

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 Li², Dongxu Li², Shafiq Joty², Ran Xu², Silvio Savarese², Caiming Xiong², and Juan Carlos Niebles²

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 (**D**iscriminative **C**ross-modal **R**easoning (**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.

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 reasoning³.

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 **Discriminatory Cross-modal Reasoning** 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.

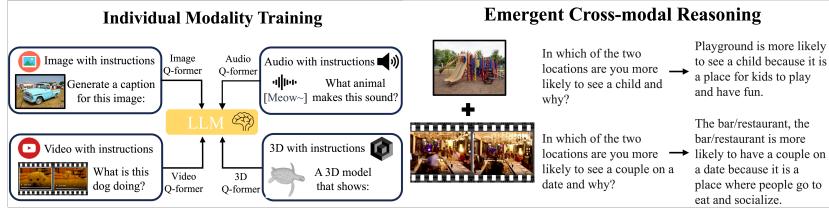


Fig. 1: Despite utilizing distinct pre-trained encoders for each modality and independently aligning them to language through individual instruction aware Q-Formers, X-InstructBLIP demonstrates emergent abilities in cross-modal comprehension.

(iii) We collect **DisCRn**, the first dataset designed for evaluating instruction-based cross-modal discriminative reasoning. Which includes $\sim 36k$ 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 pre-training [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 apart 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]

<p> [here comes the sun]</p> <p>Q: Which band plays this famous song?</p> <p>A: The Beatles</p> <p>Q: What instrument you hear?</p> <p>A: Guitar</p>	<p> [nutcracker theme]</p> <p>Q: What is the name of this famous melody?</p> <p>A: Waltz of the Snowflakes from the Nutcracker Suite by Tchaikovsky</p>	
<p> [poem recitation]</p> <p>Q: Which input is more likely to be taking place in a classroom?</p> <p>A: Audio</p> <p>Q: Which one is taking place in a kitchen?</p> <p>A: Video</p>	<p> [poem recitation]</p> <p>Q: Describe the image in detail</p> <p>A: The image shows a 3d model of a white and red animal, which appears to be a cat or a wolf it is standing on its hind legs.</p>	<p>Q: Describe the image in detail using the corresponding 3D model.</p> <p>A: A 3D model of a white tiger with red eyes and a long tail, standing on its hind legs.</p>
<p> [baaa-]</p> <p>Q: What animal is this?</p> <p>A: Cow</p> <p>Q: Using both the audio and the 3d model, what animal am I?</p> <p>A: Sheep</p>	<p> [sōkyoku]</p> <p>Q: Which country am I in?</p> <p>A: Japan</p>	<p> [children laughing]</p> <p>Q: How likely is it to hear these sounds in the location depicted in the image?</p> <p>A: It is very likely to hear these sounds in the location depicted in the image</p>
<p> [music beats]</p> <p>Q: Briefly describe the video.</p> <p>A: A 3D animation of a computer screen</p> <p>Q: What sound is playing in the audio?</p> <p>A: Music is playing</p> <p>Q: Given the audio describe the video.</p> <p>A: This is a DJ software screen shot of a DJ software on a laptop</p>	<p> [music beats]</p> <p>Q: Describe the 3D model.</p> <p>A: 3D model of a cartoon character with long legs and arms, resembling a monster</p>	<p> [music beats]</p> <p>Q: Describe the 3D model.</p> <p>A: 3D model of a triceratops dinosaur with horns and a long tail</p>
		<p>Q: Describe the image.</p> <p>A: There are four yellow parrots sitting on a branch in front of a mesh fence the birds are perched together and seem to be enjoying each other's company they are surrounded by green trees and foliage</p>

Fig. 2: Qualitative Examples: X-InstructBLIP framework effectively handles both uni-modal 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

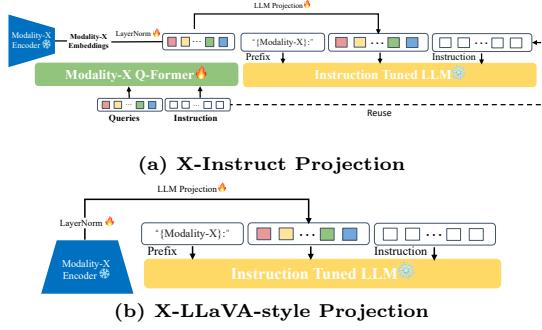


Fig. 3: Projection types explored in the X-InstructBLIP Framework. (a) is an instruction aware Q-Former projection [19] and (b) is a linear projection [59].

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 $f_\theta^M : R^{d_M} \rightarrow R^{k d_L}$ 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]: $\min_{\theta} \mathcal{L}_{\text{CE}}(\text{LLM}(x_{\text{LLM}}), h(y))$ where \mathcal{L}_{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.

