRn.给定两个不同的模态输入，选择适合查询的模态。

来自 AudioCaps 的 Cap3D [64] 和 ~24k 音频示例 [43]。补充中提供了有关数据生成和分发的详细信息。

4.2 判别性跨模态推理

X-InstructBLIP 提供了一种独特的涌现能力：尽管接受了个人模态培训，但可以跨不同模态进行推理。这凸显了该模型的多功能性和在多种模式中的潜在可扩展性。为了研究这种跨模态推理能力，我们提出了一个判别性跨模态推理 （DisCRn） 挑战数据集。如图 5 所示，该任务要求模型通过选择满足查询属性的实体来区分跨模态的两个实体的属性。这项任务要求模型不仅要区分所涉及模态的固有特征，还要考虑它们在输入中的相对位置。这种战略性强加有助于减少对简单文本匹配启发式、顺序偏差或模态之间潜在的欺骗性相关性的依赖。

为了生成数据集，我们以思路链 [98] 的方式提示 google/flan-t5-xxl 为每个数据集实例生成一组属性。然后将每个实例与数据集中的随机实体配对，使用三个示例形成（问题、答案、解释）三元组，以利用上下文学习 [7]。此创建过程中的一个关键步骤是往返一致性检查：只有当模型对生成问题的预测（给定标题）与示例答案的 Levenshtein 距离大于 0.9 时，才会将示例集成到最终数据集中。这个改进的数据集包括来自 AudioCaps 验证集的 8,802 个音频-视频样本，以及来自 Cap3D 的 5k 点云保留子集的 29,072 个图像点云实例 [64]。数据集中的每个实例都与与字幕对应的两种表示形式相结合：来自 AudioCaps 的（音频、视频）和来自 Cap3D 的（点云、图像）。鉴于数据的排列可以改变，这允许保持一组平衡的答案，不仅在答案的位置方面，而且在答案模态方面。人类在任务中的表现为 90%，表明其质量很高。更多详细信息可在补充材料中找到。

[5] Cap3D 保留了 5k 点云的子集，用于构建 DisCRn（第 4.2 节）。此排除项在字幕和 QA 中均保留。

5 Experiments

We study the effectiveness of X-InstructBLIP as a comprehensive solution for incorporating cross-modality into pre-trained frozen LLMs. Following a debrief on the implementation details in Section 5.1, Section 5.2 verifies the framework's competitiveness in individual modality-to-text tasks, and explores its emergent cross-modal reasoning ability even in the absence of joint optimization.

5.1 Implementation Details

X-InstructBLIP is built on the LAVIS library's framework [48] atop of the Vicuna v1.1 7b and 13b models [17]. We adopt `EVA-CLIP-ViT-G/14` [24] as the encoder for image and video, for audio $\texttt{BEATs}_{\texttt{iter3+}}$ [13] and for 3D `ULIP-2` [`PointBERT` backbone] [78]. In the X-Instruct setup, each Q-Former optimizes 188M trainable parameters and learns $K = 32$ query tokens with a a hidden dimension of size 768 to select a *single best model* per modality. Raw inputs undergo standardized pre-processing prior to encoding. All Q-Formers are pre-initialized with BLIP-2 stage-1 weights [19] except for the video Q-Former which is initialized from the last iteration of the corresponding image Q-Former. Details on preprocessing and training hyperparameters for each modality are included in the supplement. The X-LLaVA-style setup linear projection is uniformly initialized and tuned to match the number of trainable parameters in X-Instruct.

All models are optimized on 8 A100 40GB GPUs using AdamW [62] with β_1= 0.9, β_2= 0.999, and weight decay of 0.05. The learning rate warms up linearly over the initial 1,000 steps from 10^{-8} to 10^{-5}, followed by a cosine decay to a minimum of 0. Evaluation hyper-parameters and templates are consistent across tasks, minimally adapted to each modality as detailed in the supplement.

5.2 Results

Our primary aim is to demonstrate the adaptability of our framework across various modalities without relying on large-scale pre-training stages or joint modality data. Nevertheless, to ensure our approach's effectiveness and comparability, we juxtapose its performance to other *methods that employ projections to pre-trained frozen or partially frozen LLMs*, wherever possible. This serves as a mere *sanity check*, verifying that our method is both effective and competitive.

Individual Modality Understanding We evaluate the framework's performance across a range of single modality to text tasks, illustrating its versatility and efficacy across all four explored modalities. Tables 1, 3, 4, and 2 summarize X-InstructBLIP's out-domain performance across 3D, audio, image, and video.

3D: Table 1 shows the results on zero-shot classification on ModelNet40 [99] under two setups: classification in closed vocabulary using loss ranking [52] and open generation where the model is prompted to `describe the 3d model` and

5 实验

我们研究了 X-InstructBLIP 作为将跨模态整合到预训练冻结中的综合解决方案的有效性 LLMs。在对 Section 5.1 中的实现细节进行汇报之后，Section 5.2 验证了该框架在单个模态到文本任务中的竞争力，并探讨了即使在没有联合优化的情况下其新兴的跨模态推理能力。

5.1 实现细节

X-InstructBLIP 构建在 Vicuna v1.1、7b 和 13b 模型 [17] 之上的 LAVIS 库框架 [48]。我们采用 EVA-CLIP-ViT-G/14 [24] 作为图像和视频、音频 BEAT[13] 和 3D ULIP-2 [PointBERT 主干] [78] 的编码器。在 X-Instruct 设置中，每个 Q-Former 优化 188M 可训练参数，并学习 K = 32 个隐藏维度大小为 768 的查询标记，以为每个模态选择一个最佳模型。原始输入在编码之前经过标准化的预处理。所有 Q-Former 都使用 BLIP-2 stage-1 权重 [19] 进行了预初始化，但视频 Q-Former 除外，它是从相应图像 Q-Former 的最后一次迭代中初始化的。补充中包含有关每种模式的预处理和训练超参数的详细信息。X-LLaVA 风格的设置线性投影经过统一初始化和调整，以匹配 X-Struct 中可训练参数的数量。

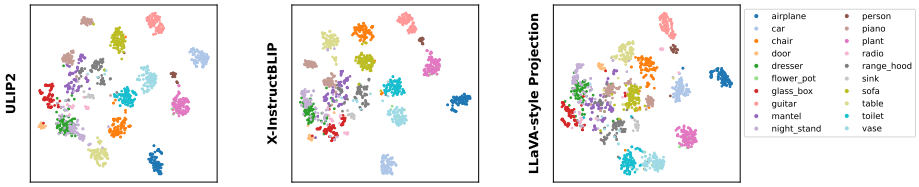
所有型号均在 8 个 A100 40GB GPU 上使用 AdamW [62] 进行了优化，β= 0.9，β= 0.999，权重衰减为 0.05。学习率在最初的 1000 步中从 10 到 10 线性预热，然后余弦衰减到最小值 0。评估超参数和模板在任务之间是一致的，最低限度地适应每种模式，如补充中所述。

5.2 结果

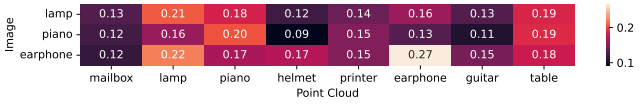
我们的主要目标是展示我们的框架在各种模态中的适应性，而无需依赖大规模的预训练阶段或联合模态数据。尽管如此，为了确保我们方法的有效性和可比性，我们将其性能与其他方法并列，这些方法尽可能采用投影来预训练冻结或部分冻结LLMs。这仅仅是一种健全性检查，验证我们的方法既有效又具有竞争力。

个体模态理解 我们评估了该框架在一系列单一模态到文本任务中的表现，说明了它在所有四种探索的模态中的多功能性和有效性。表 1、3、4 和 2 总结了 X-InstructBLIP 在 3D、音频、图像和视频方面的域外性能。

3D：表 1 显示了在两种设置下在 ModelNet40 [99] 上进行零样本分类的结果：使用损失排名 [52] 的封闭词汇分类和开放生成，其中提示模型描述 3D 模型和



(a) TSne plots of *LLM projections* for 20 randomly sampled Modelnet40 classes.



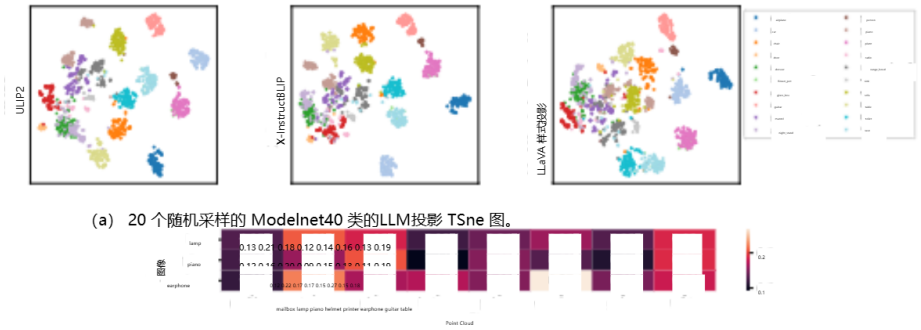
(b) Heatmap of the relative cosine similarity between image and point cloud query outputs from samples in Shapenet.

Fig. 6: Analysis on the alignment of X-InstructBLIP representations.

correctness is validated if a *single* class from the 40 candidates is present in the generation. In both projection setups, X-InstructBLIP significantly outperforms the InstructBLIP baseline, which processes a single view rendering of the point cloud. Interestingly, the X-Instruct projection setup outperforms not only the X-LLaVA-style projection but also PointLLM [32] that learns a similar projection but- unlike this set-up - employs RGB features. It also outperforms PointBindLLM [32] which trains an adapter on ImageBind [27] image encodings, and relies to the common embedding space to generalize to point clouds, showing the importance of individual modality encoders in our framework. This is further bolstered by the TSNE [65] visualization of the ULIP-2, X-Instruct and LLaVA-style Projection representations in Figure 6a showing that the LLaVA-style Projection breaks class separation leading to lower performance of 16.4 and 19.2 points in classification and open generation accuracy compared to X-Instruct Projection. We further observe, in Figure 6b, a mild effect of relative alignment between similar classes across modalities since the cosine similarity of the image and point cloud query outputs of similar classes in Shapenet are higher compared to dissimilar ones.

Audio: Table 3 shows X-InstructBLIP's performance in audio classification, question answering, and captioning tasks on ESC50 [76], ClothoAQA [57], and Clotho [22], respectively. Classification is evaluated both in close (cls) and open generation settings. Both X-InstructBLIP variants outperform ImageBindLLM in all tasks, potentially suggesting that separate encoders and audio specific training data are beneficial for audio-to-text alignment. Notably, X-LLaVA-style Projection outperforms the X-Instruct Projection on Audio QA, while underperforming in all other tasks. This is likely due to the low amount of Audio QA data priming the instruction aware projections to produce a small set of responses.

Image: While no large variations in performance are expected in comparison to InstructBLIP [19], Table 4 presents results on image captioning, visual ques-



（a） 20 个随机采样的 Modelnet40 类的LLM投影 TSne 图。

（b） Shapenet 中样本的图像和点云查询输出之间的相对余弦相似性的热图。

图 6： X-InstructBLIP 表示的对齐分析。

如果生成中存在 40 个候选类中的单个类，则会验证正确性。在这两种投影设置中，X-InstructBLIP 的性能明显优于 InstructBLIP 基线，后者处理点云的单个视图渲染。有趣的是，X-Instruct 投影设置不仅优于 X-LLaVA 风格的投影，而且优于 PointLLM [32]，后者学习了类似的投影，但与此设置不同，它采用了 RGB 功能。它还优于 PointBindLLM [32]，后者在 ImageBind [27] 图像编码上训练适配器，并依靠公共嵌入空间来推广到点云，显示了单个模态编码器在我们的框架中的重要性。图 6a 中 ULIP-2、X-Instruct 和 LLaVA 型投影表示的 TSNE [65] 可视化进一步支持了这一点，该可视化显示，与 XInstruct 投影相比，LLaVAstyle 投影打破了类分离，导致分类性能和开放生成精度降低 16.4 和 19.2 个点。在图 6b 中，我们进一步观察到跨模态的相似类之间的相对对齐的轻微影响，因为 Shapenet 中相似类的图像和点云查询输出的余弦相似性高于不同类。

音频：表 3 分别显示了 X-InstructBLIP 在 ESC50 [76]、ClothoAQA [57] 和 Clotho [22] 上的音频分类、问答和字幕任务中的表现。分类在关闭 （cls） 和开放生成设置中进行评估。X-InstructBLIP 变体在所有任务中都优于 ImageBindLLM，这可能表明单独的编码器和特定于音频的训练数据有利于音频到文本的对齐。值得注意的是，X-LLaVA 风格的投影在音频 QA 上优于 X-Instruct 投影，但在所有其他任务中表现不佳。这可能是由于音频 QA 数据量低，导致指令感知投影产生一小部分响应。

图片：虽然与 InstructBLIP 相比 [19] 预计性能不会有太大变化，但表 4 显示了图像标题、视觉问题

tion answering, MME [25], and MMVET [112] as a sanity check. While X-InstructBLIP outperforms InstructBLIP on VizWiz [6] there is a mild drop in performance overall, likely due to the lack of BLIP2 Stage-2 finetuning, and the expanded template space which introduces a trade-off of generalization and performance as shown by the increased prompt robustness of X-InstructBLIP in the supplement.

Silent Video: Table 2 evaluates X-InstructBLIP on out-of-domain video tasks. We compare performance with prominent baselines that rely on *frozen or partially frozen* LLMs and show comparable or improved performance on Video Question Answering (VQA). However, due to the nature of the instruction aware

	Close	Open
InstructBLIP (7b) [19]	31.4	23.7
InstructBLIP (13b) [19]	31.5	25.5
Point-LLMv2+(RGB) (7b) [104]	-	32.3
Point-LLMv2+(RGB)(13b) [104]	-	31.8
PointBind-LLM (7b) [32]	47.3	36.3
X-LLaVA-style Proj. (7b)	46.4	30.2
X-Instruct Proj. (7b)	62.8	49.4
X-Instruct Proj. (13b)	**65.1**	**50.0**

Table 1: Zero-shot 3D classification on Modelnet40 [87] test set.

	PT	MSVD *test*	VATEX *val*	MSVDQA *test*
FrozenBiLM [107]	✓	-	-	33.8
VideoLLaMA [113]	✓	-	-	51.6
InstructBLIP [19]	×	87.2	57.6	41.2
X-LLaVA-style Proj. (7b)	×	105.3	46.2	49.8
X-Instruct Proj. (7b)	×	116.1	**59.2**	**51.7**
X-Instruct Proj. (13b)	×	**124.3**	52.0	49.2

Table 2: Out-Domain Silent Video Results. PT denotes video pretraining stage.

	$ESC50_{close}$ Acc.	$ESC50_{open}$ Acc.	ClothoAQA EM	Clotho v1 CIDEr	Clotho v1 SPIDEr	Clotho v2 CIDEr	Clotho v2 SPIDEr
ImageBind [27]	66.9	×	×	×	×	×	×
MWAFM [30]	-	-	22.2	-	-	-	-
Pengi [20]	-	53.9	64.5	39.6	30.0	32.9	27.1
Kim et. al., 2023 [44]	-	-	-	-	-	19.2	13.3
ImageBind-LLM (7B) [35]	40.1	27.4	10.3†	3.7	5.5	3.2	5.5
X-LLaVA-style Proj. (7b)	67.4	20.3	**26.9**	25.3	16.6	22.0	14.8
X-Instruct Proj. (7b)	75.9	**38.2**	21.4	**29.4**	**19.5**	26.7	17.8
X-Instruct Proj. (13b)	**77.1**	34.6	21.7	28.7	18.8	**27.5**	**18.0**

Table 3: Out-Domain Audio Quantitative Results.

		Flickr30k	NoCaps	VizWiz	GQA	MME	MMVet
Vision	Flamingo 9B [3]	61.5	-	28.8	-	-	-
	BLIP-2 [50]	76.1	107.5	29.8	44.7	**1293.8**	22.4
	InstructBLIP [19]	**84.5**	**123.1**	34.5	**49.5**	1212.8	26.2
	MiniGPT4 (7b) [116]	-	-	-	32.2	1158.6	22.1
	LLaVA (7b) [59]	27.7	33.1	-	-	717.5	27.4
	LLaMA-adapter (13b) [114]	30.5	41.7	-	-	1222.0	-
X-Modal	PandaGPT (13b) [82]	23.0	29.7	-	-	871.2	19.6
	ImagebindLLM (7b) [35]	23.5	30.4	-	-	989.3	-
	X-LLaVA-style Proj. (7b)	6.2	22.3	27.7	41.5	866.7	17.0
	X-Instruct Proj. (7b)	82.1	117.7	34.9	48.1	891.8	29.0
	X-Instruct Proj. (13b)	74.7	114.5	**36.0**	49.2	1174.0	**35.1**

Table 4: Out-Domain Image Quantitative Results.

