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

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

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

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

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

1 Introduction

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

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

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

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

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

Our framework explores two state-of-the-art projection mechanisms on frozen LLMs - a prerequisite for maintaining separate modality training - instruction-aware Q-Formers [14] and linear projections [48]. Through an expansive evaluation on 13 benchmarks across 4 modalities we find that Q-Formers tend to exhibit higher performance on single modality tasks and versatility in distinguishing when to reason in a joint or discriminative manner in the presence of 2+ extra-linguistic modalities. Figure 2 shows illustrative results, highlighting the capabilities of our framework. To quantify and challenge this emergent ability we introduce `DisCRn` in Section 4.2, an automatically curated **Dis**criminatory **C**ross-modal **R**easo**n**ing challenge dataset requiring models to distinguish between diverse combinations of modalities, such as audio-video and 3D-image.
Our contributions are summarized as follows:
(i) We introduce an extendable framework that aligns Image, 3D, Audio, and Video to LLMs, and we benchmark its emergent cross-modal reasoning capability across two projection mechanisms. This framework does not need specific pre-training tailored to each modality. To the best of our knowledge, this is the first attempt to demonstrate that discriminative cross-modal reasoning emerges naturally through individual modality alignment to LLMs.
(ii) We introduce an automatic approach for crafting instruction-tuning datasets for a variety of modalities, leveraging only readily available captioning data and open-source language models. Contributing $\sim$ 250k samples for 3D QA data and $\sim$ 24k samples for Audio QA data.

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

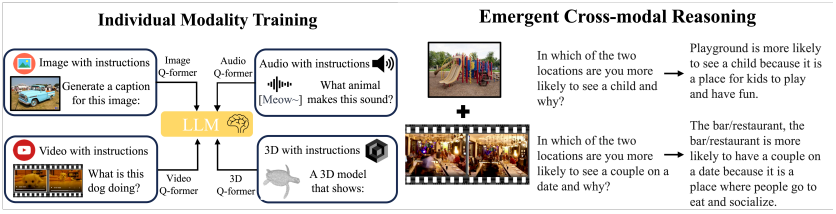


Fig. 1: Despite utilizing distinct pre-trained encoders for each modality and independently aligning them to language through individual instruction aware Q-Formers, X-InstructBLIP demonstrates emergent abilities in cross-modal comprehension.

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

2 Related Work

Vision Language Models: Recent years have seen a surge in models capable of executing a spectrum of vision-language tasks, leading to the creation of Multimodal Language Models (MLMs). These models align the static vision and language representations through various techniques, such as unified pretraining [13, 31, 37, 66, 72, 74–76, 87], vision-to-language alignment through textual feature extraction [23, 45, 63, 78, 88], vision-encoder optimization [69], and linear [16, 36, 48], transformer-based [8, 56, 59], or auto-encoder based projections [47, 89]. More relevant to this work are approaches that learn intermediate vision-informed language token representations either interleaved in LLM layers such as in Flamingo [2] and LLaMA adapter [92] or only to the input layer such as in the BLIP series [14, 40, 41] which employ Q-Former based projections, LLAVA [48], and MiniGPT4 [94] which employ linear projections.
Cross-Modal Language Models: Projection-based approaches, initially focused on images, have recently broadened to encompass audio [15, 21, 35, 68], video [3, 52, 55, 86], and 3D projections [24, 29, 84] into pre-trained large language models (LLMs). This expansion has seen the advent of unified pretraining frameworks such as mPLUG2 [83] and OnePeace [73], as well as projection-based methods for enhancing frozen LLMs, like VideoLLaMA [91] and X-LLM [7], 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 [93] adopts a training approach akin to X-LLM but utilizes a perceiver-based projection [32]. Audio-Visual LLM [64] follows a similar training paradigm of a final joint finetuning stage, but instead of maintaining a frozen LLM, it updates it using LoRA [30]. Our method is set appart by maintaining independent finetuning throughout and a frozen LLM avoiding the instability in training due to disparately aligned modality projections. Another line of work, including ImageBind-LLM [27], PandaGPT [65] and PointLLM [24] leverages