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 $\mathbb{I} = \{I_{Mt} : M \in \mathbb{M}, t \in \mathbb{T}\}$ for each task $t \in \mathbb{T}$

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

Given a modality M encoding $z_M = \text{Enc}_M(x_M)$ and task instruction $i_M \in \mathbb{I}_{Mt}$, 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 BERT_{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 $\text{LP}_M : R^{d_M} \rightarrow R^{k_{d_{\text{LLM}}}}$, 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.

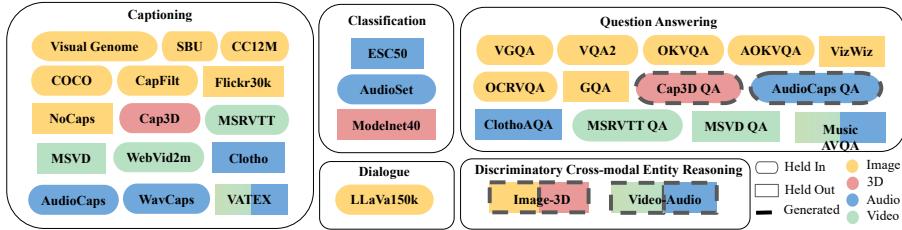


Fig. 4: Instruction Tuning and Evaluation Datasets: Oval-enclosed and square datasets are tuning and out-of-domain evaluation datasets respectively. Dashed outline is used for automatically derived datasets as described in Section 4.1.

4 Datasets

X-InstructBLIP is optimized and evaluated on a collection of pre-existing and automatically generated datasets succinctly 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 $\frac{\sqrt{|D_{M_d}|}}{\sum_{d \in [1 \dots |\mathbb{D}_M|]} \sqrt{|D_{M_d}|}}$, with minimal adjustments as justified in the supplementary material.

Instruction Data Augmentation: Extracting instruction-aware representations necessitates diverse instruction-related tasks across all modalities. Notably, datasets for 3D and audio modalities are majorly 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 $\sim 250k$ 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.

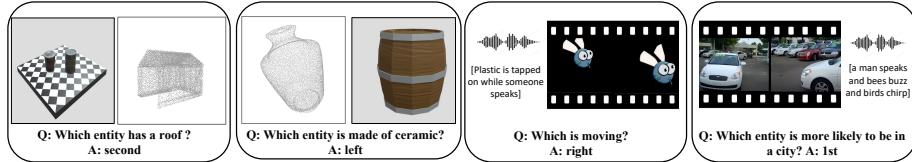


Fig. 5: **DisCRn**. Given two distinct modality inputs, select which one fits the query.

Cap3D [64]⁵ and $\sim 24k$ 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 **Discriminatory Cross-modal Reasoning (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/f1an-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.

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 `BEATsiter3+` [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 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 $\beta_1 = 0.9$, $\beta_2 = 0.999$, 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

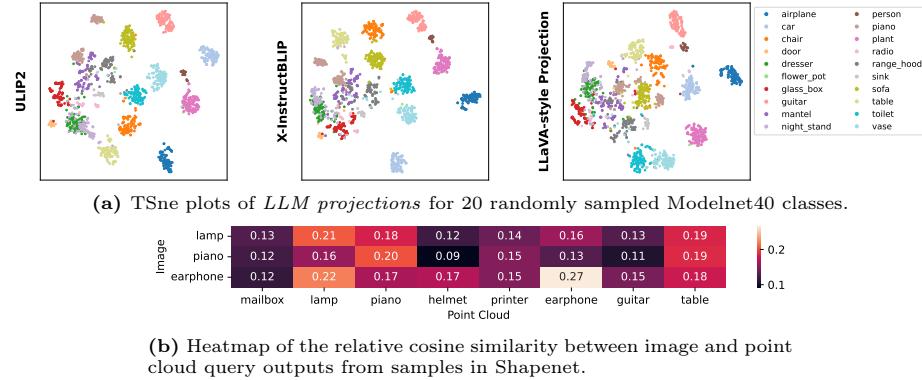


Fig. 6: Analysis on the alignment of X-InstructBLIP representations.

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-

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.

	ESC50close Acc.	ESC50open Acc.	ClothoAQA EM	Clotho v1 CIDEr	Clotho v2 CIDEr
ImageBind [27]	66.9	✗	✗	✗	✗
MWAFM [30]	-	-	22.2	-	-
Pengi [20]	-	53.9	64.5	39.6	30.0
Kim et. al., 2023 [44]	-	-	-	-	19.2
ImageBind-LLM (7B) [35]	40.1	27.4	10.3†	3.7	5.5
X-LLaVA-style Proj. (7b)	67.4	20.3	26.9	25.3	16.6
X-Instruct Proj. (7b)	75.9	38.2	21.4	29.4	19.5
X-Instruct Proj. (13b)	77.1	34.6	21.7	28.7	18.8

Table 3: Out-Domain Audio Quantitative Results.