Single Modality Quantitative Results. Underlined numbers indicate in-domain evaluations. **Bold** indicates the top zero-shot performance. Blue indicates second best zero-shot performance. Purple denotes evaluations conducted independently. Models denoted with 7b and 13b indicate the underlying LLM size. Gray shaded rows correspond to X-InstructBLIP variants, and Yellow to the LLaVA-style [59] model equivalent. CIDEr score [89] is reported for captioning, accompanied by SPIDEr [61] score for audio captioning and Top-1 accuracy for QA and classification tasks. † signifies a relaxed exact match metric where the ground truth is a substring of the prediction.

tion 回答、MME [25] 和 MMVET [112] 作为健全性检查。虽然 XInstructBLIP 在 VizWiz 上的性能优于 InstructBLIP [6]，但整体性能略有下降，这可能是由于缺乏 BLIP2 Stage-2 微调，以及扩展的模板空间，这引入了泛化和性能的权衡，如补充中 X-InstructBLIP 的提示稳健性增加所示。

无声视频：表 2 评估了域外视频任务的 X-InstructBLIP 。我们将性能与依赖于冻结或部分冻结LLMs的突出基线进行比较，并在视频问答 （VQA） 上显示出相当或改进的性能。但是，由于指令的性质

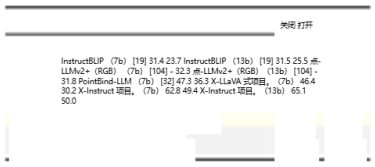


表 1：Modelnet40 [87] 测试集上的零样本 3D 分类。

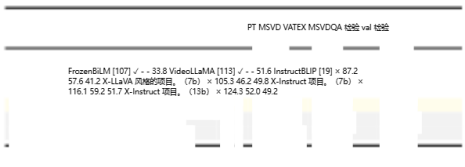


表 2：域外无声视频结果。PT 表示视频预训练阶段。

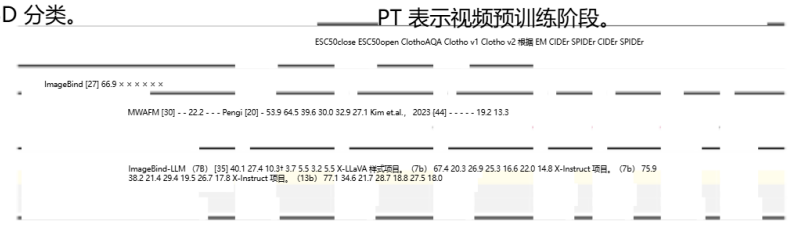


表 3：域外音频定量结果。

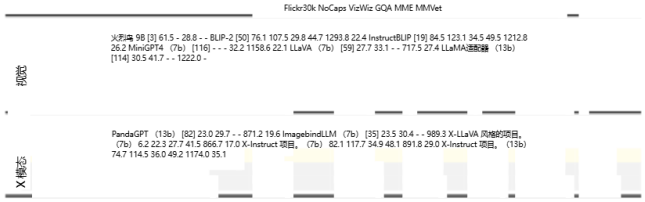


表 4：域外图像定量结果。

单模态定量结果。带下划线的数字表示域内评估。粗体表示最佳的零样本性能。蓝色表示第二好的零点性能。紫色表示独立进行的评估。用 7b 和 13b 表示的模型表示底层LLM大小。灰色阴影行对应于 X-InstructBLIP 变体，黄色对应于 LLaVA 样式 [59] 模型等效项。字幕的 CIDEr 分数 [89] 报告，音频字幕的 SPIDEr [61] 分数和 QA 和分类任务的 Top-1 准确性。†表示宽松的精确匹配指标，其中 Ground Truth 是预测的子字符串。

setup, X-InstructBLIP is tuned on other QA tasks, thus having an advantage over VideoLLaMA [113] and FrozenBiLM [107] even though it lacks video pre-training (PT). As we show in the supplement, the Video Q-Former component of X-InstructBLIP, initialized with the Image Q-Former's weights, reaches convergence in performance remarkably fast, within about 1,000 iterations. For context, VideoLLaMA is pre-trained on the entire WebVideo dataset of 2 million videos, in addition to BLIP-2 image pre-training. FrozenBiLM, on the other hand, undergoes two epochs of training on the larger WebVideo10M.

Cross-Modal Joint Reasoning: Despite each modality projection being trained individually, X-InstructBLIP shows strong joint modality reasoning, particularly under the X-Instruct Projection setting. Table 5 demonstrates X-Instruct's capability to reason jointly over video (V) and audio (A). Notably, X-Instruct Proj. (7b) is capable of synergizing inputs, displaying an improvement in performance compared to utilizing a single modality when the model is cued with different modalities both in MusicAVQA [49] and VATEX [95]. However, this is not the case for X-LLaVA-style Projection which exhibits the same or lower performance under such a cross-modal setting.

Cross-Modal Discriminative Reasoning We assess X-InstructBLIP in executing discriminatory reasoning across different modalities using our newly introduced `DisCRn` benchmark, detailed in Section 4.2. We frame the problem as a realistic open generation problem. The LLM is prefixed with the instruction:

> You are given two inputs. Select exactly one of the two by reference to its relative position (first or second, left or right) that best answers the question.

	Music AVQA *test*				VATEX *test*			
	A	V	A+V	Δ	A	V	A+V	Δ
X-LLaVA-style Proj. (7b)	36.4	34.8	36.4	0.0	4.9	46.2	35.5	-10.7
X-Instruct Proj. (7b)	13.4	27.2	28.1	1.3	6.7	59.2	**60.9**	1.7
X-Instruct Proj. (13b)	22.7	43.5	44.5	1.0	6.1	52.0	58.2	**6.2**

Table 5: Emergent Joint (A)udio-(V)ideo Reasoning. Δ denotes the difference between joint modality and best single modality score.

	A-V	Img-3D
Caption Baseline (7b)	30.8	41.8
X-LLaVA-style Proj. (7b)	**47.1**	41.4
X-Instruct Proj. (7b)	34.0	48.1
X-Instruct Proj. (13b)	45.5	**48.8**

Table 6: `DisCRN` evaluation.

In prompting X-Instruct Proj. (7b) we found that using a Q-Former captioning prompt different from the comparative prompt provided to the LLM model induces a more general representation that was more applicable for the comparative task, as such we employ this approach for the results in Table 6. This is likely due to the lack of comparative data in fine-tuning since each modality Q-Former is trained separately. Future work can explore the effect of different prompts conditioned on different parameters in the instruction-aware training setup (e.g. data, templates, joint training, and LLM partial or full optimization).

设置中，X-InstructBLIP 会针对其他 QA 任务进行调整，因此比 VideoLLaMA [113] 和 FrozenBiLM [107] 具有优势，即使它缺乏视频预训练 （PT）。正如我们在补充中所示，X-InstructBLIP 的视频 Q-Former 组件，使用图像 Q-Former 的权重进行初始化，在大约 1,000 次迭代内以非常快的速度达到性能收敛。对于上下文，除了 BLIP-2 图像预训练外，VideoLLaMA 还在包含 200 万个视频的整个 WebVideo 数据集上进行了预训练。另一方面，FrozenBiLM 在更大的 WebVideo10M 上经历了两个 epoch 的训练。

跨模态联合推理：尽管每种模态投影都是单独训练的，但 X-InstructBLIP 表现出很强的联合模态推理，尤其是在 X-Instruct Projection 设置下。表 5 演示了 X-Instruct 对视频 （V） 和音频 （A） 进行联合推理的能力。值得注意的是，X-Instruct 项目。（7b） 能够协同输入，在 MusicAVQA [49] 和 VATEX [95] 中，当模型使用不同的模态进行提示时，与使用单一模态相比，性能有所提高。但是，对于 X-LLaVA 样式投影，情况并非如此，它在这种跨模态设置下表现出相同或更低的性能。

跨模态判别推理 我们使用我们新引入的 DisCRn 基准评估 X-InstructBLIP 在不同模态中执行判别推理的情况，详见第 4.2 节。我们将问题定义为现实的开放发电问题。以LLM指令为前缀：

> 您将获得两个输入。根据两者的相对位置（第一个或第二个、左或右）选择最能回答问题的选项之一。

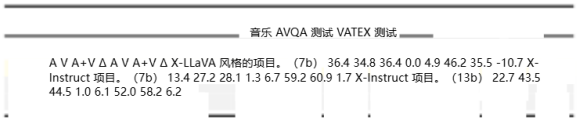


表 5：紧急关节 （A）udio-（V）ideo 推理。Δ 表示联合模态和最佳单模态评分之间的差异。

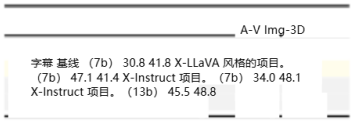


表 6：DisCRN 评估。

在提示 X-Instruct Proj.（7b） 我们发现，使用与提供给LLM模型的比较提示不同的 Q-Former 字幕提示会诱导更普遍的表示，更适用于比较任务，因此我们对表 6 中的结果采用这种方法。这可能是由于微调中缺乏比较数据，因为每种模态 Q-Former 都是单独训练的。未来的工作可以探索在指令感知训练设置（例如数据、模板、联合训练以及LLM部分或全部优化）中以不同参数为条件的不同提示的效果。

For the video-audio comparison, we select two frames for each modality to allow for a more balanced generation influence.

To benchmark our model's capabilities, we incorporate a robust captioning baseline by substituting the query outputs with captions corresponding to the modalities using the Vicuna 7b model. For images, 3D, and video modalities, we elicit captions by prompting InstructBLIP [19] to `Describe the image/video.` For 3D, a randomly chosen rendering view of the point cloud is provided to InstructBLIP. For video we follow [19] and sample four frames and concatenate their output representations as input to the model. For audio we use WavCaps [100].

While the X-InstructBLIP framework produces models that outpefrom the strong captioning baseline by a significant margin, there is no conclusive remark on which of the two projection types is more suitable for cross-modal discriminative reasoning[6]. X-LLaVA Proj. outperforms X-Instruct Proj. on Audio-Video, likely due to its stronger Audio QA performance also reported in Table 3. For image-3D the opposite is true, signifying the intuitive result that the individual modality performance plays a role in cross-modal reasoning abilities. It is worth noting, however, that X-Instruct Proj. exhibits the ability to switch from discriminative to joint reasoning, by either discriminating or combining the inputs to generate a response. As seen in Table 5 this is not the case for X-LLaVA style projections, suggesting that the instruction aware representations might prime the LLM to respond more aptly to the task in question.

5.3 Ablations

Prefix Effect: We explore the effect of prefixing the modality input with a modality specific prefix in Table 7. We compare performance of X-Instruct Proj. (7b) with X-Instruct Proj.$_{\text{no-prefix}}$ which is trained similarly to X-Instruct Proj. (7b), with the distinction that the modality type is not prepended to the modality's LLM input tokens before feeding into the LLM for training and inference. In both audio and 3D single modality tasks removing the prefix consistently hurts the performance. This improvement is likely due to the fact that the Q-Former is relieved from the extra burden to encode the type of modality and instead reserves bandwidth for semantic information. Including the prefix also allows the model to learn to combine modalities better as shown by the improved performance over the single modality for MusicAVQA and VATEX. Initially, it was theorized that the model's inability to differentiate tokens corresponding to each modality, treating them instead as a continuous stream, might be the cause for this behavior. However, the results from the image-3D cross-modal reasoning task where the prefix-less model outperforms the prefixed one by 10 points challenge this view. It appears that the inclusion of cues may be prompting the model to encode modality-specific information, which is beneficial in joint reasoning scenarios. This specialized encoding does not, however, prime the model

[6] It is worth noting that using a small sub-sample of the data we observed that the task is prompt sensitive, mainly in the language only setting. We leave it to future work to systematically evaluate the model's ability on the task based on different prompts and in-context examples.

对于视频-音频比较，我们为每种模态选择两帧，以实现更平衡的生成影响。

为了对模型的能力进行基准测试，我们通过使用 Vicuna 7b 模型将查询输出替换为与模态相对应的标题来整合强大的字幕基线。对于图像、3D 和视频模态，我们通过提示 InstructBLIP [19] 来描述图像/视频来引出字幕。对于 3D，将向 InstructBLIP 提供随机选择的点云渲染视图。对于视频，我们按照 [19] 对 4 帧进行采样，并将它们的输出表示连接起来作为模型的输入。对于音频，我们使用 WavCaps [100]。

虽然 X-InstructBLIP 框架生成的模型明显优于强字幕基线，但对于两种投影类型中的哪一种更适合跨模态判别推理，没有决定性的评论。X-LLaVA 项目。优于 X-Instruct Proj。在音频-视频上，可能是由于其更强的音频 QA 性能也在表 3 中报告。对于 image-3D 来说，情况正好相反，它表示个人模态性能在跨模态推理能力中起着重要作用的直观结果。然而，值得注意的是，X-Instruct Proj.表现出从判别式推理切换到联合推理的能力，通过判别或组合输入来生成响应。如表 5 所示，X-LLaVA 风格的投影并非如此，这表明指令感知表示可能会促使 更LLM恰当地响应所讨论的任务。

5.3 消融

前缀效应：我们在表 7 中探讨了在模态输入前加上模态特定前缀的效果。我们比较了 X-Instruct Proj 的性能。（7b） 使用 X-Instruct Proj.no-prefix，其训练方式与 X-Instruct Proj 类似。（7b），区别在于模态类型在输入LLM用于训练和推理之前没有添加到模态的LLM输入标记前面。在音频和 3D 单模态任务中，删除前缀始终会损害性能。这种改进可能是由于 QFormer 减轻了编码模态类型的额外负担，而是为语义信息保留了带宽。包含前缀还允许模型学习更好地组合模态，如 MusicAVQA 和 VATEX 的单一模态相比性能的改进所示。最初，理论上认为，该模型无法区分对应于每种模态的标记，而是将它们视为连续流，这可能是造成这种行为的原因。然而，图像 3D 跨模态推理任务的结果，其中无前缀模型的性能比有前缀的 1 模型高出 10 个百分点，这挑战了这一观点。看起来，包含线索可能会促使模型对特定于模态的信息进行编码，这在联合推理场景中是有益的。但是，这种专门的编码并不会启动模型

[6] 值得注意的是，使用一小部分数据子样本，我们观察到该任务对提示敏感，主要是在仅语言设置中。我们留给未来的工作，根据不同的提示和上下文中的示例系统地评估模型在任务上的能力。

Modality	Task	X-Instruct Proj.	X-Instruct Proj.$_{\text{no-prefix}}$
3D	Modelnet40 *close*	**62.8**	60.9
	Modelnet40 *open*	**49.4**	46.7
Audio	ESC50 *close*	**75.9**	67.5
	ESC50 *open*	**38.2**	36.0
	ClothoAQA	**15.4**	9.9
	Clotho v1/v2	**29.4**/**26.7**	26.9/24.5
Audio+Video	MusicAVQA (A/V/A+V/Δ)	**13.4**/27.2/**28.1**/1.3	8.9/**27.3**/22.3/-5.0
	VATEX (A/V/A+V/Δ)	6.7/59.3/**60.9**/1.7	**6.8**/**59.5**/58.3/-1.2
	DisCRn	**34.0**	31.4
Image+3D	DisCRn	48.1	**57.7**

Table 7: Ablation: Prefix Effect

	ESC50$_{\text{close}}$	ESC50$_{\text{open}}$	ClothoAQA	Clotho v1	Clotho v2
X-Instruct Proj. (7b)	**75.9**	**38.2**	**15.4**	**29.4**	26.7
X-Instruct Proj. (7b)$_{\text{no-init}}$	70.0	37.8	11.9	29.3	**27.4**

Table 8: Out-Domain Audio Quantitative Results.

to recognize and process characteristics usually associated with other modalities, required for enhanced performance in contrastive tasks. The underlying rationale is that the language model, already tuned to generate modality-relevant outputs, leads the Q-Former to primarily receive feedback on modality-specific generation during training, also accounting for the improvements in single modality.

BLIP-2 Initialization We also explore the effectiveness of the BLIP-2 initialization by training the audio Q-Former in X-Instruct Proj. (7b) using a random initialization approach denoted as X-Instruct Proj. (7b)$_{\text{no-init}}$. Table 8 demonstrates the benefits of this prior, indicating that it's possible to integrate new modalities into our framework without extensive pre-training, since from the modalities considered, audio is the least likely to benefit from image-text pre-training. Future research should delve into the effects of modality-specific pre-training, as they are outside of our scope. The most significant improvement is observed in question answering, indicating that BLIP-2 weight initialization appears to enhance instruction awareness more than direct audio-language alignment, corroborated by the gap in closed vocabulary classification performance.

6 Conclusion

This study introduces X-InstuctBLIP, a scalable framework for independently aligning the representation of multiple modalities to that of a frozen LLM demonstrating competitive results compared to leading methods across all addressed modalities. The framework exhibits emergent cross-modal reasoning, despite separate modality training. To test this emergent ability a new cross-modal discriminatory reasoning task `DisCRn` is introduced to show that the framework yields models that can outperform strong captioning baselines across all four examined modalities. Despite the effectiveness of the method, the task remains an open challenge. We also find complexities and unanswered questions within each modality, paving the way for future explorations across and within modalities.

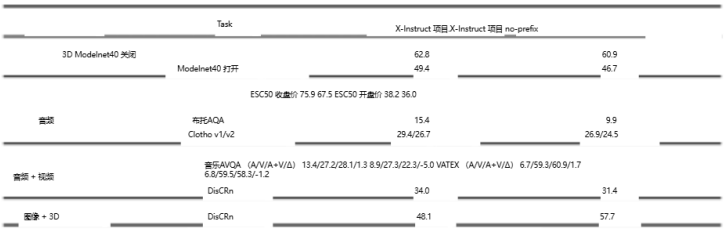


表 7：消融：前缀效应

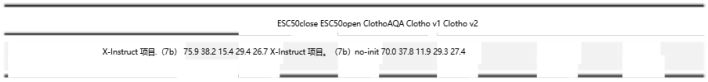


表 8：域外音频定量结果。

识别和处理通常与其他模式相关的特征，这些特征是增强对比任务性能所必需的。基本原理是，语言模型已经被调整为生成与模态相关的输出，导致 Q-Former 在训练期间主要接收有关特定模态生成的反馈，同时也考虑了单一模态的改进。

BLIP-2 初始化我们还通过在 X-Instruct Proj 中训练音频 Q-Former 来探索 BLIP-2 初始化的有效性。（7b） 使用表示为 X-Instruct Proj 的随机初始化方法。（7b）不初始化。表 8 展示了这种先验的好处，表明可以在不进行大量预训练的情况下将新模态集成到我们的框架中，因为从所考虑的模态来看，音频最不可能从图像文本预训练中受益。未来的研究应该深入研究特定模式预训练的效果，因为它们超出了我们的研究范围。在问答中观察到最显着的改进，表明 BLIP-2 权重初始化似乎比直接的音频语言对齐更能增强指令意识，封闭词汇分类性能的差距证实了这一点。

6 总结

本研究介绍了 X-InstuctBLIP，这是一个可扩展的框架，用于独立地将多种模态的表示与冷冻LLM的表示对齐，与所有所解决模态的领先方法相比，显示出有竞争力的结果。尽管进行了单独的模态训练，但该框架表现出了涌现的跨模态推理。为了测试这种涌现能力，引入了一个新的跨模态判别推理任务 DisCRn，以表明该框架产生的模型可以在所有四种检查模态中胜过强字幕基线。尽管该方法有效，但这项任务仍然是一个公开的挑战。我们还在每种模态中发现了复杂性和未解之谜，为未来跨模态和模态内部的探索铺平了道路。

Supplementary Material for X-InstructBLIP: A Framework for Aligning Image, 3D, Audio, Video to LLMs and its Emergent Cross-modal Reasoning

Artemis Panagopoulou[1⋆], Le Xue[2,⋆⋆], Ning Yu[2,⋆⋆], Junnan Li[2], Dongxu Li[2], Shafiq Joty[2], Ran Xu[2], Silvio Savarese[2], Caiming Xiong[2], and Juan Carlos Niebles[2]

[1] University of Pennsylvania
artemisp@seas.upenn.edu
[2] Salesforce AI Research

1 Data Generation

1.1 Instruction Tuning Data Augmentation

For the audio and 3D modalities, the available range of tasks for instruction tuning is relatively limited. To address this challenge we follow a common paradigm in the literature [101] and extract question-answer pairs from captioning datasets, specifically from captions consisting of 10 words or more. Figure 1 delineates the procedure to automatically generate question answering data from captioning datasets. The `google/flan-t5-xxl` model from huggingface-transformers is employed, and is prompted to produce candidate single-word answers based on the caption. Subsequently, the model is tasked with generating a relevant question using the answer and context as inputs. The method of round-trip-consistency [75] is utilized to retain only those question-answer pairs that align with the context. This alignment is verified by ensuring that the Levenshtein partial similarity between the predicted and initial answers is greater than 0.90, calculated using the Fuzzy Wuzzy Python package. Subsequently, we apply a string matching post-processing to filter out instances that do not conform to the prescribed format. As a result, 250,070/1,157 suitable training/validation examples are derived from an initial 661,576/5,000 3D-caption samples from Cap3D [64], and 24,156/1,653 training/validation examples are derived from 38,695/1,900 original audio-caption samples from AudioCaps [43]. Moreover, for 3D data, it is imperative to ensure that the question-answer pairs do not allude to color. This is due to the fact that the 3D encoder does not capture color characteristics. To achieve this, the language model is directed to reformulate the captions by omitting any references to color, prompted as: `Rewrite the sentence {caption} by eliminating any color mentions`, prior to implementing

⋆ Work done while interning at Salesforce Research
⋆⋆ Equal mentorship contribution.