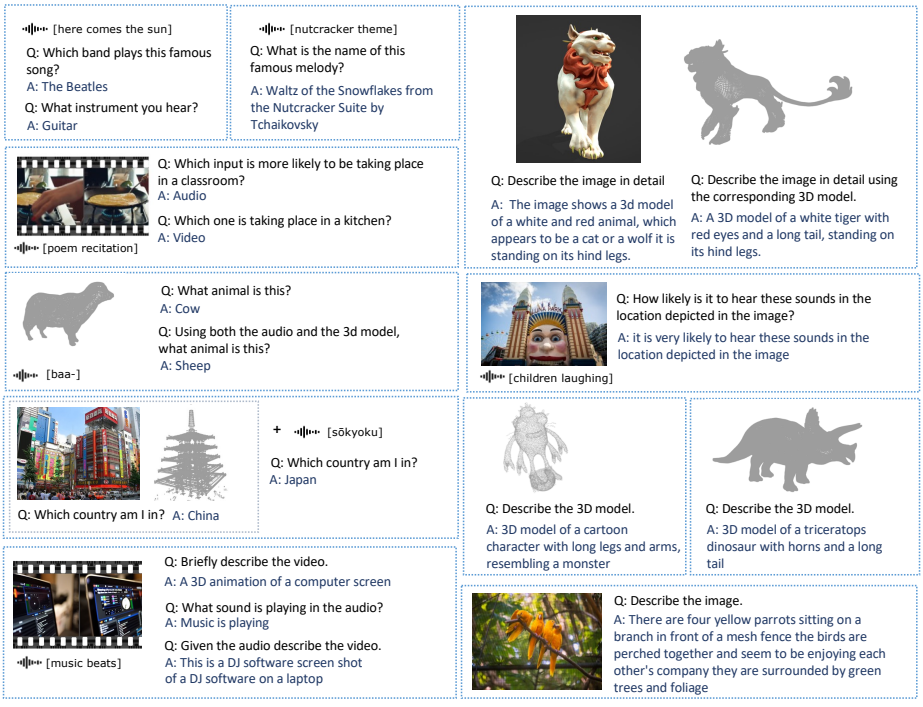
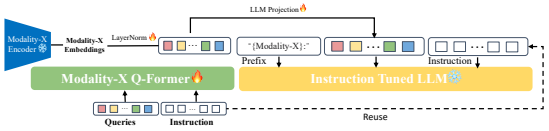


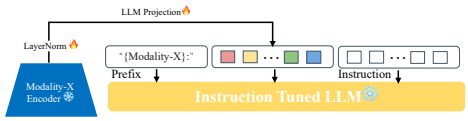
Fig. 2: Qualitative Examples: X-InstructBLIP framework effectively handles both unimodal and cross-modal tasks without training on joint data.

unified representations such as ImageBind [20] to only implicitly align additional modalities to LLMs by only training on image-text pairs. Contemporary works such as AnyMAL [57] and OneLLM [26] 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 [4], deliberating over a series of slides [67], responding to queries necessitating cross-modal reasoning across images and tables [44], or executing a range of instruction-based tasks involving multiple image inputs [44]. 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



(a) X-Instruct Projection



(b) X-LLaVA-style Projection

Fig. 3: Projection types explored in the X-InstructBLIP Framework. (a) is an an instruction aware Q-Former projection [14] and (b) is a linear projection [48].