	Flickr30k	NoCaps	VizWiz	GQA	MME	MMVet
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	-
PandaGPT (13b) [82]	23.0	29.7	-	-	871.2	19.6
ImagebindLLM (7b) [35]	23.5	30.4	-	-	980.3	-
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.

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).

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 outperform 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 reasoning⁶. 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._{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.

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

Table 7: Ablation: Prefix Effect

	ESC50 _{close}	ESC50 _{open}	ClothoAQA	Clotho v1	Clotho v2
X-Instruct Proj. (7b)	75.9	38.2	15.4	29.4	26.7
X-Instruct Proj. (7b) _{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)_{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-InstructBLIP, 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.

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 Li², Dongxu Li², Shafiq Joty², Ran Xu², Silvio Savarese², Caiming Xiong², and Juan Carlos Niebles²

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.

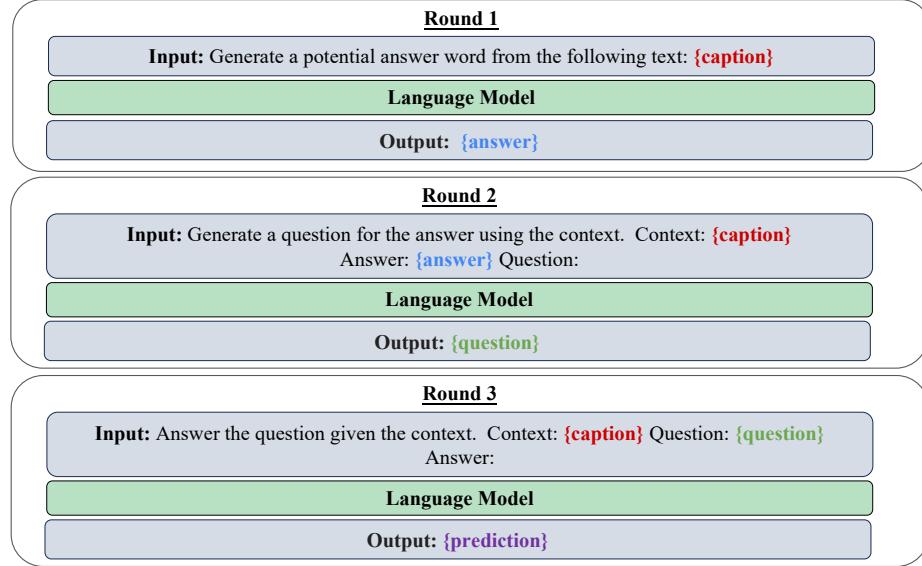


Fig. 1: Round-Trip-Consistency Prompting for QA Datasets in 3D and Audio.

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

	Caption	Question	Answer
Audio	A woman speaks while types a keyboard;	What is the woman typ- ing 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	Applauds after the man speaks?	
	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	
	<i>train</i>	<i>val</i>	<i>train</i>	<i>val</i>
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