X-InstructBLIP 的补充材料：图像、3D、音频、视频对齐 LLMs及其紧急跨模态推理的框架

Artemis Panagopoulou、Le Xue、Ning Yu、Junnan Li、Dongxu Li、Shafiq Joty、Ran Xu、Silvio Savarese、Caim Xiong 和 Juan Carlos Niebles

[1] 宾夕法尼亚大学
artemisp@seas.upenn.edu
[2] Salesforce AI 研究

1 数据生成

1.1 指令调优数据增强

对于音频和 3D 模态，指令调优的可用任务范围相对有限。为了应对这一挑战，我们遵循文献中的常见范式 [101] 并从字幕数据集中提取问答对，特别是从由 10 个或更多单词组成的字幕中提取问答对。图 1 描述了从字幕数据集自动生成问答数据的过程。采用了来自 huggingface-transformers 的 google/flan-t5-xxl 模型，并提示根据标题生成候选单字答案。随后，模型的任务是使用答案和上下文作为输入生成相关问题。往返一致性的方法 [75] 用于仅保留那些与上下文一致的问答对。通过确保预测答案和初始答案之间的 Levenshtein 部分相似性大于 0.90（使用 Fuzzy Wuzzy Python 包计算）来验证这种对齐方式。随后，我们应用字符串匹配后处理来过滤掉不符合规定格式的实例。结果，250,070/1,157个合适的训练/验证样本来自Cap3D [64]的初始661,576/5,000个3D字幕样本，24,156/1,653个训练/验证样本来自AudioCaps的38,695/1,900个原始音频字幕样本[43]。此外，对于 3D 数据，必须确保问答对不暗示颜色。这是因为 3D 编码器无法捕获颜色特征。为此，语言模型被指示通过省略对 color 的任何引用来重新构建标题，提示为：在实现之前，通过消除任何颜色提及来重写句子 {caption}

---

[*] 在 Salesforce Research 实习期间完成的工作
[**] 平等的导师贡献。

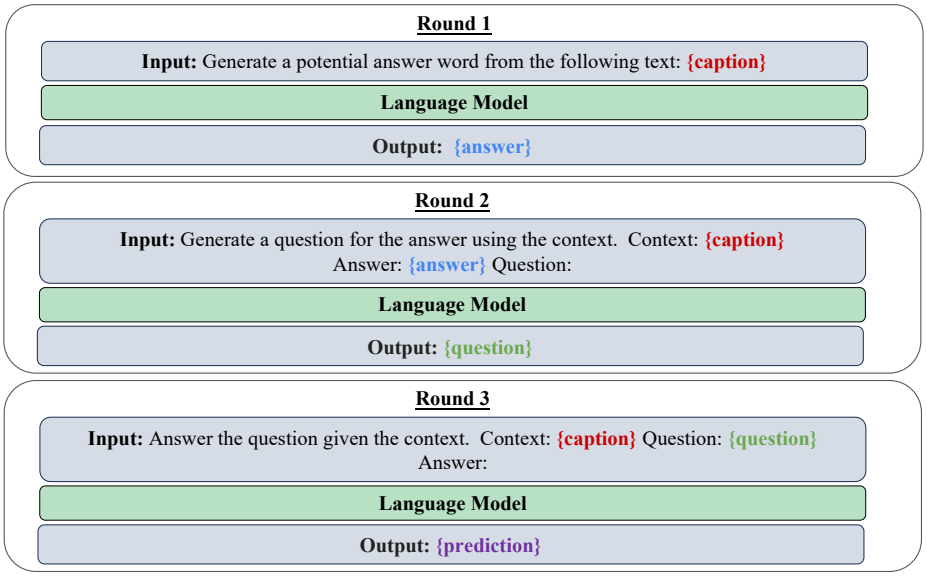


Fig. 1: Round-Trip-Consistency Prompting for QA Datasets in 3D and Audio.

the round-trip-consistency check. A short human evaluation on 50 samples for each modality shows that 90% of the generated audio and 82% of the 3D data is correct. Table 1 presents a random sample of the generated data and table 2 provides an overview of the datasets's distribution statistics. It is worth noting that the error cases are typically due to non-sensical questions, rather than wrong answers. For example the following pairs were marked as non-sensical: *What is the sewing machine running at? speed*, *What does the steam whistle do? hisses*, *What is the 3D model of a brick wall with holes and stacked cubes, resembling? elements*, and *What is the hat with? pattern*.

1.2 Cross-modal Discriminative Reasoning Data Generation

To assess the cross-modal reasoning capabilities of X-InstructBLIP, we devised a unique task that repurposes existing captioning datasets, specifically focusing on data representable in multiple modalities. We chose the AudioCaps [43] validation dataset and reserved a subset of 5k examples from Cap3D [64] as our validation dataset, ensuring that the 3D projection is not exposed to this subset during the training phase in either captioning or 3DQA settings.

The audio data from AudioCaps originates from Youtube videos, allowing us to download the corresponding video files using their YouTube IDs. For Cap3D, we employed the associated point clouds and randomly selected one rendered image from the available eight view angles.

A depiction of the data generation procedure, also outlined in the main text, is provided in Figure 2. During the evaluation, we maintain a balance, ensuring

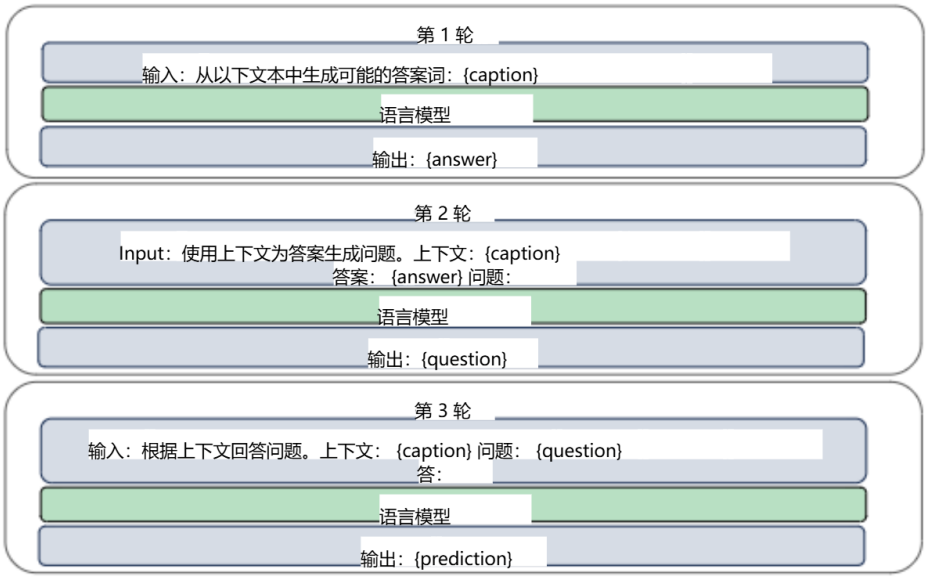


图 1：3D 和音频中 QA 数据集的往返一致性提示。

往返一致性检查。对每种模态的 50 个样本进行简短的人工评估表明，生成的音频中有 90% 和 3D 数据有 82% 是正确的。表 1 显示了生成数据的随机样本，表 2 提供了数据集分布统计数据的概述。值得注意的是，错误情况通常是由于无意义的问题，而不是错误的答案。例如，以下对被标记为无意义：缝纫机在什么位置运行？速度，蒸汽哨子是做什么的？嘶嘶地说道，带有孔洞和堆叠立方体的砖墙的 3D 模型是什么样子的？元素和帽子用什么？模式。

1.2 跨模态判别推理数据生成

为了评估 X-InstructBLIP 的跨模态推理能力，我们设计了一项独特的任务，重新利用现有的字幕数据集，特别关注以多种模态表示的数据。我们选择了 AudioCaps [43] 验证数据集，并从 Cap3D [64] 中保留了 5k 样本的子集作为我们的验证数据集，确保在训练阶段，无论是字幕还是 3DQA 设置，3D 投影都不会暴露在这个子集中。

AudioCaps 的音频数据来自 Youtube 视频，允许我们使用他们的 YouTube ID 下载相应的视频文件。对于 Cap3D，我们采用了相关的点云，并从可用的 8 个视角中随机选择一个渲染图像。

图 2 中提供了数据生成过程的描述，也包含在正文中。在评估过程中，我们保持平衡，确保

	Caption	Question	Answer
Audio	A woman speaks while types a keyboard;	What is the woman typing on?	Keyboard
	A man are talking while multiple dogs are barking around them;	What is the dog doing?	Barking
	A man speaks and a crowd applauds, he continues talking;	What does the crowd do after the man speaks?	Applauds
	A plane flies in the distance as a man speaks and metal clinks.	What does the metal do?	Clinks
3D	A 3D model of a wooden chair and stool with a chained bucket on it	What is on the stool?	Bucket
	A 3D model of a moss-covered stone, resembling a leaf, paper map, and rock	What is covering the stone?	Moss
	A balloon with a string attached, featuring a teddy bear and a cat face on it	What is the object with a string attached?	Balloon
	A 3D model of various food items,including an oyster, a piece of fruit, and different forms of eggs.	What is the food item that is a shellfish?	oyster

Table 1: Automatically Generated QA examples from Captioning Data.

Dataset	`AudioCapsQA`		`Cap3DQA`	
	train	*val*	*train*	*val*
Size	24,156	1,274	250,070	1,157
Distinct Questions	10,010	810	67,001	953
Distinct Answers	1,636	374	4,555	451
Avg. Question Length (words)	6.0	6.1	6.8	7.0
Vocabulary Size	2,951	723	12,771	1,022

Table 2: QA Generated Dataset Statistics

each option (A or B) serves as the ground truth 50% of the time. Given that this problem is structured as an open vocabulary generation task, we expanded the ground truth answer space to include synonyms and equivalent expressions, such as `[{answer modality}, left, 1st, 1, first, input 1, entity 1, object 1, input A, entity A, object A]` and `[{answer modality}, right, 2nd, second, input 2, entity 2, object 2, input B, entity B, object B]`, corresponding to whether the first or the second input is the ground truth. The human performance on a subsample of 100 examples of the dataset is 90%. Table 3 provides an overview of the datasets's distribution statistics.

Dataset	Audio-Video	Video-3D
Total Size	8,802	28,173
Number of Distinct Questions	1,212	3,100
Average Question Length	5.8 words	6.6 words
Question Vocabulary Size	701 words	1,272 words

Table 3: `DisCRn`: Discriminative Cross-modal Reasoning Dataset Statistics

	标题	问题解答	
音频	一个女人在打键盘时说话; 那个女人在做什么?		键盘
	一名男子正在交谈，而多只狗在他们周围吠叫;	狗在做什么？叫	
	一个人说话，一群人鼓掌，他继续说话;男人说话后，人群在做什么?		赞扬
	一架飞机从远处飞过，一个男人说话，金属叮当作响。	金属有什么作用？叮当声	
3D	木椅和凳子的 3D 模型，上面有一个带链的水桶	凳子上有什么？桶	
	长满苔藓的石头的 3D 模型，类似于树叶、纸质地图和岩石	什么覆盖了石头?	Moss
	一个系有绳子的气球，上面有一只泰迪熊和一张猫脸	附加了字符串的对象是什么?	气球
	各种食物的 3D 模型，包括牡蛎、一块水果和不同形式的鸡蛋。	什么是贝类的食物?	牡蛎

表 1：从字幕数据自动生成的 QA 示例。

数据集	AudioCapsQA	Cap3DQA	
火车	val	火车	val
面积	24,156	1,274	250,070 1,157
不同问题	10,010	810	67,001 953
不同的答案	1,636	374	4,555 451
平均问题时长（字数）	6.0	6.1	6.8 7.0
词汇量	2,951	723	12,771 1,022

表 2：QA 生成的数据集统计数据

确保每个选项（A 或 B）在 50% 的时间内作为基本实况。鉴于这个问题是作为一个开放词汇生成任务构建的，我们扩展了真实答案空间，包括同义词和等效表达式，例如 [{answer modality}，left，1st，1，first，input 1，entity 1，object 1，input A，entity A，object A] 和 [{answer modality}，right，2nd，second，input 2，entity 2，object 2，输入 B、实体 B、对象 B]，对应于第一个输入还是第二个输入是真实值。人类在数据集的 100 个样本的子样本上的表现为 90%。表 3 提供了数据集的分布统计数据的概述。

数据集	音视频	视频 - 3D	
总面积	8,802	28,173	
不同问题的数量	1,212	3,100	
平均问题时长	5.8 字	6.6 字	
问题词汇量	701 个单词	1,272 个单词	

表 3：DisCRn：判别性跨模态推理数据集统计

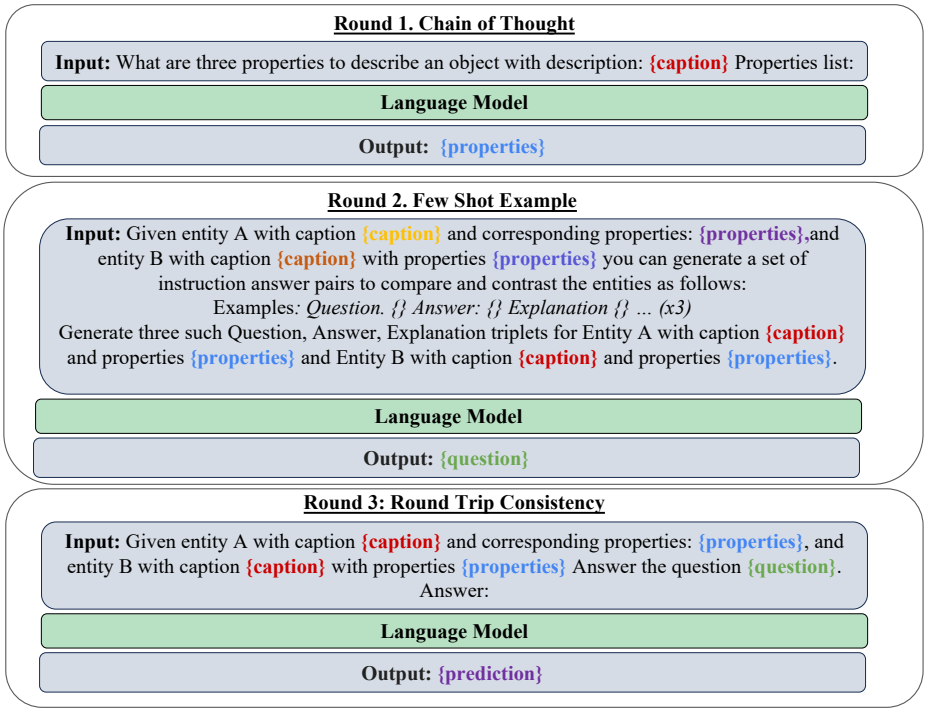


Fig. 2: Cross-modal Discriminative Reasoning Dataset Generation Framework

2 Video Q-Former Fine-Tuning Versus Image Initialization

To explore the impact of further training the Image Q-Formers on video data, Table 4 presents the results of evaluating video tasks using the weights from the Image Q-Formers. It is evident that training on video data enhances performance. However, it's worth noting that the Video Q-Formers reach convergence at an earlier stage (15k and 5k iterations for Vicuna7b and Vicuna13b, respectively). This is likely because the Q-Formers have already achieved semantic understanding during the image alignment phase, requiring minimal additional training to capture the nuances of sequential video projections. The higher drop in performance in MSVD [10] captioning compared to VATEX [95] is likely due to the closer similarity between MSVD and the held-in MSRVTT [103] dataset distribution. There is a notably lower drop in performance for Video QA tasks, owing to the more constraint nature of the task - training on videos would not significantly increase the performance since the answer is typically constrained in one frame [8], and as such processing that frame would be almost equivalent to processing it in the image. The improvement probably stems from identifying the answer across a longer sequence of query tokens.

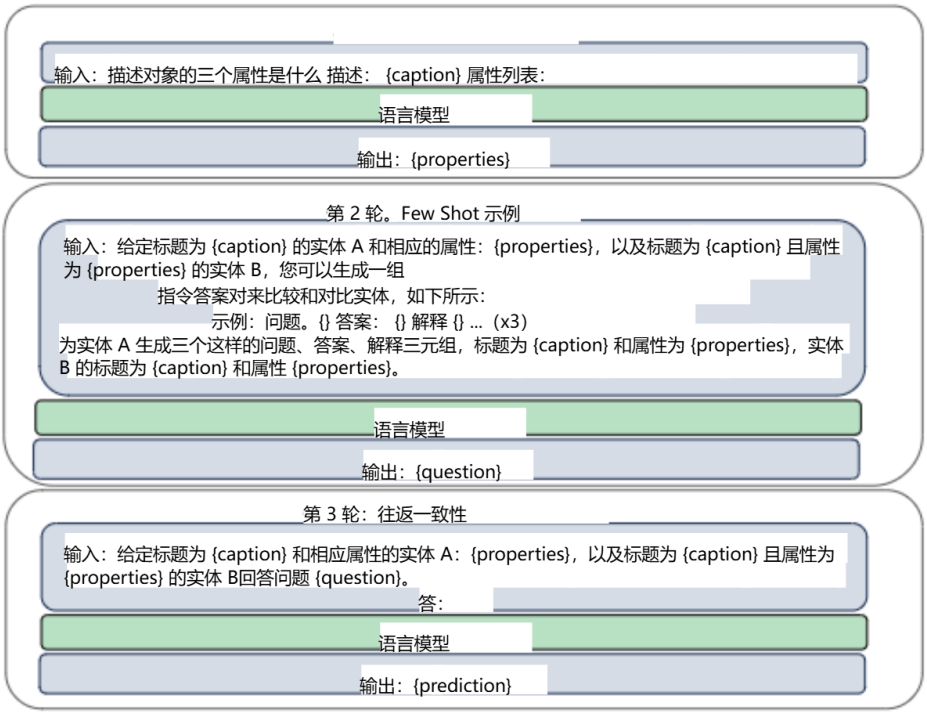


图 2：跨模态判别推理数据集生成框架

2 视频 Q-Former 微调与图像初始化

为了探索进一步训练图像Q-Formers对视频数据的影响，表4显示了使用Image Q-Formers的权重评估视频任务的结果。很明显，对视频数据进行训练可以提高性能。然而，值得注意的是，视频 Q-Formers 在早期阶段达到收敛（Vicuna7b 和 Vicuna13b 分别为 15k 和 5k 迭代）。这可能是因为 Q-Formers 在图像对齐阶段已经实现了语义理解，只需要最少的额外培训即可捕捉顺序视频投影的细微差别。与 VATEX [95] 相比，MSVD [10] 字幕的性能下降幅度更大，这可能是由于 MSVD 与保留的 MSRVTT [103] 数据集分布之间的相似性更接近。由于任务的约束性质更强，视频 QA 任务的性能明显下降 - 视频训练不会显著提高性能，因为答案通常被限制在一帧 [8] 中，因此处理该帧几乎等同于在图像中处理它。这种改进可能源于在较长的查询标记序列中确定答案。

3 In-Domain Evaluations

Table 6 presents in-domain performance for a sample of datasets seen in training across all four modalities. It's important to clarify that when we refer to 'in-domain,' we are specifically referring to datasets that were sampled during the training process. However, it's crucial to note that this does not constitute explicit fine-tuning, as there is no guarantee that the Q-Former has encountered the entirety of the dataset during its training.