and video [1,39], 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 [10, 11, 28, 33, 40, 42, 43, 61], even in cross-modal settings [20, 25, 58], 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 [14,48] to an arbitrary number of modalities through independent fine-tuning of modality-specific projections to a frozen LLM, further broken down in Algorithm 1. X-InstructBLIP's alignment framework involves the following steps: **(1)** For each modality, collect an instruction tuning dataset suite $(x, y) \in \mathbb{D}_M$ s.t. $x = (x_M, x_T)$ is a tuple of a modality input and text, and y is the expected text output. **(2)** Let Enc_M be a modality encoder to R^{d_M} and Enc_T be a mapping from text to the LLM's embedding space R^{d_L}. Optimize a single separate projection module $f_\theta^M : R^{d_M} \rightarrow R^{kd_L}$ for each modality M on $\mathbb{D}_M$ while maintaining the parameters of the LLM frozen, where k is the number of LLM input tokens corresponding to the non-linguistic input. For sequential data, such as video and audio, we extract $N \times k$ query tokens; each frame is encoded and processed separately by the projection module. **(3)** The model is optimized under a causal language modeling objective [49]: $\min_\theta \mathcal{L}_{\text{CE}}\Big(\text{LLM}(x_{\text{LLM}}), \text{h}(y)\Big)$ where $\mathcal{L}_{\text{CE}}$ is the cross entropy loss, θ the Q-Former parameters, y is the target sequence, $\text{LLM}(x_{\text{LLM}})$ is the LLM's prediction.

X-Instruct Projection: Figure 3a highlights all components associated with learning instruction aware Q-Former projections [14] 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 $\mathbb{I} = \{I_{M_t} : M \in \mathbb{M}, t \in \mathbb{T}\}$ for each task $t \in \mathbb{T}$

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

Given a modality M encoding $z_M = \text{Enc}_M(x_M)$ and task instruction $i_M \in \mathbb{I}_{M_t}$, the Q-Former module transforms a set of k learnable embeddings $\mathbf{Q}_M = \{\mathbf{q}_{M_1} \dots \mathbf{q}_{M_K}\}$ termed *input query tokens* into instruction-aware representations of $\mathbf{Q}'_M = \text{QF}_M(\mathbf{Q}_M, z_M, i_M)$. The Q-Former module consists of two transformer submodules that share the same self-attention layers: one submodule interacts with the output of the modality encoder Enc_M and the other is a $\text{BERT}_{\text{base}}$ text transformer that serves as both an encoder and decoder. Each Q-Former is initialized with the pre-trained weights from BLIP-2 [40], 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 [48] to cater multiple modalities, similarly to the instruction aware Q-Former. Figure 3b depicts the architecture of this simple projection which linearly transforms the outputs of the modality encoder directly to the input embedding space of the LLM. Formally, the model consists of a single linear projection layer $\text{LP}_M : R^{d_M} \rightarrow R^{kd_{\text{LLM}}}$, where d_{LLM} is the LLM's embedding dimension. To compare the two projection types we match the number of trainable parameters for each modality, and maintain the training set-up.

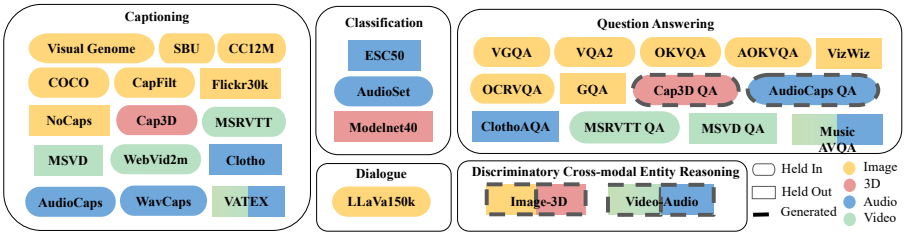


Fig. 4: Instruction Tuning and Evaluation Datasets: Oval-enclosed and square datasets are tuning and out-of-domain evaluation datasets respectively. Dashed outline is used for automatically derived datasets as described in Section 4.1.

4 Datasets

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

4.1 Fine-tuning Datasets

Existing Datasets: Figure 4 illustrates the datasets utilized for both instruction tuning and evaluation. A detailed breakdown of the dataset statistics and formats can be found in the supplementary material. For each dataset in $\mathbb{D}_M$, the collection of held-in datasets specific to modality M, a modified sampling strategy from [14] is adopted accommodating a broader range of modalities. The sampling probability for any given dataset $D_{M_d} \in \mathbb{D}_M$ is $\frac{\sqrt{|D_{M_d}|}}{\sum_{d \in [1 \ldots |\mathbb{D}_M|]} \sqrt{|D_{M_d}|}}$, with minimal adjustments as justified in the supplementary matierial.