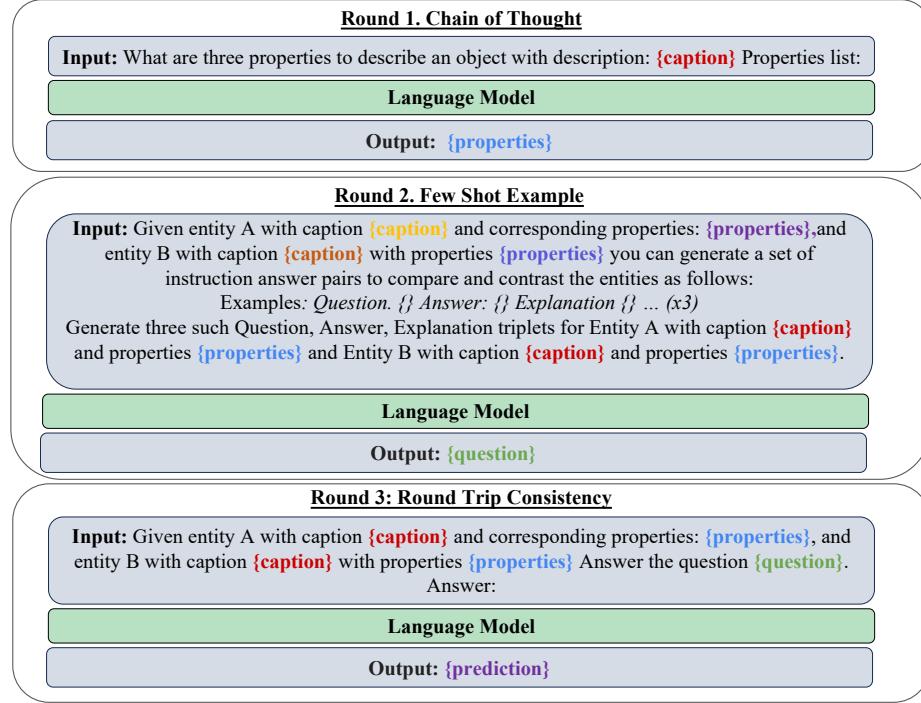


Fig. 2: Cross-modal Discriminative Reasoning Dataset Generation Framework

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.

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 <i>test</i>	VATEX <i>val</i>	MSVD QA <i>test</i>
X-LLaVA Style Proj. (7b)	105.3	46.2	49.8
X-LLaVA Style Proj. (7b)[<i>image</i>]	16.4	10.7	23.2
X-Instruct Proj. (7b)	116.1	59.2	51.7
X-Instruct Proj. (7b) [<i>image</i>]	42.4(\downarrow 73.7)	28.1(\downarrow 30.1)	39.7(\downarrow 12.0)
X-Instruct Proj.-no-prefix (7b)[<i>image</i>]	62.0(\downarrow 56.7)	52.6(\downarrow 6.9)	38.8(\downarrow 11.7)
X-Instruct Proj. (13b)	124.3	52.0	49.2
X-Instruct Proj. (13b) [<i>image</i>]	78.7(\downarrow 45.6)	53.5(\uparrow 1.5)	36.0(\downarrow 13.2)

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

	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. 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 <i>test</i>	COCO <i>val</i>	COCO <i>test</i>	Cap3D <i>val</i>	Cap3D <i>qa-val</i>	MSRVTT <i>val</i>	MSRVTT <i>test</i>	MSRVTT <i>val</i>	MSRVTT <i>test</i>	QA <i>val</i>	QA <i>test</i>	AudioCaps <i>qa-val</i>
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.

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 than InstructBLIP. The reason is that we maintain conformity with the other Q-Formers and do not initialize 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

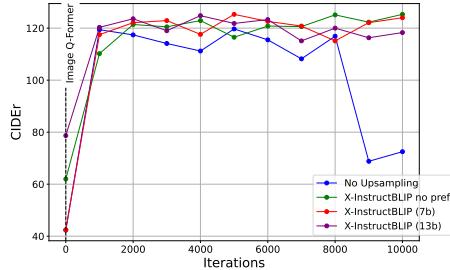


Fig. 3: CIDEr score on MSVD (out-domain) over training iterations on Video Q-Former initialized from Image Q-Former. Most performance gains are achieved within only 1000 iterations.

	MSVD <i>test</i>	VATEX <i>val</i>	MSVD QA <i>test</i>
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.

	Flickr30k <i>test</i>	NoCaps <i>val-all</i>	Zero-Shot <i>test-dev</i>	VizWiz <i>balanced test-dev</i>	GQA <i>test</i>	OKVQA <i>test</i>	In-Domain COCO <i>val</i>	In-Domain COCO <i>test</i>
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.

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

	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		
Image	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
Image	Dialogue	LLaVA150k [59]	394,276 image-instruction pairs
	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
		Cap3D QA**	250,070 question-answer pairs
	Caption	MSRVTT [103]	130,260 video-caption pairs
		WebVid2M [4]	2M video-caption pairs
Video	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

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

Audio Instruction Templates	
QA	<ul style="list-style-type: none"> "{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." "Respond to 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	<ul style="list-style-type: none"> "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	<ul style="list-style-type: none"> "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 feelings 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 pre-trained 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.

3D Instruction Templates	
QA	<ul style="list-style-type: none"> "{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	<ul style="list-style-type: none"> "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

Video Instruction Templates	
QA	<ul style="list-style-type: none"> "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	<ul style="list-style-type: none"> "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 23rd 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*

- 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=gJLAf04KUq>
 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., Serbanet, 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)

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=nBZBPXdJ1C>
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 multimodality 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)

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>

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.: MMCQA: 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-Nava, D., Hudson, D.A., Zelfikman, E., Durmus, E., Ladzhak, 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., Khatab, 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=i04LZibEqW>, 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 crowd-sourced 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=m1xRLIy7kc>
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=w0H2xGH1kw>
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=Hyg0vbWC>
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)

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)

81. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual lilm 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.tilm-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=mL8Q900amV>
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.: Multi-modal 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.: One-peace: 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 pre-training for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 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)

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>