4 Prompt Robustness

Table 5 compares performance between InstructBLIP (7b) and X-Instruct Proj. (7b) on NoCaps [1], using prompts not encountered in the optimization of either model. While X-InstructBLIP exhibits some performance variability, it maintains more than half the standard deviation of InstructBLIP. This variance can be attributed to the expanded vocabulary in our templates, allowing the Q-Former to better associate an instruction with a specific task. For example, in the case of prompt P2: *"Provide a recap of what is happening in the picture"*, InstructBLIP maintains high performance as it closely resembles an in-domain prompt *"Use a few words to illustrate what is happening in the picture"*. Note that the performance drop in InstructBLIP is mostly attributed to the language model resorting to generating longer descriptions when the Q-Former outputs have not captured the task, resulting in hallucinations in later stages of generation.

5 Training Details

Prior to encoding, raw inputs undergo standardized pre-processing: images are resized to 224×224 resolution with random cropping and normalization; audio files undergo mono conversion and filter bank pre-processing followed by normalization as in [13] over two 5-second frames; videos are uniformly sampled to 5 frames subject to the same pre-processing as images, and 3D point clouds are uniformly sampled and normalized to 8k points as in [78, 106]. All modality Q-Formers are pre-initialized with BLIP-2 [19] stage-1 weights except for the video

	MSVD *test*	VATEX *val*	MSVD QA *test*
X-LLaVA Style Proj. (7b)	105.3	46.2	49.8
X-LLaVA Style Proj. (7b)[*image*]	16.4	10.7	23.2
X-Instruct Proj. (7b)	116.1	**59.2**	**51.7**
X-Instruct Proj. (7b) [*image*]	42.4(↓73.7)	28.1(↓30.1)	39.7(↓12.0)
X-Instruct Proj.$_{\text{no-prefix}}$ (7b)[*image*]	62.0(↓56.7)	52.6(↓6.9)	38.8(↓11.7)
X-Instruct Proj. (13b)	**124.3**	52.0	49.2
X-Instruct Proj. (13b) [*image*]	78.7(↓45.6)	53.5(↑1.5)	36.0(↓13.2)

Table 4: Impact of Training Image Q-Formers on Video. Models labeled [*image*] utilize the Image Q-Former for video alignment.

3 域内评估

表 6 显示了在所有四种模态的训练中看到的数据集样本的域内性能。需要澄清的是，当我们提到“域内”时，我们特指在训练过程中采样的数据集。但是，请务必注意，这并不构成显式微调，因为无法保证 Q-Former 在训练期间遇到了整个数据集。

4 迅速稳健

表 5 比较了 InstructBLIP （7b） 和 X-Instruct Proj 之间的性能。（7b） 在 NoCaps [1] 上，使用在任一模型优化中未遇到的提示。虽然 X-InstructBLIP 表现出一些性能变化，但它保持了 InstructBLIP 标准差的一半以上。这种差异可以归因于我们模板中扩展的词汇表，使 QFormer 能够更好地将指令与特定任务相关联。例如，在提示 P2 的情况下： “Provide a recap of what is happening in the picture”，InstructBLIP 保持高性能，因为它与域内提示 “Use a few words to illturify what is happening in the picture” 非常相似。请注意，InstructBLIP 中的性能下降主要归因于语言模型在 Q-Former 输出未捕获任务时求助于生成更长的描述，从而导致生成后期出现幻觉。

5 培训详情

在编码之前，原始输入经过标准化的预处理：图像大小调整为 224 × 224 分辨率，并进行随机裁剪和标准化;音频文件在两个 5 秒的帧中进行单声道转换和滤波器组预处理，然后进行归一化，如 [13] 所示;视频被均匀采样为 5 帧，并接受与图像相同的预处理，并且 3D 点云被均匀采样并标准化为 8k 点，如 [78， 106]。除视频外，所有模态 QFormer 都使用 BLIP-2 [19] stage-1 权重进行了预初始化

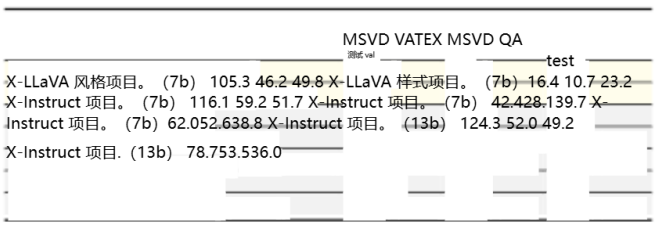


表 4：训练图像 Q-Former 对视频的影响。标记为 [image] 的模型使用 Image Q-Former 进行视频对齐。

	InstructBLIP (7b)	X-Instruct Proj. (7b)
P1	1.0	**88.0**
P2	**121.9**	109.7
P3	0.9	**54.9**
P4	5.4	**112.7**
P5	0.8	**111.5**
Avg	26.3	**83.0**
Std.	43.8	**20.8**

P1 In a few words describe the basic features of this image.
P2 Provide a recap of what is happening in the picture.
P3 I'd like to hear your interpretation of this image. What do you see?
P4 Provide a verbal snapshot of what's happening in this image.
P5 Please articulate the elements and context of this image

Table 5: Robustness to unseen prompts on NoCaps (*val*) [1].

Q-Former which is initialized from the last iteration of the corresponding image Q-Former and optimized for 15k/5k steps for the Vicuna 7b and 13b models respectively.

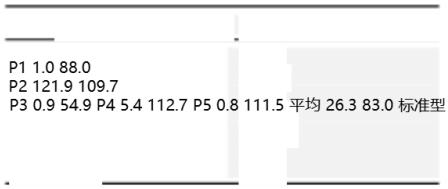
Table 7 compiles the training hyperparameters employed for each modality and model. The X-Instruct Proj. $_{\text{no-prefix}}$ variant is trained similarly to X-Instruct Proj., with the notable distinction that the modality type is not prepended to the modality's query outputs, both during training and inference. Following [19] that noted that sampling ratios play an important role in training we perform some minor modifications in the sampling ratios that we show in tables 9 and 8 are effective in improving performance. The decisions are discussed further below. It is worth noting that due to the large amount of experiments consisting of all modalities, we did not exhaust all possibilities, and there may be better training configurations. We leave this to future work to be explored.

As each modality exhibits unique characteristics, we have customized the training approach for each one. For instance, the 3D and Audio projections are trained for the maximum number of iterations specified in Table 7.

The Vicuna7b Image projection undergoes training for 735k iterations, utilizing normalized data sampling. Additionally, an extra 40k iterations are performed with the sampling ratio of COCO Captions [9] set to 3.0 while keeping the other ratios consistent with the original sampling. This adjustment leverages the clean annotations of COCO Captions, mitigating noise introduced by

	Image			3D		Video				Audio		
	OKVQA	COCO		Cap3D		MSRVTT		MSRVTT QA		AudioCaps		
	test	*val*	*test*	*val*	*qa-val*	*val*	*test*	*val*	*test*	*val*	*test*	*qa-val*
Finetuned SOTA	66.1 [21]	-	155.1 [47]	-	-	-	80.3 [102]	-	48.0 [102]	-	78.1 [14]	-
InstructBLIP (T5xl)	48.6	137.7	140.2	-	-	44.1	44.0	25.0	22.3	-	-	-
InstructBLIP (T5xxl)	47.8	139.1	140.8	-	-	41.5	47.8	25.6	21.4	-	-	-
InstructBLIP (7b)	57.3	141.0	142.3	-	-	28.1	31.1	22.1	18.7	-	-	-
InstructBLIP (13b)	56.3	139.1	141.0	-	-	36.7	37.1	24.8	20.2	-	-	-
X-LLaVA Style Proj. (7b)	28.5	126.0	118.1	126.7	39.9	55.5	53.1	41.0	41.4	44.3	46.1	53.2
X-Instruct Proj. (7b)	52.5	137.7	138.2	142.1	53.6	61.0	57.6	44.6	42.1	44.6	67.9	41.2
X-Instruct Proj. (13b)	51.9	128.2	128.7	148.8	54.9	57.7	52.2	36.4	36.1	54.2	53.7	37.4

Table 6: In-Domain performance across modalities.



P1 用几句话描述一下这张图片的基本特征。
P2 提供图片中发生的情况的回顾。
P3 我想听听你对这张图片的解释。您看到了什么?
P4 提供此图像中所发生情况的口头快照。P5 请阐明此图像的元素和背景

表 5：NoCaps 上看不见的提示的鲁棒性 （val ） [1]。

Q-Former 从相应图像 Q-Former 的最后一次迭代开始初始化，并分别针对 Vicuna 7b 和 13b 型号的 15k/5k 步进行了优化。

表 7 编译了用于每种模态和模型的训练超参数。X-Instruct 项目。variant 的训练方式与 X-Instruct Proj. 类似，但显着区别在于，在训练和推理期间，模态类型都不会添加到模态的查询输出中。在 [19] 指出采样率在训练中起着重要作用之后，我们对表 9 和表 8 中所示的采样率进行了一些细微的修改，以有效提高性能。这些决定将在下文进一步讨论。值得注意的是，由于由所有模态组成的大量实验，我们并没有穷尽所有可能性，可能会有更好的训练配置。我们把这个问题留给未来的工作去探索。

由于每种模式都表现出独特的特征，因此我们为每种模式定制了训练方法。例如，3D 和 Audio 投影针对表 7 中指定的最大迭代次数进行训练。

Vicuna7b 图像投影利用归一化数据采样进行了 735k 迭代的训练。此外，在 COCO 字幕 [9] 的采样率设置为 3.0 的情况下，执行额外的 40k 迭代，同时保持其他比率与原始采样一致。此调整利用了 COCO 字幕的干净注释，减轻了

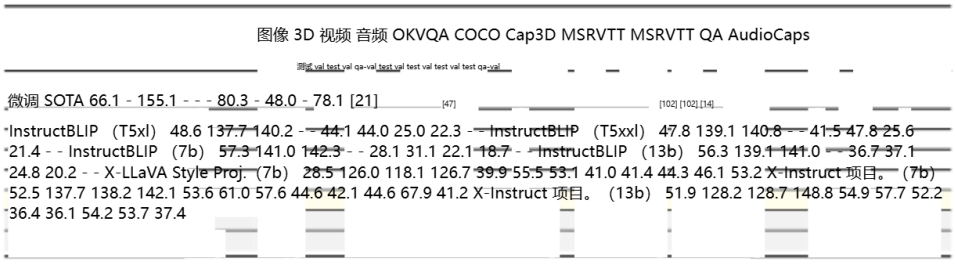


表 6：跨模式的域内性能。

larger image datasets. However, this upsampling technique is not applied to the Vicuna13b Image Q-Former, since it appears to lower out of distribution performance in non-captioning tasks as shown in table 9. It could be that due to the smaller batch size, Vicuna13b is less sensitive to noisy data, since it effectively sees less of them. In both cases, the last checkpoint from the iterations specified in Table 7 is chosen, with guidance from the COCO Captions validation dataset. Note that we optimize the Image Q-Former for 10 times more iterations that InstructBLIP. The reason is that we maintain conformity with the other Q-Formers and do not intitialize the cross-attention layers from BLIP-2 pretraining nor we allow for stage-2 training. Nevertheless, we show that with enough iterations, the cross attention layers can be learned equivalently without the need of the contrastive auxiliary losses of BLIP-2 nor stage-2 training.

The Vicuna7b video projections are initialized from the best Vicuna7b image projection and undergoes validation every 5k iterations on the MSRVTT captioning [103] dataset. The selection process involves choosing the checkpoint that precedes any drop in performance during the subsequent validation rounds even if there is a better performing checkpoint later on in training, to avoid overfitting to the MSRVTT skeletal captions. Table 8 quantitatively shows our observations. Due to the initialization of the video Q-Former with the well trained image Q-Former, the noisy captions of WebVid2M reduce the performance instead of improving it. However, this is corrected with cleaner data.

Similarly, the Vicuna13b video Q-Former is initialized from the best checkpoint of the Vicuna13b Image Q-Former and validated every 1k iterations. While we let the Vicuna7b and 13b video Q-Formers train for 15k and 25k respectively, we observe early convergence at 15k and 5k iterations likely due to the pre-initialization with the Image Q-Former. During training, 5 frames are sampled for the Vicuna7b Video Q-Former, while 4 frames are sampled for the Vicuna13b to reduce computational demands. Figure 3 shows that the video performance converges in 1k iterations on an out of domain video captioning dataset.

The best training approach for each model was empirically identified, and it is beyond the scope of the paper to rigorously analyze the reasons of the differences in training across modalities. We leave this to future work.

(7b/13b)	Image	Audio	3D	Video*
Iterations	775k/880k	65k/300k	65k/300k	15k/5k
Batch Size	64/16	64/16	128/32	32/8

Table 7: Training hyperparameters. * Video projection is initialized from Image Projection. Parameters for 7b/13b model respectively.

6 Evaluation Hyperparameters

During the evaluation of X-InstructBLIP, we adhere to a consistent set of hyperparameters, with minor variations to accommodate the distinct needs of each

较大的图像数据集。然而，这种上采样技术并不适用于 Vicuna13b Image Q-Former，因为它似乎会降低非字幕任务中的分发性能，如表 9 所示。可能是由于批处理较小，Vicuna13b 对嘈杂数据不太敏感，因为它实际上看到的数据较少。在这两种情况下，都会在 COCO Captions 验证数据集的指导下，从表 7 中指定的迭代中选择最后一个检查点。请注意，我们优化了 Image Q-Former 的迭代次数是 InstructBLIP 的 10 倍。原因是我们与其他 Q-Former 保持一致，并且不会初始化 BLIP-2 预训练中的交叉注意力层，也不允许进行阶段 2 训练。尽管如此，我们表明，通过足够的迭代，交叉注意力层可以等效地学习，而不需要 BLIP-2 或阶段 2 训练的对比辅助损失。

Vicuna7b 视频投影是从最佳 Vicuna7b 图像投影初始化的，并在 MSRVTT 字幕 [103] 数据集上每 5k 迭代一次进行验证。选择过程包括在后续验证轮次中选择性能下降之前的检查点，即使在训练后期有性能更好的检查点，以避免过度拟合 MSRVTT 骨骼字幕。表 8 定量显示了我们的观察结果。由于视频 Q-Former 的初始化使用了训练有素的图像 QFormer，因此 WebVid2M 的杂点字幕不仅没有提高性能，反而降低了性能。但是，这可以通过更清晰的数据进行更正。

同样，Vicuna13b 视频 Q-Former 从 Vicuna13b 图像 Q-Former 的最佳检查点初始化，并每 1k 次迭代验证一次。虽然我们让 Vicuna7b 和 13b 视频 Q-Former 分别训练 15k 和 25k，但我们观察到 15k 和 5k 迭代的早期收敛可能是由于图像 Q-Former 的预初始化。在训练期间，Vicuna7b Video Q-Former 采样 5 帧，而 Vicuna13b 采样 4 帧，以减少计算需求。图 3 显示，视频性能在域外视频字幕数据集上以 1k 迭代收敛。

每个模型的最佳训练方法都是根据经验确定的，严格分析不同模式训练差异的原因超出了本文的范围。我们把这个问题留给未来的工作。

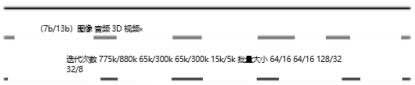


表 7：训练超参数。视频投影是从 Image Projection 初始化的。7b/13b 型号的参数。

6 评估超参数

在评估 X-InstructBLIP 期间，我们遵循一组一致的超参数，并有细微的变化以适应每个超参数的不同需求

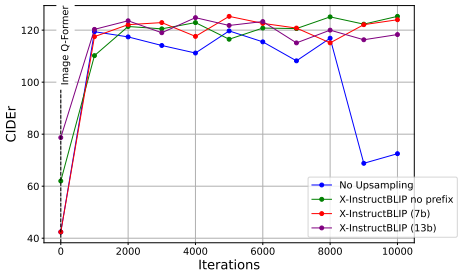


Fig. 3: CIDEr score on MSVD (out-domain) over training iterations on Video Q-Former initialized from Image Q-Former. Most performance gains are achieved within only 1000 iterations.

	MSVD *test*	VATEX *val*	MSVD QA *test*
X-Instruct Proj. (7b)	118.2	58.5	52.5
X-Instruct Proj. (7b) -upsample	73.3(↓44.9)	41.6(↓16.9)	49.1(↓3.4)

Table 8: Effect of MSRVTT Upsampling (at 10k iterations)

task. A comprehensive list of these configurations is presented in table 10. In every experiment, we utilize Beam Search for generation, setting the beam size to 5, repetition penalty and temperature equal to 1.5 and 1 respectively. For tasks involving contrastive reasoning across video-audio modalities, a balanced representation and computational efficiency are achieved by querying two frames from both video and audio. The length penalty is typically configured to 1 for long caption tasks, -1 for Visual Question Answering (VQA) tasks requiring short answers, and 0 for short caption tasks. The minimum and maximum length constraints are adapted based on the task: for captions, we maintain a range of 10 to 80; for short-answer VQA tasks, the range is set from 1 to 10; for variable-length captions, the range is between 1 and 80. In the case of the InstructBLIP baseline for video datasets, we borrow the recommended inference setup of sampling 4 frames for the captioning baselines of MSVD and VATEX with the prompt `A video that shows` and the same generation hyperparameters as X-InstructBLIP.