Instruction Data Augmentation: Extracting instruction-aware representations necessitates diverse instruction-related tasks across all modalities. Notably, datasets for 3D and audio modalities are marjorly caption-centric. To address this, we leverage the open-source large language model `google/flan-t5-xxl` [79] to automatically generate question-answer pairs for the 3D and audio modalities from their corresponding captions. The process begins by prompting the model with captions to generate potential answers. These answers are then used to prompt the model to generate candidate questions. If the model's response to a question, using the caption as context aligns closely with the initial answer, the example is added to our dataset, yielding ∼250k 3D examples from

[4] The term *discriminative reasoning*, adapted from [85], refers to the ability to distinguish between sets of inputs, as opposed to *joint reasoning*, the synthesis of information from aligned sources.

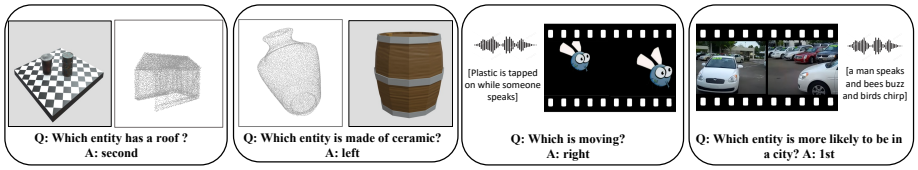


Fig. 5: `DisCRn`. Given two distinct modality inputs, select which one fits the query.

Cap3D [53] [5] and ∼24k audio examples from AudioCaps [34]. Details about the data generation and distribution are provided in the supplement.

4.2 Discriminative Cross-modal Reasoning

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

To generate the dataset, we prompt `google/flan-t5-xxl` in a Chain-of-Thought [80] 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 [6]. 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 [53]. 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 [38] atop of the Vicuna v1.1 7b and 13b models [12]. We adopt `EVA-CLIP-ViT-G/14` [18] as the encoder for image and video, for audio $\texttt{BEATs}_{\texttt{iter3+}}$ [9] and for 3D `ULIP-2` [`PointBERT` backbone] [62]. In the X-Instruct setup, each Q-Former optimizes 188M trainable parameters and learns $K = 32$ query tokens with a a hidden dimension of size 768 to select a *single best model* per modality. Raw inputs undergo standardized pre-processing prior to encoding. All Q-Formers are pre-initialized with BLIP-2 stage-1 weights [14] 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.

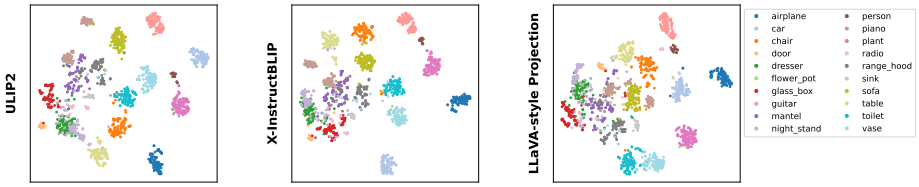
All models are optimized on 8 A100 40GB GPUs using AdamW [51] with β_1= 0.9, β_2= 0.999, and weight decay of 0.05. The learning rate warms up linearly over the initial 1,000 steps from 10^{-8} to 10^{-5}, followed by a cosine decay to a minimum of 0. Evaluation hyper-parameters and templates are consistent across tasks, minimally adapted to each modality as detailed in the supplement.

5.2 Results

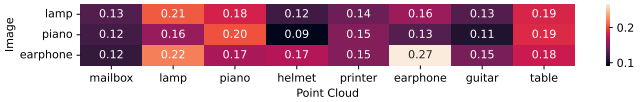
Our primary aim is to demonstrate the adaptability of our framework across various modalities without relying on large-scale pre-training stages or joint modality data. Nevertheless, to ensure our approach's effectiveness and comparability, we juxtapose its performance to other *methods that employ projections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

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

3D: Table 1 shows the results on zero-shot classification on ModelNet40 [81] under two setups: classification in closed vocabulary using loss ranking [42] and open generation where the model is prompted to `describe the 3d model` and



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



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

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

correctness is validated if a *single* class from the 40 candidates is present in the generation. In both projection setups, X-InstructBLIP significantly outperforms the InstructBLIP baseline, which processes a single view rendering of the point cloud. Interestingly, the X-Instruct projection setup outperforms not only the X-LLaVA-style projection but also PointLLM [24] that learns a similar projection but- unlike this set-up - employs RGB features. It also outperforms PointBindLLM [24] which trains an adapter on ImageBind [20] 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 [54] 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 [60], ClothoAQA [46], and Clotho [17], 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 [14], Table 4 presents results on image captioning, visual question

answering, MME [19], and MMVET [90] as a sanity check. While X-InstructBLIP outperforms InstructBLIP on VizWiz [5] 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) [14]	31.4	23.7
InstructBLIP (13b) [14]	31.5	25.5
Point-LLMv2+(RGB) (7b) [84]	-	32.3
Point-LLMv2+(RGB)(13b) [84]	-	31.8
PointBind-LLM (7b) [24]	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 [70] test set.

	PT	MSVD *test*	VATEX *val*	MSVDQA *test*
FrozenBiLM [86]	✓	-	-	33.8
VideoLLaMA [91]	✓	-	-	51.6
InstructBLIP [14]	×	87.2	57.6	41.2
X-LLaVA-style Proj. (7b)	×	105.3	46.2	49.8
X-Instruct Proj. (7b)	×	116.1	**59.2**	**51.7**
X-Instruct Proj. (13b)	×	**124.3**	52.0	49.2

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

	$ESC50_{close}$ Acc.	$ESC50_{open}$ Acc.	ClothoAQA EM	Clotho v1 CIDEr	Clotho v1 SPIDEr	Clotho v2 CIDEr	Clotho v2 SPIDEr
ImageBind [20]	66.9	×	×	×	×	×	×
MWAFM [22]	-	-	22.2	-	-	-	-
Pengi [15]	-	53.9	64.5	39.6	30.0	32.9	27.1
Kim et. al., 2023 [35]	-	-	-	-	-	19.2	13.3
ImageBind-LLM (7B) [27]	40.1	27.4	10.3†	3.7	5.5	3.2	5.5
X-LLaVA-style Proj. (7b)	67.4	20.3	**26.9**	25.3	16.6	22.0	14.8
X-Instruct Proj. (7b)	75.9	**38.2**	21.4	**29.4**	**19.5**	26.7	17.8
X-Instruct Proj. (13b)	**77.1**	34.6	21.7	28.7	18.8	**27.5**	**18.0**

Table 3: Out-Domain Audio Quantitative Results.

		Flickr30k	NoCaps	VizWiz	GQA	MME	MMVet
Vision	Flamingo 9B [2]	61.5	-	28.8	-	-	-
	BLIP-2 [40]	76.1	107.5	29.8	44.7	**1293.8**	22.4
	InstructBLIP [14]	**84.5**	**123.1**	34.5	**49.5**	1212.8	26.2
	MiniGPT4 (7b) [94]	-	-	-	32.2	1158.6	22.1
	LLaVA (7b) [48]	27.7	33.1	-	-	717.5	27.4
	LLaMA-adapter (13b) [92]	30.5	41.7	-	-	1222.0	-
X-Modal	PandaGPT (13b) [65]	23.0	29.7	-	-	871.2	19.6
	ImagebindLLM (7b) [27]	23.5	30.4	-	-	989.3	-
	X-LLaVA-style Proj. (7b)	6.2	22.3	27.7	41.5	866.7	17.0
	X-Instruct Proj. (7b)	82.1	117.7	34.9	48.1	891.8	29.0
	X-Instruct Proj. (13b)	74.7	114.5	**36.0**	49.2	1174.0	**35.1**

Table 4: Out-Domain Image Quantitative Results.