	Zero-Shot					In-Domain	
	Flickr30k *test*	NoCaps *val-all*	VizWiz *test-dev*	GQA *balanced test-dev*	OKVQA *test*	COCO *val*	COCO *test*
X-Instruct Proj. (7b)	82.1	117.7	34.9	48.1	52.5	137.7	138.2
X-Instruct Proj. (7b)-coco	79.4(↓2.7)	116.5(↓1.2)	34.2(↓0.7)	48.2(↑0.1)	52.3(↓0.2)	133.5(↓4.2)	134.3(↓3.9)
X-Instruct Proj. (13b)	74.7	114.5	36.0	49.2	51.9	128.2	128.7
X-Instruct Proj. (13b)+coco	83.8(↑9.1)	118.7(↑4.2)	32.0(↓4.0)	47.0(↓2.2)	44.6(↓7.3)	138.2(↑10.0)	139.0(↑10.3)

Table 9: Effect of COCO upsampling.

 0 2000 4000 6000 8000 10000 迭代

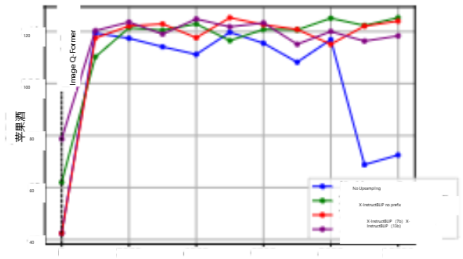


图 3：从图像 Q-Former 初始化的视频 QFormer 上训练迭代中 MSVD（域外）的 CIDEr 分数。大多数性能提升仅在 1000 次迭代中即可实现。

	MSVD VATEX MSVD QA			
	测试 val		test	
X-Instruct 项目.（7b）	118.2 58.5 52.5 X-Instruct 项目。（7b）	上采样		
73.341.649.1				

表 8：MSRVTT 上采样的效果（10k 迭代时）

任务。表 10 中列出了这些配置的完整列表。在每个实验中，我们都使用 Beam Search 来生成，将光束大小设置为 5，重复惩罚和温度分别等于 1.5 和 1。对于涉及跨视频-音频模态的对比推理的任务，通过从视频和音频中查询两帧来实现平衡的表示和计算效率。对于长字幕任务，长度损失通常配置为 1，对于需要简短答案的视觉问答 （VQA） 任务，长度损失通常配置为 -1，对于短字幕任务，长度损失通常配置为 0。最小和最大长度约束根据任务进行调整：对于字幕，我们保持 10 到 80 的范围;对于简答 VQA 任务，范围设置为 1 到 10;对于可变长度字幕，范围介于 1 到 80 之间。对于视频数据集的 InstructBLIP 基线，我们借用了推荐的推理设置，即为 MSVD 和 VATEX 的字幕基线采样 4 帧，并提示 A video that showing 和与 X-InstructBLIP 相同的生成超参数。

	零点									域内			
	Flickr30k		NoCaps		VizWiz	GQA	OKVQA		COCO				
	test	val-all	test-dev	平衡 test-dev	测试 val	test							
X-Instruct 项目（7b）	82.1	117.7	34.9	48.1	52.5	137.7	138.2	X-Instruct 项目。（7b）	-可可	79.4 (12.7)	116.5 (11.2)	34.2 (10.7)	48.2 (10.1)
52.3 (10.2)	133.5 (14.2)	134.3 (13.9)	X-Instruct 项目。（13b）	74.7	114.5	36.0	49.2	51.9	128.2	128.7	X-Instruct 项目。（13b）	+可可	83.8 (19.1)
118.7 (14.2)	32.0 (14.0)	47.0 (12.2)	44.6 (17.3)	138.2 (110.0)	139.0 (110.3)								

表 9：COCO 上采样的效果。

	Dataset	Split	Prompt	Len. Penalty	Min Len.	Max Len.
Image	Flickr30k [88]	*test*: 1,000 images	A short description	1.	10	80
	NoCaps [1]	*val*: 4,500 images *out-domain*: 1,413 images	A short description	1.	10	80
	COCO * [9]	*train*: 566,747 image-caption pairs *val*: 5,000 images *test*: 5,000 images	A short description.	1.	10	80
	VizWiz [6]	*test-dev*: 8,000 image-question pairs	based on the given image respond to {question}	-1.	1	10
	OKVQA [68]	*test*: 5,046 examples	based on the given image respond to {question} answer	-1.	1	10
	GQA [40]	*balanced test-dev*: 12,578 image-question pairs	based on the given image respond to {question}	-1.	1	10
3D	Modelnet40 [99]	*test*: 2,468 point clouds	Describe the 3d model. A 3d model of	-1.	1	3
	Modelnet40†		Describe the 3d model.	0.	10	80
Audio	Clotho [22]	*eval* (v1): 1,045 audios *val* (v2): 1,045 audios	A short description.	0.	10	80
	ClothoAQA [57]	*test*: 2,838 audio-question pairs	{question}	-1.	1	10
	ESC50 [76]	*test*: 2,000 audios	Describe the audio. An audio of	1.	10	80
	AudioCaps* [43]	*train*: 38,695 audio-caption pairs *val*: 380 audios	A short description	0.	1	80
Video	MSVD [10]	*test*: 670 images[1]	A short description	1.	10	80
	MSRVTT * [103]	*train*: 130,260 video-caption pairs *val*: 497 videos *test*: 2,990 videos	A short description	1.	10	80
	MSVD QA [101]	*test*: 13,157 video-question pairs	based on the given video respond to {question}	-1.	1	10
A+V	MusicAVQA [49]	*val*: 3,698 examples *test*: 7,402 video-question pairs	Question: {question} Answer:	-1.	1	10
	VATEX [95]	*val*: 3,000 videos	A short description	1.	10	80

Table 10: Hyperparameters used on each of the evaluation datasets. Underlined datasets are in-domain evaluations. * datasets are used for best checkpoint selection. Blue text is provided as input to the LLM but not the Q-Former.

7 Instruction Tuning Suite

Table 11 presents a comprehensive list of datasets employed in the instruction tuning process for X-InstructBLIP, accompanied by their corresponding dataset sizes. Datasets labeled with ** have been generated automatically through the round-trip-consistency procedure. Datasets marked with • indicate instances of data loss resulting from file corruption or expired links.

8 Prompt Templates

X-InstructBLIP has undergone fine-tuning using a diverse array of instruction templates, tailored to cover a wide spectrum of tasks and modalities. For reference, the specific templates corresponding to each modality can be found in the following tables: Table 12 for images, Table 13 for audio, Table 14 for 3D,

	数据集拆分	提示	Len. 罚款	Min Len.	Max Len.
图像	Flickr30k [88] 测试：1,000 张图片	简短描述 1.10 80			
	无帽 [1] VAL：4,500 张图片 out-domain： 1,413 张图片	简短描述 1.10 80			
	可可 ∗ [9] train：566,747 个图像-标题对 val：5,000 张图像 test：5,000 张图像	简短的描述。1. 10 80			
	VizWiz [6] test-dev： 基于给定图像的 8,000 个图像问题对	将 spond 转换为 {question}	-1. 1 10		
	OKVQA [68] 测试：基于给定图像的 5,046 个示例	spond 到 {question} 答案	-1. 1 10		
	GQA [40] 平衡测试开发： 12,578 基于给定图像的 image-question 对 re-	将 spond 转换为 {question}	-1. 1 10		
3D	Modelnet40 [99] 测试：2,468 个点云 描述 3D 模型。一个 3d	型号	-1. 1 3		
	型号net40†	描述 3D 模型。0. 10 80			
音频	布雷托 [22] EVAL （v1）：1,045 个音频 val （v2）：1,045 个音频	简短的描述。0. 10 80			
	ClothoAQA [57] 测试：2,838 音频问题对	{问题}	-1. 1 10		
	ESC50 [76] 测试：2,000 个音频 描述音频。一个 au-	dio 为	1. 10 80		
	音频电容 ∗ [43] train：38695 对音频-字幕 val：380 个音频	简短描述 0.1 80			
视频	MSVD [10] 测试：670 张图像1	简短描述 1.10 80			
	MSRVTT ∗ [103] train：130260 个视频-字幕对 val：497 个视频测试：2990 个视频	简短描述 1.10 80			
	MSVD QA [101] 测试：基于给定视频回复的 13,157 个视频问题对	将 spond 转换为 {question}	-1. 1 10		
A+V	MusicAVQA [49] val： 3,698 示例测试： 7,402 个视频问题对 问题： {question} An-	转弯：	-1. 1 10		
	VATEX [95] val： 3,000 个视频	简短描述 1.10 80			

表 10：每个评估数据集上使用的超参数。带下划线的数据集是域内评估。数据集用于最佳检查点选择。
蓝色文本作为 Q-Former 的输入提供，LLM而不是 Q-Former。

7 指令调整套件

表 11 显示了 X-InstructBLIP 的指令调整过程中使用的数据集的完整列表，并附有相应的数据集大小。标记为 的数据集是通过往返一致性过程自动生成的。标记为 的数据集表示由于文件损坏或链接过期而导致的数据丢失实例。

8 提示模板

X-InstructBLIP 已使用各种指令模板进行了微调，这些模板专为涵盖广泛的任务和模式而量身定制。作为参考，下表中提供了每种模态对应的具体模板：图像表 12，音频表 13,3D 表 14，

	Task	Dataset	Training Size
Image	Caption	CapFilt14M [50]	13,873,136 image-caption pairs
		Conceptual Captions 12M [9]	6,029,862 image-caption pairs•
		MS COCO Dataset [9]	566,747 image-caption pairs
		SBU Captions [74]	859,739 image-caption pairs
		Visual Genome Captions [46]	821,774 image-caption pairs
	QA	AOK VQA [79]	17,056 question-answer pairs
		OK VQA [68]	9,009 question-answer pairs
		OCR VQA [69]	1,002,146 question-answer pairs
		Visual Genome QA [46]	1,440,069 question-answer pairs
		VQAV2 [29]	658,104 question-answer pairs
	Dialogue	LLaVA150k [59]	394,276 image-instruction pairs
Audio	Caption	AudioCaps [43]	38,701 audio-caption pairs•
		WAVCaps [100]	297,341 audio-caption pairs•
	QA	AudioCaps QA**	24,158 question-answer pairs
	Classification	AudioSet balanced train [26]	14,141 labeled audios•
3D	Caption	Cap3D [64]	651,576 point cloud-caption pairs
	QA	Cap3D QA**	250,070 question-answer pairs
Video	Caption	MSRVTT [103]	130,260 video-caption pairs
		WebVid2M [4]	2M video-caption pairs
	QA	MSRVTT QA [101]	149,075 question-answer

Table 11: Datasets for Instruction Tuning: This table presents datasets used for instruction tuning, along with their associated task types and sizes. •Missing data results from expired links and corrupted files. ** Datasets marked with double asterisks are generated automatically within this study.

and Table 15 for videos. Compared to InstructBLIP [19] caption templates have increased from 13 to 32, while question-answering templates have grown from 10 to 21. These enhancements have been strategically incorporated to foster greater adaptability of the model to a wide range of user instructions.

9 Ethics Statement

In this research, we present a framework for aligning multiple modalities with a frozen large language model (LLM). Our methodology strictly involves the use of publicly available and free datasets, ensuring we do not engage in the collection of private data. However, it is crucial to acknowledge that publicly sourced datasets carry implicit biases [23, 71, 110]. These biases reflect historical and societal inequalities, potentially influencing the model's outputs. Our framework builds upon a pre-existing frozen LLM. While this approach benefits from the extensive knowledge encoded within the LLM, it is important to recognize that such models can inherently propagate biases present in their training data. Additionally, there is a non-negligible risk of generating false or misleading information. While there exist tools to measure language model toxicity such as Helm [55], their evaluation datasets are constrained in the language modality, and hence are not applicable to measure toxicity across modalities which is the

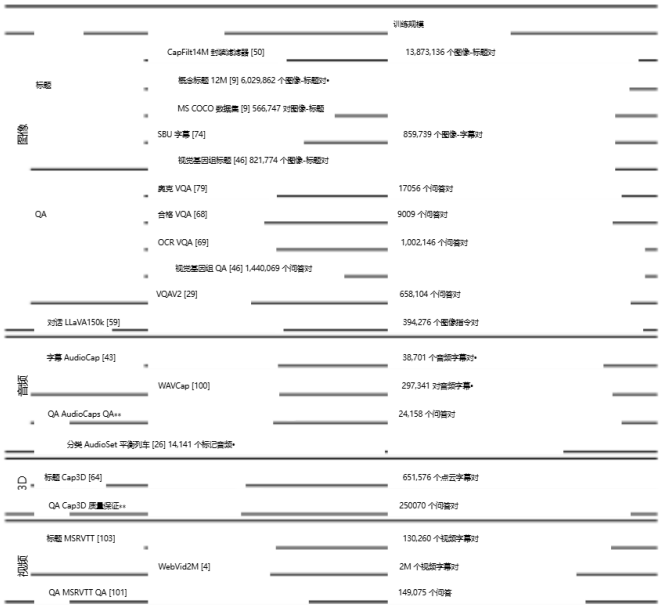


表 11：用于指令优化的数据集：此表显示了用于指令优化的数据集，以及它们相关的任务类型和大小。缺少数据是由于链接过期和文件损坏造成的。标有双星号的数据集是在此研究中自动生成的。

和 表 15 为视频。与 InstructBLIP 相比 [19] 字幕模板从 13 个增加到 32 个，而问答模板从 10 个增加到 21 个。这些增强功能已被战略性地纳入，以促进模型对各种用户指令的更大适应性。

9 道德声明

在这项研究中，我们提出了一个框架，用于将多种模态与冻结的大型语言模型 （）LLM 对齐。我们的方法严格涉及使用公开可用的免费数据集，确保我们不参与私人数据的收集。然而，承认公开来源的数据集带有隐性偏见是至关重要的 [23， 71， 110]。这些偏差反映了历史和社会不平等，可能会影响模型的输出。我们的框架建立在预先存在的 frozen LLM.虽然这种方法受益于 LLM中编码的广泛知识，但重要的是要认识到此类模型本身可以传播其训练数据中存在的偏差。此外，还存在产生虚假或误导性信息的风险。虽然存在测量语言模型毒性的工具，例如 Helm [55]，但它们的评估数据集在语言模态中受到限制，因此不适用于跨模态的毒性测量，即

	Image Instruction Templates
QA	"{question}" "Q: {question} A:" "Answer the following question: {question}" "Question: {question} Answer:" "How would you answer {question}?" "What is the answer to the question {question}?" "Answer the question based on the image. Question: {question} Answer: " "Instruction: Answer the following question by reference to the input image. Question: {question} Answer:" "Given the photo, what is the answer to the question {question}?" "What's your response to the query {question}?" "Please provide an answer to {question}" "Respond to the query {question}" "Based on the given image, respond to {question}" "Question: {question} What's your response?" "Consider the following query: {question}" "Could you help answer the question {question}?" "Referencing the provided image, can you answer the question {question}?" "With respect to the image shown, please answer {question}" "What's your answer to {question} in the context of the provided image?" "Question (refer to the image for context): {question} Answer:" "In response to the question {question}, what would your answer be?"
Caption	"A short caption:" "A short description:" "A photo of" "A photo that shows" "A picture of" "A picture that shows" "An image of" "A image that shows" "Write a short description." "Write a description for the image." "Provide a description of what is presented in the image." "Briefly describe the content of the image." "Can you briefly explain what you see in the image?" "Could you use a few words to describe what you perceive in the image?" "Please provide a short description of the image." "Using language, provide a short account of the image." "Use a few words to illustrate what is happening in the photo." "Write a description for the photo." "Provide a description of what is presented in the photo." "Briefly describe the content of the photo." "Can you briefly explain what you see in the photo?" "Could you use a few words to describe what you perceive in the photo?" "Please provide a short description of the picture." "Using language, provide a short account of the picture." "Use a few words to illustrate what is happening in the picture." "Write a description for the picture." "Provide a description of what is presented in the picture." "Briefly describe the content of the picture." "Can you briefly explain what you see in the picture?" "Could you use a few words to describe what you perceive in the picture?" "Please provide a short description of the picture." "Using language, provide a short account of the picture." "Use a few words to illustrate what is happening in the picture."

Table 12: Instruction-tuning templates for image tasks

focus of this work. We leave the generation of cross-modal datasets for toxicity and bias measurement as a future research direction.

Users of our framework should be aware of these limitations and exercise caution, particularly in applications where the accuracy and impartiality of outputs are critical. We advocate for responsible use of our framework, especially in sensitive contexts. Users should critically assess and verify the model's outputs and consider the potential for reinforcing biases or spreading misinformation. Furthermore, we commit to transparency regarding our model's capabilities and limitations. All code, data, and model weights will be released to ensure reproducibility and encourage external evaluation and subsequent research.