Single Modality Quantitative Results. Underlined numbers indicate in-domain evaluations. **Bold** indicates the top zero-shot performance. Blue indicates second best zero-shot performance. Purple denotes evaluations conducted independently. Models denoted with 7b and 13b indicate the underlying LLM size. Gray shaded rows correspond to X-InstructBLIP variants, and Yellow to the LLaVA-style [48] model equivalent. CIDEr score [71] is reported for captioning, accompanied by SPIDEr [50] 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 [91] and FrozenBiLM [86] 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 [39] and VATEX [77]. 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 [14] to `Describe the image/video.` For 3D, a randomly chosen rendering view of the point cloud is provided to InstructBLIP. For video we follow [14] and sample four frames and concatenate their output representations as input to the model. For audio we use WavCaps [82].

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

5.3 Ablations

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

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

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

Table 7: Ablation: Prefix Effect

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

Table 8: Out-Domain Audio Quantitative Results.

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

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

6 Conclusion

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

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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 visual questions. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. pp. 333–342 (2010)
6. 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)
7. 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)
8. 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)
9. 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
10. 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)
11. 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)
12. 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)
13. 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)
14. 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
15. Deshmukh, S., Elizalde, B., Singh, R., Wang, H.: Pengi: An audio language model for audio tasks. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=gJLAfO4KUq
16. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: PaLM-e: An embodied multimodal language model. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 8469–8488. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/driess23a.html
17. 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)
18. 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)
19. 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)
20. 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)
21. Gong, Y., Luo, H., Liu, A.H., Karlinsky, L., Glass, J.R.: Listen, think, and understand. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=nBZBPXdJlC
22. Guangyao li, Yixin Xu, D.H.: Multi-scale attention for audio question answering. Proc. INTERSPEECH (2023)
23. 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)
24. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., et al.: Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:2309.00615 (2023)
25. 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)
26. 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)
27. 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)

28. 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)
29. 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
30. 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)
31. 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
32. 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)
33. 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
34. 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)
35. 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)
36. 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
37. 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
38. 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
39. 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)
40. 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)
41. 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
42. 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)
43. 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)
44. Li, Y., Li, W., Nie, L.: MMCoQA: Conversational question answering over text, tables, and images. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4220–4231. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.290, https://aclanthology.org/2022.acl-long.290
45. 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)
46. Lipping, S., Sudarsanam, P., Drossos, K., Virtanen, T.: Clotho-aqa: A crowdsourced dataset for audio question answering. In: 2022 30th European Signal Processing Conference (EUSIPCO). pp. 1140–1144. IEEE (2022)
47. Liu, H., Yan, W., Abbeel, P.: Language quantized autoencoders: Towards unsupervised text-image alignment. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=mlxRLIy7kc
48. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=w0H2xGHlkw
49. 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-
50. 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)
51. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
52. 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

53. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3d captioning with pretrained models. In: Proceedings of the NeurIPS 2023 (2023)
54. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
55. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2023)
56. 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
57. 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)
58. 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)
59. 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
60. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)
61. 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)
62. Salesforce: Ulip. https://github.com/salesforce/ULIP (2022), accessed: 2023-07-1
63. 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)
64. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual llm for video understanding. arXiv preprint arXiv:2312.06720 (2023)
65. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: PandaGPT: One model to instruction-follow them all. In: Hazarika, D., Tang, X.R., Jin, D. (eds.) Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23. Association for Computational Linguistics, Prague, Czech Republic (Sep 2023), https://aclanthology.org/2023.tllm-1.2
66. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=mL8Q9OOamV
67. 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)

68. 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
69. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
70. 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)
71. 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)
72. 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)
73. 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)
74. 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)
75. 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)
76. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
77. 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)
78. 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)
79. 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
80. 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)
81. 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)
82. XinhaoMei: Wavcaps. https://github.com/XinhaoMei/WavCaps (2023), accessed: 2023-07-1

83. 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)
84. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds (2023)
85. 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
86. 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)
87. 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)
88. 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)
89. 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
90. 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)
91. 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)
92. 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
93. 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)
94. 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