10 Reproducibility Statement

In alignment with the principles of open science and to foster reproducibility, transparency, and further research, we promise to provide open source access to all the resources associated with our study, including: a complete, documented, and public codebase with all the scripts, models, preprocessing, and evaluation

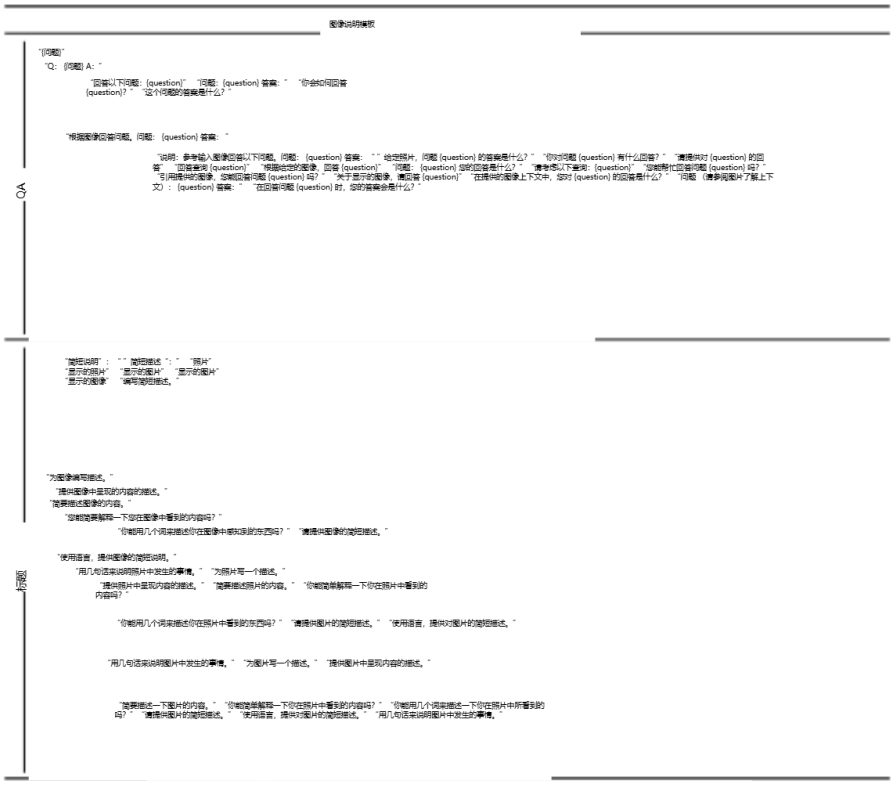


表 12：图像任务的指令调整模板

因此不适用于测量作为本研究重点的跨模式的毒性。我们将生成用于毒性和偏差测量的跨模态数据集作为未来的研究方向。

我们框架的用户应了解这些限制并谨慎行事，尤其是在输出的准确性和公正性至关重要的应用中。我们倡导负责任地使用我们的框架，尤其是在敏感环境中。用户应该批判性地评估和验证模型的输出，并考虑强化偏见或传播错误信息的可能性。此外，我们承诺对模型的功能和限制保持透明。所有代码、数据和模型权重都将发布，以确保可重复性并鼓励外部评估和后续研究。

10 可重复性声明

根据开放科学的原则，并促进可重复性、透明度和进一步研究，我们承诺提供对与我们的研究相关的所有资源的开源访问，包括：包含所有脚本、模型、预处理和评估的完整、记录在案的公共代码库

Audio Instruction Templates	
QA	"{question} "Question: {question} Answer:" "Q: {question} A:" "Based on the audio, {question}" "Answer the following question based on the audio: {question}" "Question: {question} Provide an answer based on the audio." "How would you answer {question} based on the audio?" "What is the answer to the question {question} using the audio as a reference?" "Answer the question using the audio. Question: {question} Answer: " "Instruction: Answer the following question by referencing the audio. Question: {question} Answer:" "Given the audio, what is the answer to the question {question}?" "What's your response to the query {question} considering the audio?" "Please provide an answer to {question} using the audio as context." " "Respond to the query {question} based on the audio content." "Based on the provided audio, respond to {question}" "Question: {question} What's your response using the audio for context?" "Consider the following query and the audio: {question}" "Could you help answer the question {question} using the audio as reference?" "Referencing the provided audio, can you answer the question {question}?" "With respect to the audio provided, please answer {question}" "What's your answer to {question} in the context of the provided audio?" "Question (refer to the audio for context): {question} Answer:" "In response to the question {question}, what would your answer be based on the audio?" "Given the audio, how would you respond to {question}?" "Taking the audio into consideration, what is your response to {question}?" "Based on the audio, how would you answer {question}?"
Classification	"Classify the following audio:" "What is the category of this audio clip?" "Identify the content of the following audio:" "Provide a classification for the audio." "Analyze and categorize the following audio." "Describe the category of the given audio." "Determine the type of this audio clip." " "Can you classify what you hear in the audio?" "What type of audio is this?" "How would you classify this audio clip?" "Please identify the category of the following audio:" "What category does the following audio fall into?" "Classify the sounds in this audio clip."
Caption	"A short caption:" "A short description:" "An audio of" "An audio that shows" "Write a short description." "Write a description for the audio." "Provide a description of what is presented in the audio." "Briefly describe the content of the audio." "Can you briefly explain what you hear in the audio?" "Could you use a few words to describe what you perceive in the audio?" "Please provide a short description of the audio." "Using language, provide a short account of the audio." " "Use a few words to illustrate what is happening in the audio." "Describe briefly the contents of the audio." "Please provide a brief summary of the audio." "What does the audio contain?" "What can you hear in the audio?" "What sounds are present in the audio?" "Summarize the audio in a few words." "Write a brief summary of the audio content." "Could you provide a concise explanation of the audio's contents?" "Describe what the audio represents." "What is the audio depicting?" "In a few words, describe what you hear in the audio."

Table 13: Instruction-tuning templates for audio tasks

code necessary to replicate the experiments. We will be further releasing the pretrained model weights along side the exact evaluation configs that generated the results cited in the paper. We show our commitment to reproducibility through an extensive supplementary section that highlights details of training and evaluation. Furthermore, all experiments were completed with prespecified random seeds that will also be made available in the experiment configuration files. Finally, we will release all datasets collected for this study for public download, as well as the code used to generate them. In addition to providing these resources, we pledge to maintain them and offer requisite support for any queries or clarifications related to the provided resources, contributing to a supportive and inclusive research environment.

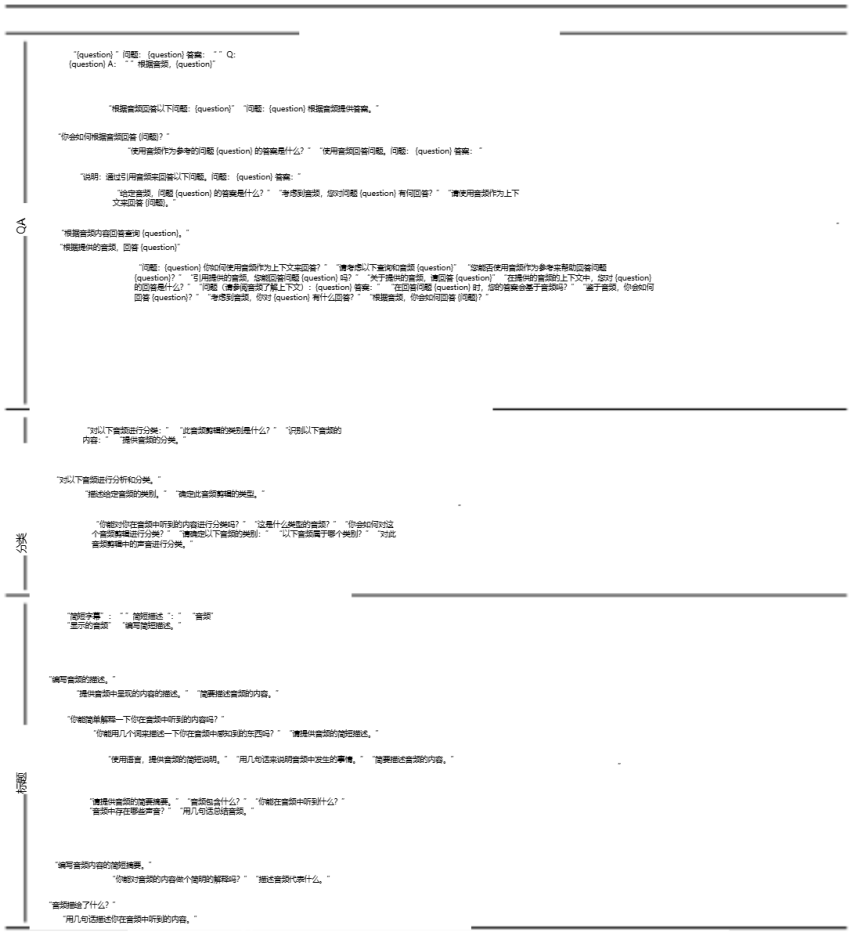


表 13：音频任务的指令调整模板

复制实验所需的代码。我们将进一步发布预训练模型权重以及生成论文中引用结果的确切评估配置。我们通过一个广泛的补充部分来展示我们对可重复性的承诺，该部分重点介绍了培训和评估的细节。此外，所有实验都是使用预先指定的随机种子完成的，这些种子也将在实验配置文件中提供。最后，我们将发布为本研究收集的所有数据集，以及用于生成这些数据集的代码，供公众下载。除了提供这些资源外，我们还承诺维护它们，并为与所提供资源相关的任何疑问或澄清提供必要的支持，从而为支持性和包容性的研究环境做出贡献。

3D Instruction Templates	
QA	"{question}" "Question: {question} Answer:" "Q: {question} A:" "Based on the 3D model, {question}" "Answer the following question based on the 3D model: {question}" "Question: {question} Provide an answer based on the 3D model." "How would you answer {question} based on the 3D model?" "What is the answer to the question {question} using the 3D model as a reference?" "Answer the question using the 3D model. Question: {question} Answer: " "Instruction: Answer the following question by referencing the 3D model. Question: {question} Answer:" "Given the 3D model, what is the answer to the question {question}?" "What's your response to the query {question} considering the 3D model?" "Please provide an answer to {question} using the 3D model as context." "Respond to the query {question} based on the 3D model content." "Based on the provided 3D model, respond to {question}" "Question: {question} What's your response using the 3D model for context?" "Consider the following query and the 3D model: {question}" "Could you help answer the question {question} using the 3D model as reference?" "Referencing the provided 3D model, can you answer the question {question}?" "With respect to the 3D model provided, please answer {question}" "What's your answer to {question} in the context of the provided 3D model?" "Question (refer to the 3D model for context): {question} Answer:" "In response to the question {question}, what would your answer be based on the 3D model?" "Given the 3D model, how would you respond to {question}?" "Taking the 3D model into consideration, what is your response to {question}?" "Based on the 3D model, how would you answer {question}?"
Caption	"A short caption:" "A short description:" "A 3D model of" "A 3D model that shows" "Write a short description." "Write a description for the 3D model." "Provide a description of what is presented in the 3D model." "Briefly describe the content of the 3D model." "Can you briefly explain what you see in the 3D model?" "Could you use a few words to describe what you perceive in the 3D model?" "Please provide a short description of the 3D model." "Using language, provide a short account of the 3D model." "Use a few words to illustrate what is happening in the 3D model." "Describe briefly the contents of the 3D model." "Please provide a brief summary of the 3D model." "What does the 3D model contain?" "What can you identify in the 3D model?" "What structures are present in the 3D model?" "Summarize the 3D model in a few words." "Write a brief summary of the 3D model content." "Could you provide a concise explanation of the 3D model's contents?" "Describe what the 3D model represents." "What is the 3D model depicting?" "In a few words, describe what you see in the 3D model."

Table 14: Instruction-tuning templates for 3D tasks

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8948–8957 (2019)
2. Alamri, H., Hori, C., Marks, T.K., Batra, D., Parikh, D.: Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In: DSTC7 at AAAI2019 Workshop. vol. 2 (2018)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
5. Bansal, A., Zhang, Y., Chellappa, R.: Visual question answering on image sets. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 51–67. Springer (2020)
6. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: nearly real-time answers to

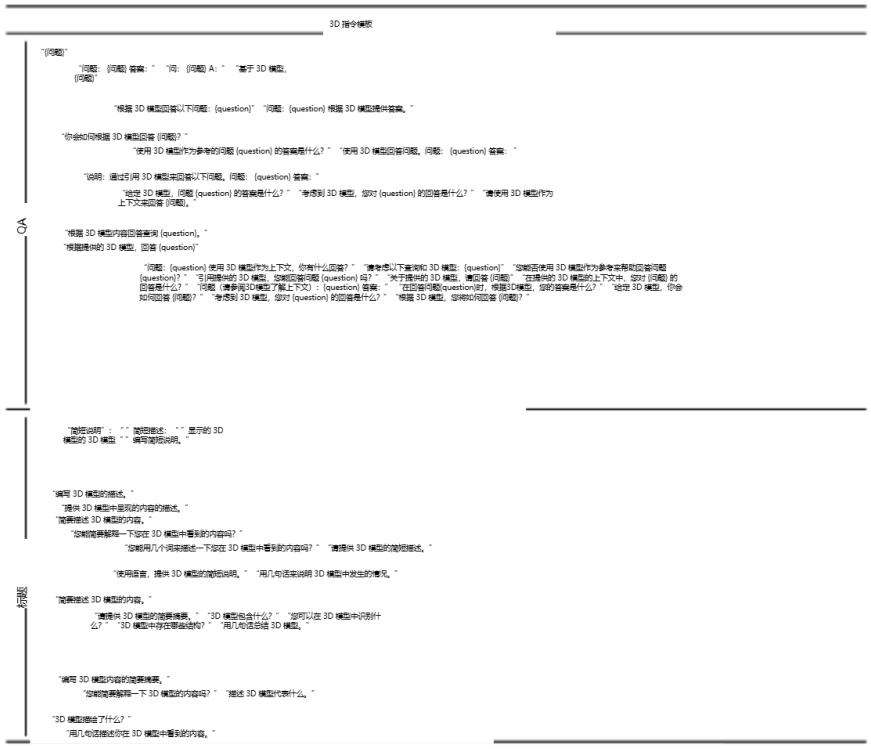


表 14：用于 3D 任务的指令调整模板

引用

1. Agrawal，H.，Desai，K.，Wang，Y.，Chen，X.，Jain，R.，Johnson，M.，Batra，D.，Parikh，D.，Lee，S.，Anderson，P.：nocaps：大规模的新对象标题。收录于：IEEE 计算机视觉国际会议论文集。页码 8948– 8957 （2019）

2. Alamri，H.，Hori，C.，Marks，TK，Batra，D.，Parikh，D.：在 dstc7 中生成自然语言的视听场景感知对话 （avsd） 轨道。在：AAAI2019 研讨会的 DSTC7。第 2 卷 （2018）
3.Alayrac，JB，Donahue，J.，Luc，P.，Miech，A.，Barr，I.，Hasson，Y.，Lenc，K.，Mensch，A.，Millican，K.，Reynolds，M.，et al.：火烈鸟：用于小样本学习的视觉语言模型。神经信息处理系统进展 35，23716–23736 （2022） 4.Bain，M.，Nagrani，A.，Varol，G.，Zisserman，A.：时间冻结：用于端到端检索的联合视频和图像编码器。收录于：IEEE/CVF 计算机视觉国际会议论文集。第 1728–1738 页 （2021） 5.Bansal，A.，Zhang，Y.，Chellappa，R.：图像集的视觉问答。收录于：Computer Vision–ECCV 2020：第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23 日至 28 日，会议记录，第 XXI 部分 16.第 51-67 页。斯普林格 （2020） 6.Bigham，J.P.，Jayant，C.，Ji，H.，Little，G.，Miller，A.，Miller，R.，Miller，R.，Tatarowicz，A.，White，B.，White，S.，et al.：Vizwiz：近乎实时的答案

	Video Instruction Templates
QA	"Given the video, {question}"
	"Q: {question} A:"
	"Answer the following question based on the video: {question}"
	"Question: {question} Answer:"
	"How would you answer {question} after watching the video?"
	"What is the answer to the question {question} after viewing the video?"
	"Answer the question based on the video. Question: {question} Answer: "
	"Instruction: Answer the following question by reference to the input video. Question: {question} Answer:"
	"Given the video, what is the answer to the question {question}?"
	"What's your response to the query {question} after watching the video?"
	"Please provide an answer to {question} after watching the video"
	"Respond to the query {question} based on the video"
	"Based on the given video, respond to {question}"
	"Question: {question} What's your response after watching the video?"
	"Consider the following query: {question}"
	"Could you help answer the question {question}?"
	"Referencing the provided video, can you answer the question {question}?"
	"With respect to the video shown, please answer {question}"
	"What's your answer to {question} in the context of the provided video?"
	"Question (refer to the video for context): {question} Answer:"
	"In response to the question {question}, what would your answer be after viewing the video?"
Caption	"A short caption for the video:"
	"A short description of the video:"
	"A video of"
	"A video that shows"
	"Describe the video briefly."
	"Write a description for the video."
	"Provide a description of what is presented in the video."
	"Briefly describe the content of the video."
	"Can you briefly explain what you see in the video?"
	"Could you use a few words to describe what you perceive in the video?"
	"Please provide a short description of the video."
	"Using language, provide a short account of the video."
	"Use a few words to illustrate what is happening in the video."

Table 15: Instruction-tuning templates for audio tasks

visual questions. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. pp. 333–342 (2010)

7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
8. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the "Video" in Video-Language Understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
9. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021)
10. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
11. Chen, F., Han, M., Zhao, H., Zhang, Q., Shi, J., Xu, S., Xu, B.: X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. arXiv preprint arXiv:2305.04160 (2023)
12. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18030–18040 (2022)
13. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., Wei, F.: BEATs: Audio pre-training with acoustic tokenizers. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of

: Vizwiz：对 28 A. Panagopoulou 等人的近乎实时的回答。

表 15：音频任务的指令调整模板

视觉问题。收录于：第 23 届年度 ACM 用户界面软件和技术研讨会论文集。第 333–342 页（2010）

7. Brown，T.，Mann，B.，Ryder，N.，Subbiah，M.，Kaplan，JD，Dhariwal，P.，Neelakantan，A.，Shyam，P.，Sastry，G.，Askell，A.，et al.：语言模型是少数学习者。神经信息处理系统的进展 33，1877–1901（2020）8.Buch，S.，Eyzaguirre，C.，Gaidon，A.，Wu，J.，Fei-Fei，L.，Niebles，J.C.：重新审视视频语言理解中的“视频”。载于：IEEE/CVF 计算机视觉和模式识别会议（CVPR）会议记录（2022 年）9.Changpinyo，S.，Sharma，P.，Ding，N.，Soricut，R.：概念 12m：推动网络规模图像文本预训练以识别长尾视觉概念。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。

第 3558–3568 页（2021）10.Chen，D.，Dolan，W.B.：收集高度并行的数据用于释义评估。收录于：计算语言学协会第 49 届年会论文集：人类语言技术。第 190-200 页（2011）11.Chen，F.，Han，M.，Zhao，H.，Zhang，Q.，Shi，J.，Xu，S.，Xu，B.：X-llm：通过将多模态视为外语来引导高级大型语言模型。arXiv 预印本 arXiv：2305.04160（2023）12.Chen，J.，Guo，H.，Yi，K.，Li，B.，Elhoseiny，M.：Visualgpt：用于图像描述的预训练语言模型的数据高效改编。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。页。

18030–18040 (2022) 13.Chen，S.，Wu，Y.，Wang，C.，Liu，S.，Tompkins，D.，Chen，Z.，Che，W.，Yu，X.，Wei，F.：BEATs：使用声学分词器进行音频预训练。在：Krause，A.，Brunskill，E.，Cho，K.，Engelhardt，B.，Sabato，S.，Scarlett，J.（eds.）第 40 届机器学习国际会议论文集。会议记录

Machine Learning Research, vol. 202, pp. 5178–5193. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/chen23ag.html
14. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=scYa9DYUAy
15. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: Simclr: A simple framework for contrastive learning of visual representations. In: International Conference on Learning Representations. vol. 2 (2020)
16. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
17. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023) (2023)
18. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. pp. 1931–1942. PMLR (2021)
19. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=vvoWPYqZJA
20. Deshmukh, S., Elizalde, B., Singh, R., Wang, H.: Pengi: An audio language model for audio tasks. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=gJLAfO4KUq
21. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-e: An embodied multimodal language model. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 8469–8488. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/driess23a.html
22. Drossos, K., Lipping, S., Virtanen, T.: Clotho: An audio captioning dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 736–740. IEEE (2020)
23. Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I.: A survey on bias in visual datasets. Computer Vision and Image Understanding **223**, 103552 (2022)
24. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)
25. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
26. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)

机器学习研究，第 202 卷，第 5178-5193 页。PMLR（2023 年 7 月 23 日至 29 日）、https://proceedings.mlr.press/v202/chen23ag.html

14 的会议记录。Chen，S.，Li，H.，Wang，Q.，Zhao，Z.，Sun，M.，Zhu，X.，Liu，J.：VAST：视觉音频-字幕-文本全模态基础模型和数据集。在：三十-第七届神经信息处理系统会议（2023 年），https：openreview.net/forum?id=scYa9DYUAy

15. Chen，T.，Kornblith，S.，Norouzi，M.，Hinton，G.：Simclr：视觉表征对比学习的简单框架。在：学习表征国际会议。第 2 卷（2020）16.Chen，T.，Kornblith，S.，Norouzi，M.，Hinton，G.：视觉表征对比学习的简单框架。在：机器学习国际会议。第 1597-1607 页。PMLR（2020）17.Chiang，W.L.，Li，Z.，Lin，Z.，Sheng，Y.，Wu，Z.，Zhang，H.，Zheng，L.，Zhuang，S.，Zhuang，Y.，Gonzalez，J.E.，et al.：Vicuna：一个开源聊天机器人，以 90%* 的 chatgpt 质量给 gpt-4 留下深刻印象。请参阅 https://vicuna。LMSYS 的。组织（2023 年 4 月 14 日访问）（2023）18.Cho，J.，Lei，J.，Tan，H.，Bansal，M.：通过文本生成统一视觉和语言任务。在：机器学习国际会议。第 1931-1942 页。

PMLR（2021）19.戴，W.，李，J.，李，D.，Tiong，A.，赵，J.，王，W.，李，B.，Fung，P.，Hoi，S.：InstructBLIP：通过指令调整实现通用视觉语言模型。在：第 37 届神经信息处理系统会议

（2023 年），https://openreview.net/forum?id=vvoWPYqZJA

20. Deshmukh，S.，Elizalde，B.，Singh，R.，Wang，H.：Pengi：用于音频任务的音频语言模型。收录于：第 37 届神经信息处理会议系统（2023），https://openreview.net/forum?id=gJLAfO4KUq

21. Driess，D.，Xia，F.，Sajjadi，M.S.M.，Lynch，C.，Chowdhery，A.，Ichter，B.，Wahid，A.，Tompson，J.，Vuong，Q.，Yu，T.，Huang，W.，Chebotar，Y.，Sermanet，P.，Duckworth，D.，Levine，S.，Vanhoucke，V.，Hausman，K.，Toussaint，M.，Greff，K.，Zeng，A.，Mordatch，I.，Florence，P.：PaLM-e：一种具身的多模态语言模型。在：Krause，A.，Brunskill，E.，Cho，K.，Engelhardt，B.，Sabato，S.，Scarlett，J.（eds.）第 40 届机器学习国际会议论文集。机器学习研究论文集，第 202 卷，第 8469-8488 页。PMLR（23–2023 年 7 月 29 日）、https://proceedings.mlr.press/v202/driess23a.html

22. Drossos，K.，Lipping，S.，Virtanen，T.：Clotho：音频字幕数据集。在：ICASSP 2020-2020 IEEE 声学、语音和信号处理国际会议（ICASSP）。第 736-740 页。国际电气工程师学会（2020）23.Fabbrizzi，S.，Papadopoulos，S.，Ntoutsi，E.，Kompatsiaris，I.：视觉数据集中偏差的调查。计算机视觉和图像理解 223，103552（2022）24.方，Y.，王，W.，谢，B.，孙，Q.，吴，L.，王晓，黄，T.，王晓，曹，Y.：Eva：探索大规模掩蔽视觉表示学习的极限。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。第 19358–19369 页（2023）25.Fu，C.，Chen，P.，Shen，Y.，Qin，Y.，Zhang，M.，Lin，X.，Yang，J.，Zheng，X.，Li，K.，Sun，X.，et al.：Mme：多模态大型语言模型的综合评估基准。arXiv 预印本 arXiv：2306.13394（2023）26.Gemmeke，J.F.，Ellis，DP，Freedman，D.，Jansen，A.，Lawrence，W.，Moore，R.C.，Plakal，M.，Ritter，M.：音频集：音频事件的本体和人为标记数据集。收录于：2017 年 IEEE 声学、语音和信号处理国际会议（ICASSP）。第 776-780 页。IEEE（2017 年）

27. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (June 2023)
28. Gong, Y., Luo, H., Liu, A.H., Karlinsky, L., Glass, J.R.: Listen, think, and understand. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=nBZBPXdJlC
29. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
30. Guangyao li, Yixin Xu, D.H.: Multi-scale attention for audio question answering. Proc. INTERSPEECH (2023)
31. Gui, L., Wang, B., Huang, Q., Hauptmann, A.G., Bisk, Y., Gao, J.: Kat: A knowledge augmented transformer for vision-and-language. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 956–968 (2022)
32. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., et al.: Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:2309.00615 (2023)
33. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Audioclip: Extending clip to image, text and audio. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 976–980. IEEE (2022)
34. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. arXiv preprint arXiv:2312.03700 (2023)
35. Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al.: Imagebind-llm: Multi-modality instruction tuning. arXiv preprint arXiv:2309.03905 (2023)
36. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
37. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-LLM: Injecting the 3d world into large language models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=YQA28p7qNz
38. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
39. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., Wei, F.: Language is not all you need: Aligning perception with language models. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=UpN2wfrLec
40. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
41. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: International conference on machine learning. pp. 4651–4664. PMLR (2021)

27. Girdhar， R.， El-Nouby， A.， Liu， Z.， Singh， M.， Alwala， K.V.， Joulin， A.， Misra， I.： Imagebind：一个嵌入空间来绑定它们。收录于：IEEE/CVF 计算机视觉和模式识别（CVPR） 会议记录。

第 15180-15190 页（2023 年 6 月），第 28 页。Gong， Y.， Luo， H.， Liu， A.H.， Karlinsky， L.， Glass， J.R.： 倾听、思考和理解。在：第十二届学习表征国际会议

（2024 年），https://openreview.net/forum?id=nBZBPXdJlC

29. Goyal， Y.， Khot， T.， Summers-Stay， D.， Batra， D.， Parikh， D.：让 VQA 中的 V 很重要：提升图像理解在视觉问答中的作用。载于：计算机视觉和模式识别会议 （CVPR） （2017） 30.Guangyao Li， Yixin Xu， D.H.： 音频问答的多尺度注意力。INTERSPEECH （2023） 31.Gui， L.， Wang， B.， Huang， Q.， Hauptmann， AG， Bisk， Y.， Gao， J.：Kat：用于视觉和语言的知识增强转换器。收录于：计算语言学协会北美分会 2022 年会议论文集：人类语言技术。第 956-968 页 （2022）：32。郭振军， 张振军， 朱振军， 唐振军， 马振军， 韩振军， 陈振军， 高平， 李振军， 李振华， 等人： 点绑定和点-llm： 将点云与多模态对齐，用于 3D 理解、生成和指令遵循。arXiv 预印本 arXiv：2309.00615 （2023） 33.Guzhov， A.， Raue， F.， Hees， J.， Dengel， A.： 音频剪辑：将剪辑扩展到图像、文本和音频。在：ICASSP 2022-2022 IEEE 声学、语音和信号处理国际会议 （ICASSP）。第 976-980 页。IEEE （2022 年） 34.Han， J.， Gong， K.， Zhang， Y.， Wang， J.， Zhang， K.， Lin， D.， Qiao， Y.， Gao， P.， Yue， X.： Onellm：一个框架，使所有模态与语言保持一致。arXiv 预印本 arXiv：2312.03700 （2023） 35.韩杰， 张， R.， 邵， W.， 高， P.， 徐， P.， 肖， H.， 张， K.， 刘， C.， 温， S.， 郭， Z.， et al.： Imagebind-llm： 多模态指令调优.arXiv 预印本 arXiv：2309.03905 （2023） 36.He， K.， Fan， H.， Wu， Y.， Xie， S.， Girshick， R.： 无监督视觉表示学习的动量对比。收录于：计算机视觉和模式识别 IEEE/CVF 会议论文集。第 9729–9738 页 （2020） 37.Hong， Y.， Zhen， H.， Chen， P.， Zheng， S.， Du， Y.， Chen， Z.， Gan， C.： 3d-LLM： 将 3d 世界注入大型语言模型。在：第 37 届神经信息处理系统会议 （2023 年），https://openreview。网/

forum？ id=YQA28p7qNz

38. 胡， E.J.， 沃利斯， P.， 朱艾伦， Z.， 李， Y.， 王， S.， 王， L.， 陈， W.， et al.： 劳拉：大型语言模型的低秩适应。载于：学习表征国际会议（2021 年） 39.黄， S.， 董， L.， 王， W.， 郝， Y.， 辛哈尔， S.， 马， S.， 吕， T.， 崔， L.， 穆罕默德， OK， 帕特拉， B.， 刘， Q.， 阿加瓦尔， K.， 池， Z.， 比约克， J.， 乔杜里， V.， 索姆， S.， 宋， X.， 魏， F.： 语言不是你所需要的全部：使感知与语言模型保持一致。收录于：第 37 届神经信息会议

加工系统 （2023），https://openreview.net/forum?id=UpN2wfrLec

40. Hudson， DA， Manning， CD： Gqa：用于真实世界视觉推理和作文问答的新数据集。收录于：计算机视觉和模式识别 IEEE/CVF 会议论文集。第 6700–6709 页 （2019）：41.Jaegle， A.， Gimeno， F.， Brock， A.， Vinyals， O.， Zisserman， A.， Carreira， J.： 感知：具有迭代注意力的一般感知。在：机器学习国际会议。第 4651-4664 页。PMLR （2021 年）

42. Jiang, C., Ye, W., Xu, H., Huang, S., Huang, F., Zhang, S.: Vision language pre-training by contrastive learning with cross-modal similarity regulation. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14660–14679. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.819, https://aclanthology.org/2023.acl-long.819
43. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 119–132 (2019)
44. Kim, M., Sung-Bin, K., Oh, T.H.: Prefix tuning for automated audio captioning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
45. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 17283–17300. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/koh23a.html
46. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
47. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., Si, L.: mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 7241–7259. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.488, https://aclanthology.org/2022.emnlp-main.488
48. Li, D., Li, J., Le, H., Wang, G., Savarese, S., Hoi, S.C.: LAVIS: A one-stop library for language-vision intelligence. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). pp. 31–41. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-demo.3
49. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19108–19118 (2022)
50. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: 40th International Conference on Machine Learning (2023)
51. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/li22n.html

PMLR（2021）42.江，C.，叶，W.，徐，H.，黄，S.，黄，F.，张，S.：视觉语言通过对比学习和跨模态相似性调节进行预训练。在：Rogers，A.，Boyd-Graber，J.，Okazaki，N.（eds.）计算语言学协会第 61 届年会论文集（第 1 卷：长篇论文）。第 14660-14679 页。计算语言学协会，加拿大多伦多（2023 年 7 月）。https://doi.org/10.18653/v1/2023.acl-long.819、https：//：//aclanthology.org/2023.acl-long.819

43. Kim，CD，Kim，B.，Lee，H.，Kim，G.：Audiocaps：为野外音频生成字幕。收录于：计算语言学协会北美分会 2019 年会议论文集：人类语言技术，第 1 卷（长篇和短篇论文）。第 119-132 页（2019）：44。Kim，M.，Sung-Bin，K.，Oh，T.H.：自动音频字幕的前缀调整。在：ICASSP 2023-2023 IEEE 声学、语音和信号处理国际会议（ICASSP）。第 1-5 页。IEEE（2023 年）45。Koh，JY，Salakhutdinov，R.，Fried，D.：将语言模型接地到多模态输入和输出的图像。在：Krause，A.，Brunskill，E.，Cho，K.，Engelhardt，B.，Sabato，S.，Scarlett，J.（eds.）第 40 届机器学习国际会议论文集。机器学习研究论文集，第 202 卷，第 17283-17300 页。PMLR（2023 年 7 月 23 日至 29 日），https://proceedings.mlr。press/v202/koh23a.html

46. Krishna，R.，Zhu，Y.，Groth，O.，Johnson，J.，Hata，K.，Kravitz，J.，Chen，S.，Kalantidis，Y.，Li，L.J.，Shamma，DA，et al.：视觉基因组：使用众包密集图像注释连接语言和视觉。国际计算机视觉杂志 123，32–73（2017）47。李，C.，徐，H.，田，J.，王W.，严，M.，毕，B.，叶J.，陈，H.，徐G.，曹Z.，张J.，黄S.，黄，F.，周，J.，Si，L.：mPLUG：通过跨模态跳跃连接进行有效和高效的视觉语言学习。在：Goldberg，Y.，Kozareva，Z.，Zhang，Y.（eds.）2022 年自然语言处理经验方法会议论文集。第 7241-7259 页。计算语言学协会，阿拉伯联合酋长国阿布扎比（2022 年 12 月）。https://doi.org/10.18653/v1/2022.emnlp- main.488，https：//：//aclanthology.org/2022.emnlp-main.488

48. Li，D.，Li，J.，Le，H.，Wang，G.，Savarese，S.，Hoi，SC：LAVIS：语言视觉智能的一站式库。收录于：计算语言学协会第 61 届年会论文集（第 3 卷：系统演示）。第 31-41 页。Association for Computational Linguistics，加拿大，多伦多（2023 年 7 月），https://aclanthology.org/2023.acl-demo.3

49. 李国，魏，Y.，田，Y.，徐，C.，温，J.R.，胡，D.：学习在动态视听场景中回答问题。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。第 19108–19118 页（2022）：50。Li，J.，Li，D.，Savarese，S.，Hoi，S.：Blip-2：使用冻结图像编码器和大型语言模型进行引导语言图像预训练。在：第 40 届机器学习国际会议（2023 年）51.Li，J.，Li，D.，Xiong，C.，Hoi，S.：BLIP：用于统一视觉-语言理解和生成的引导语言图像预训练。在：Chaudhuri，K.，Jegelka，S.，Song，L.，Szepesvari，C.，Niu，G.，Sabato，S.（eds.）第 39 届机器学习国际会议论文集。机器学习研究论文集，第 162 卷，第 12888-12900 页。PMLR（2022 年 7 月 17 日至 23 日），https://proceedings.mlr.press/v162/li22n.html

52. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
53. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
54. Li, Y., Li, W., Nie, L.: MMCoQA: Conversational question answering over text, tables, and images. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4220–4231. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.290, https://aclanthology.org/2022.acl-long.290
55. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C.A., Manning, C.D., Re, C., Acosta-Navas, D., Hudson, D.A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N.S., Khattab, O., Henderson, P., Huang, Q., Chi, R.A., Xie, S.M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., Koreeda, Y.: Holistic evaluation of language models. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=iO4LZibEqW, featured Certification, Expert Certification
56. Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C., Yuan, L.: Revive: Regional visual representation matters in knowledge-based visual question answering. Advances in Neural Information Processing Systems **35**, 10560–10571 (2022)
57. Lipping, S., Sudarsanam, P., Drossos, K., Virtanen, T.: Clotho-aqa: A crowdsourced dataset for audio question answering. In: 2022 30th European Signal Processing Conference (EUSIPCO). pp. 1140–1144. IEEE (2022)
58. Liu, H., Yan, W., Abbeel, P.: Language quantized autoencoders: Towards unsupervised text-image alignment. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=mlxRLIy7kc
59. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=w0H2xGHlkw
60. Liu*, P.J., Saleh*, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., Shazeer, N.: Generating wikipedia by summarizing long sequences. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=HygOvbWC-
61. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. In: Proceedings of the IEEE international conference on computer vision. pp. 873–881 (2017)
62. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
63. Luo, R., Zhao, Z., Yang, M., junwei dong, Li, D., Wang, T., Qiu, M., Hu, L., zhongyu wei: Valley: Video assistant with large language model enhanced ability (2024), https://openreview.net/forum?id=bjyf5FyQ0a
64. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models. In: Proceedings of the NeurIPS 2023 (2023)

32 A. Panagopoulou 等人，52.Li， J.， Selvaraju， R.， Gotmare， A.， Joty， S.， Xiong， C.， Hoi， SCH： 融合前对齐：使用动量蒸馏进行视觉和语言表征学习。

神经信息处理系统的进展 34， 9694–9705 （2021） 53.李晓东， 尹晓东， 李俊杰， 张平， 胡晓明， 张俊杰， 胡俊杰， 胡晓东， 董俊杰， 魏俊俊， 等人： 奥斯卡：视觉语言任务的对象语义对齐预训练。在：计算机视觉-ECCV 2020：第 16 届欧洲会议，英国格拉斯哥，2020 年 8 月 23 日至 28 日，会议记录，第 XXX 部分 16。第 121-137 页。施普林格 （2020） 54.Li， Y.， Li， W.， Nie， L.： MMCoQA：文本、表格和图像的对话式问答。收录于：计算语言学协会第 60 届年会论文集（第 1 卷：长篇论文）。第 4220– 4231 页。计算语言学协会，爱尔兰都柏林（2022 年 5 月）。

https://doi.org/10.18653/v1/2022.acl-long.290，https://aclanthology。org/2022.acl-long.290 55.梁，P.，Bommasani，R.，李，T.，齐普拉斯，D.，索伊鲁，D.，安永，M.，张，Y.，纳拉亚南，D.，吴，Y.，库马尔，A.，纽曼，B.，袁B.，严，B.，张，C.，科斯格罗夫，C.A.，曼宁，C.D.，Re，C.，阿科斯塔-纳瓦斯，D.，哈德逊，D.A.，泽利克曼，E.，杜尔姆斯，E.，拉达克，F.，荣，F.，任，H.，姚，H.，王，J.，桑塔南， K.， Orr， L.， Zheng， L.， Yuksekgonul， M.， Suzgun， M.， Kim， N.， Guha， N.， Chatterji， NS， Khattab， O.， Henderson， P.， Huang， Q.， Chi， RA， Xie， SM， Santurkar， S.， Ganguli， S.， Hashimoto， T.， Icard， T.， Zhang， T.， Chaudhary， V.， Wang， W.， Li， X.， Mai， Y.， Zhang， Y.， Koreeda， Y.： 语言模型的整体评估。机器学习研究汇刊 （2023），

https://openreview.net/forum?id=iO4LZibEqW， 特色认证， 专家认证 56.Lin， Y.， Xie， Y.， Chen， D.， Xu， Y.， Zhu， C.， Yuan， L.： 复兴：区域视觉表示在基于知识的视觉问答中很重要。神经信息处理系统进展 35， 10560–10571 （2022） 57。 Lipping， S.， Sudarsanam， P.， Drossos， K.， Virtanen， T.： Clotho-aqa：用于音频问答的众包数据集。收录于：2022 年第 30 届欧洲信号处理会议 （EUSIPCO）。第 1140-1144 页。 IEEE（2022 年）58。Liu， H.， Yan， W.， Abbeel， P.： 语言量化自动编码器：迈向无监督文本图像对齐。收录于：第 37 届神经信息会议

加工系统 （2023），https://openreview.net/forum?id=mlxRLIy7kc

59. Liu， H.， Li， C.， Wu， Q.， Lee， YJ.：视觉指令调整。在：三十-第七届神经信息处理系统会议（2023 年），https：openreview.net/forum?id=w0H2xGHlkw

60. Liu*， PJ， Saleh*， M.， Pot， E.， Goodrich， B.， Sepassi， R.， Kaiser， L.， Shazeer， N.：通过总结长序列生成维基百科。在：学习表征国际会议 （2018），https://openreview.net/forum?id=-HygOvbWC-

61. Liu， S.， Zhu， Z.， Ye， N.， Guadarrama， S.， Murphy， K.： 通过蜘蛛的策略梯度优化改进图像标题。收录于：IEEE 计算机视觉国际会议论文集。第 873-881 页 （2017）： 62。 Loshchilov， I.， Hutter， F.： 解耦权重衰减正则化。载于：学习表征国际会议（2018 年） 63. 罗瑞， 赵 Z.， 杨 M.， 董俊伟， 李 D.， 王 T.， 邱 M.， 胡 L.， 魏中宇： 谷：具有大语言模型增强能力的视频助手

（2024 年），https://openreview.net/forum?id=bjyf5FyQOa

64. Luo， T.， Rockwell， C.， Lee， H.， Johnson， J.： 使用预训练模型进行可扩展的 3D 字幕。在： NeurIPS 2023 会议记录 （2023）

65. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
66. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2023)
67. Mañas, O., Rodriguez Lopez, P., Ahmadi, S., Nematzadeh, A., Goyal, Y., Agrawal, A.: MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In: Vlachos, A., Augenstein, I. (eds.) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2523–2548. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.eacl-main.185, https://aclanthology.org/2023.eacl-main.185
68. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
69. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: ICDAR (2019)
70. Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.F., Murugesan, P., Heidari, P., Liu, Y., et al.: Anymal: An efficient and scalable any-modality augmented language model. arXiv preprint arXiv:2309.16058 (2023)
71. Motoki, F., Neto, V.P., Rodrigues, V.: More human than human: Measuring chatgpt political bias. Public Choice pp. 1–21 (2023)
72. Nagrani, A., Seo, P.H., Seybold, B., Hauth, A., Manen, S., Sun, C., Schmid, C.: Learning audio-video modalities from image captions. In: European Conference on Computer Vision. pp. 407–426. Springer (2022)
73. Najdenkoska, I., Zhen, X., Worring, M.: Meta learning to bridge vision and language models for multimodal few-shot learning. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=3oWo92cQyxL
74. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc. (2011), https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf
75. Paranjape, B., Lamm, M., Tenney, I.: Retrieval-guided counterfactual generation for qa. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1670–1686 (2022)
76. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)
77. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
78. Salesforce: Ulip. https://github.com/salesforce/ULIP (2022), accessed: 2023-07-1
79. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
80. Shao, Z., Yu, Z., Wang, M., Yu, J.: Prompting large language models with answer heuristics for knowledge-based visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14974–14983 (2023)

在：NeurIPS 2023 会议记录 （2023） 65。Van der Maaten， L.， Hinton， G.： 使用 t-sne 可视化数据。机器学习研究杂志 9（11） （2008） 66。Maaz， M.， Rasheed， H.， Khan， S.， Khan， F.S.： 视频聊天：通过大型视觉和语言模型实现详细的视频理解（2023 年） 67.Mañas， O.， Rodriguez Lopez， P.， Ahmadi， S.， Nematzadeh， A.， Goyal， Y.， Agrawal， A.：MAPL：单峰预训练模型的参数高效适应，用于视觉语言小镜头提示。在： Vlachos， A.， Augenstein， I.（eds.） 计算语言学协会欧洲分会第 17 届会议论文集。第 2523-2548 页。计算语言学协会，克罗地亚杜布罗夫尼克（2023 年 5 月）。 https://doi.org/10.18653/v1/2023。

eacl-main.185 中，https://aclanthology.org/2023.eacl-main.185
68. Marino， K.， Rastegari， M.， Farhadi， A.， Mottaghi， R.： Ok-vqa：需要外部知识的视觉问答基准。载于：计算机视觉和模式识别会议 （CVPR） （2019） 69.Mishra， A.， Shekhar， S.， Singh， AK， Chakraborty， A.： Ocr-vqa：通过阅读图像中的文本进行视觉问答。在：ICDAR （2019） 70。Moon， S.， Madotto， A.， Lin， Z.， Nagarajan， T.， Smith， M.， Jain， S.， Yeh， CF， Murugesan， P.， Heidari， P.， Liu， Y.， et al.： Anymal：一种高效且可扩展的模态增强语言模型。arXiv 预印本 arXiv：2309.16058 （2023） 71.Motoki， F.， Neto， VP， Rodrigues， V.： 比人类更人性化：衡量 chatgpt 政治偏见。公共选择，第 1-21 页（2023 年）：72。Nagrani， A.， Seo， PH， Seybold， B.， Hauth， A.， Manen， S.， Sun， C.， Schmid， C.： 从图像字幕中学习音频视频模式。在：欧洲计算机视觉会议。第 407-426 页。施普林格 （2022） 73.Najdenkoska， I.， Zhen， X.， Worring， M.： 元学习在多模态小样本学习中架起视觉和语言模型的桥梁。在：第十一届学习表征国际会议（2023 年），https://openreview。

net/forum？id=3oWo92cQyxL
74. Ordonez， V.， Kulkarni， G.， Berg， T.： Im2text：使用 100 万张带标题的照片描述图像。在：Shawe-Taylor， J.， Zemel， R.， Bartlett， P.， Pereira， F.， Weinberger， K.（eds.） 神经信息处理系统的进展

第 24 卷。Curran Associates， Inc. （2011 年），https://proceedings.neurips.cc/论文/2011/文件/5dd9db5e033da9c6fb5ba83c7a7ebea9 -Paper.pdf
75. Paranjape， B.， Lamm， M.， Tenney， I.：用于 qa 的检索导向反事实生成。收录于：计算语言学协会第 60 届年会论文集（第 1 卷：长篇论文）。第 1670–1686 页 （2022）：76。 Piczak， KJ： Esc：环境声音分类数据集。收录于：第 23 届 ACM 多媒体国际会议论文集。第 1015-1018 页 （2015） 77.Radford， A.， Kim， JW， Hallacy， C.， Ramesh， A.， Goh， G.， Agarwal， S.， Sastry， G.， Askell， A.， Mishkin， P.， Clark， J.， et al.： 从自然语言监督中学习可转移的视觉模型。在：机器学习国际会议。

第 8748-8763 页。PMLR （2021） 78.销售人员：Ulip。https://github.com/salesforce/ULIP （2022），访问时间：2023 年-

07-1
79. Schwenk， D.， Khandelwal， A.， Clark， C.， Marino， K.， Mottaghi， R.： A-okvqa：使用世界知识进行视觉问答的基准。在：欧洲计算机视觉会议。第 146-162 页。施普林格（2022） 80。Shao， Z.， Yu， Z.， Wang， M.， Yu， J.： 使用答案启发式方法提示大型语言模型，以进行基于知识的视觉问答。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。页码： 14974– 14983 （2023）

81. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual llm for video understanding. arXiv preprint arXiv:2312.06720 (2023)
82. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: PandaGPT: One model to instruction-follow them all. In: Hazarika, D., Tang, X.R., Jin, D. (eds.) Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23. Association for Computational Linguistics, Prague, Czech Republic (Sep 2023), https://aclanthology.org/2023.tllm-1.2
83. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=mL8Q9OOamV
84. Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., Saito, K.: Slidevqa: A dataset for document visual question answering on multiple images. In: AAAI (2023)
85. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., MA, Z., Zhang, C.: SALMONN: Towards generic hearing abilities for large language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=14rn7HpKVk
86. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
87. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019)
88. Van Zwol, R.: Flickr: Who is looking? In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'07). pp. 184–190. IEEE (2007)
89. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
90. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: Git: A generative image-to-text transformer for vision and language. Transactions on Machine Learning Research (2022)
91. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: Onepeace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
92. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
93. Wang, T., Ge, Y., Zheng, F., Cheng, R., Shan, Y., Qie, X., Luo, P.: Accelerating vision-language pretraining with free language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23161–23170 (June 2023)
94. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)

14974– 14983 （2023） 34 A. Panagopoulou 等人。

81. Shu， F.， Zhang， L.， 江， H.， 谢， C.： 用于视频理解llm的视听。arXiv 预印本 arXiv：2312.06720 （2023） 82.Su， Y.， Lan， T.， Li， H.， Xu， J.， Wang， Y.， Cai， D.： PandaGPT：一个模型来指导 - 遵循它们。在：Hazarika， D.， Tang， X.R.， Jin， D.（eds.） 第一届驯服大型语言模型研讨会论文集：交互式助手时代的可控性！第 11-23 页。计算语言学协会，捷克共和国布拉格（2023 年 9 月），https://aclanthology.org/2023。

TLLM-1.2 的

83. Sun， Q.， Yu， Q.， Cui， Y.， Zhang， F.， Zhang， X.， Wang， Y.， Gao， H.， Liu， J.， Huang， T.， Wang， X.： 鹌鹑：多模态中的生成式预训练。在：
第十二届学习表征国际会议（2024 年），https：
openreview.net/forum?id=mL8Q9OOamV

84. Tanaka， R.， Nishida， K.， Nishida， K.， Hasegawa， T.， Saito， I.， Saito， K.： Slidevqa：用于对多个图像进行文档视觉问答的数据集。在：AAAI （2023） 85。Tang， C.， Yu， W.， Sun， G.， Chen， X.， Tan， T.， Li， W.， Lu， L.， MA， Z.， Zhang， C.： SALMONN：迈向大型语言模型的通用听力能力。在：

第十二届学习表征国际会议（2024 年），https：
openreview.net/forum?id=14rn7HpKVk

86. Tsimpoukelli， M.， Menick， J.L.， Cabi， S.， Eslami， S.， Vinyals， O.， Hill， F.： 使用冻结语言模型的多模态小样本学习。神经信息处理系统进展 34， 200–212 （2021） 87。 Uy， M.A.， Pham， Q.H.， 华， B.S.， Nguyen， T.， Yeung， S.K.： 重新审视点云分类：基于真实世界数据的新基准数据集和分类模型。收录于：IEEE/CVF 计算机视觉国际会议论文集。第 1588–1597 页（2019 年）：第 88 页。Van Zwol， R.： Flickr： 谁在看？在： IEEE/WIC/ACM 国际网络智能会议 （WI'07）。第 184-190 页。IEEE（2007 年）89。 Vedantam， R.， Lawrence Zitnick， C.， Parikh， D.： 苹果酒：基于共识的图像描述评估。收录于：IEEE 计算机视觉和模式识别会议论文集。第 4566–4575 页 （2015） 90.王，J.，杨，Z.，胡，X.，李，L.，林，K.，甘，Z.，刘，Z.，刘，C.，王，L.：Git：用于视觉和语言的生成图像到文本转换器。机器学习研究汇刊 （2022） 91.王平， 王， S.， 林， J.， 白， S.， 周， X.， 周， J.， 王， X.， 周， C.： 一个和平：探索一种面向无限模式的一般表示模型。

arXiv 预印本 arXiv：2305.11172 （2023） 92.王平， 杨 A.， 门 R.， 林 J.， 白 S.， 李 Z.， 马 J.， 周 C.， 周 J.， 杨 H.： Ofa：通过简单的序列到序列学习框架统一架构、任务和模式。在：机器学习国际会议。第 23318-23340 页。PMLR （2022） 93.Wang， T.， Ge， Y.， Zheng， F.， Cheng， R.， Shan， Y.， Qie， X.， Luo， P.： 通过免费语言建模加速视觉语言预训练。收录于：IEEE/CVF 计算机视觉和模式识别 （CVPR） 会议记录。

第 23161-23170 页（2023 年 6 月），第 94 页。Wang， W.， Bao， H.， Dong， L.， Bjorck， J.， Peng， Z.， Liu， Q.， Aggarwal， K.， Mohammed， OK， Singhal， S.， Som， S.， et al.：图像作为外语：视觉和视觉语言任务的 Beit 预训练。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。页码： 19175–19186 （2023）

95. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
96. Wang, Z., Chen, C., Li, P., Liu, Y.: Filling the image information gap for vqa: Prompting large language models to proactively ask questions. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 2874–2890 (2023)
97. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=gEZrGCozdqR
98. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
99. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
100. XinhaoMei: Wavcaps. https://github.com/XinhaoMei/WavCaps (2023), accessed: 2023-07-1
101. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
102. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., Zhou, J.: Mplug-2: A modularized multi-modal foundation model across text, image and video. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
103. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
104. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds (2023)
105. Xu, W., Chen, K., Zhao, T.: Discriminative reasoning for document-level relation extraction. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1653–1663. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.144, https://aclanthology.org/2021.findings-acl.144
106. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1179–1189 (2023)
107. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems **35**, 124–141 (2022)
108. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision. pp. 521–539. Springer (2022)
109. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3081–3089 (2022)

95. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex：用于视频和语言研究的大规模、高质量的多语言数据集。载于：IEEE/CVF 计算机视觉国际会议 （ICCV） 论文集（2019 年 10 月） 96.Wang, Z., Chen, C., Li, P., Liu, Y.：填补 vqa 的图像信息空白：促使大型语言模型主动提出问题。在：计算语言学协会的调查结果：EMNLP 2023。第 2874–2890 页 （2023） 97。Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, AW, Lester, B., Du, N., Dai, A.M., Le, Q.V.: 微调语言模型是零样本学习者。在：学习表征国际会议 （2022），https://openreview.net/forum?

id=gEZrGCozdqR

98. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., 周, D., et al.: 在大型语言模型中提示引发推理的思维链。

神经信息处理系统进展 35， 24824–24837 （2022） 99。Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d 形状网：体积形状的深度表示。收录于：IEEE 计算机视觉和模式识别会议论文集。第 1912-1920 页 （2015）：100。XinhaoMei: Wavcaps.https://github.com/XinhaoMei/WavCaps （2023），访问时间：2023-07-1 101。Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: 通过逐渐细化对外观和动作的关注进行视频问答。在： Pro-

第 25 届 ACM 多媒体国际会议的结束。第 1645–1653 页 （2017）

102. 徐海， 叶Q.， 严明， 石Y.， 叶J.， 徐英， 李建， 毕乙， 钱Q.， 王文， 徐建， 张建军， 黄建军， 黄建军， 黄建军， 周建军： Mplug-2： 模块化多

跨文本、图像和视频的模态基础模型。收录于：第 40 届机器学习国际会议论文集。ICML'23，JMLR.org （2023）

103. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt：用于桥接视频和语言的大型视频描述数据集。在：IEEE 计算机会议论文集

视觉和模式识别。第 5288-5296 页 （2016） 104。Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm：使大型语言模型能够理解点云 （2023） 105.Xu, W., Chen, K., Zhao, T.: 文档级关系提取的判别推理。在：计算语言学协会的调查结果：ACL-

IJCNLP 2021 年。第 1653-1663 页。计算语言学协会，在线（2021 年 8 月）。https://doi.org/10.18653/v1/2021.findings-acl.144，https://aclanthology.org/2021.findings-acl.144

106. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip：学习语言、图像和

用于 3D 理解的点云。收录于：IEEE/CVF 计算机视觉和模式识别会议论文集。第 1179–1189 页 （2023）

107. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.：通过冻结的双向语言模型进行零镜头视频问答。神经信息的进展

加工系统 35， 124–141 （2022） 108。Yang, Z., Gan, Z., Wang, J., 胡, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab：统一文本和框输出，用于扎根视觉语言建模。

在：欧洲计算机视觉会议。第 521-539 页。施普林格 （2022） 109。Yang, Z., Gan, Z., Wang, J., 胡, X., 卢, Y., 刘, Z., 王, L.: gpt-3 用于少数镜头基于知识的 vqa 的实证研究，收录于：AAAI 会议记录

人工智能会议。第 36 卷，第 3081–3089 页（2022 年）

110. Yeh, K.C., Chi, J.A., Lian, D.C., Hsieh, S.K.: Evaluating interfaced llm bias. In: Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023). pp. 292–299 (2023)
111. Yu, L., Cheng, Y., Wang, Z., Kumar, V., Macherey, W., Huang, Y., Ross, D.A., Essa, I., Bisk, Y., Yang, M.H., Murphy, K.P., Hauptmann, A.G., Jiang, L.: SPAE: Semantic pyramid autoencoder for multimodal generation with frozen LLMs. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=CXPUg86A1D
112. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
113. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. Empirical Methods in Natural Language Processing 2023, Demo Track (2023)
114. Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Li, H., Gao, P., Qiao, Y.: LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=d4UiXAHN2W
115. Zhao, Z., Guo, L., Yue, T., Chen, S., Shao, S., Zhu, X., Yuan, Z., Liu, J.: Chatbridge: Bridging modalities with large language model as a language catalyst. arXiv preprint arXiv:2305.16103 (2023)
116. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=1tZbq88f27

3081–3089 （2022） 36 A. Panagopoulou 等人。

110. Yeh， KC， Chi， JA， Lian， D.C.， Hsieh， S.K.：评估界面偏差llm。收录于：第 35 届计算语言学和语音会议论文集

处理 （ROCLING 2023）。第 292-299 页 （2023）：111。Yu， L.， Cheng， Y.， Wang， Z.， Kumar， V.， Macherey， W.， Huang， Y.， Ross， D.A.

Essa， I.， Bisk， Y.， Yang， M.H.， Murphy， K.P.， Hauptmann， A.G.， 江， L.：SPAE： 用于冻结LLMs多模态生成的语义金字塔自动编码器。在：第 37 届神经信息处理系统会议（2023 年），https://openreview.net/forum?id=CXPUg86A1D

112. Yu， W.， Yang， Z.， Li， L.， Wang， J.， Lin， K.， Liu， Z.， Wang， X.， Wang， L.： Mm-vet：评估大型多模态模型的集成功能。arXiv 预印本

arXiv：2308.02490 （2023） 113.Zhang， H.， Li， X.， Bing， L.： 视频骆驼：用于视频理解的指令调整视听语言模型。自然语言中的实证方法

Processing 2023，演示轨道 （2023） 114.张 R.， 韩 J.， 刘 C.， 周 A.， 卢 P.， 李 H.， 高 P.， 乔 Y.： LLaMAadapter：大型语言模型的高效微调，零初始化 at-

tention 的 intent 中。在：第十二届学习表征国际会议

（2024 年），https://openreview.net/forum?id=d4UiXAHN2W

115. Zhao， Z.， Guo， L.， Yue， T.， Chen， S.， Shao， S.， Zhu， X.， Yuan， Z.， Liu， J.： Chatbridge：以大型语言模型作为语言催化剂的桥接模态。

arXiv 预印本 arXiv：2305.16103 （2023） 116.Zhu， D.， Chen， J.， Shen， X.， Li， X.， Elhoseiny， M.： MiniGPT-4：使用先进的大型语言模型增强视觉语言理解。在：

第十二届学习表征国际会议（2024 年），https：openreview.net/forum?id=1tZbq88f27