Mixture of Probes: Learning with Privileged Modalities in Multimodal LLMs Through Probing

Dominick Reilly[1,2*] **Qiyu Wu**[1†] **Hiromi Wakaki**[1] **Srijan Das**[2] **Yuki Mitsufuji**[1]
[1] Sony Group Corporation
[2] University of North Carolina at Charlotte
dreilly1@charlotte.edu
{qiyu.wu, hiromi.wakaki, yuhki.mitsufuji}@sony.com
https://github.com/Sony/MoP

Abstract

Multimodal Large Language Models (MLLMs) are typically designed under the assumption that all modalities available during training will also be accessible at inference. However, many real-world settings violate this assumption, requiring models to operate under a *privileged modality setting*, where auxiliary modalities are available only during training. While these modalities contain valuable information, existing MLLMs largely fail to leverage them effectively, as they treat modalities as interchangeable inputs rather than sources of complementary supervision. We propose **Mixture of Probes (MoP)**, a novel framework that disentangles *modality-specific* and *modality-general* signals within the MLLM, allowing the model to preserve modality-dependent structure while learning transferable representations across modalities. At its core, MoP achieves this through a structured probing mechanism that extracts and organizes information from intermediate representations of a shared modality encoder, rather than relying only on final-layer alignment as done in existing MLLMs. To support this disentanglement, we further introduce **MoP Cross-modal Training (MoP-X)**, a training strategy for MoP centered around a probe disentanglement loss that prevents probe collapse and encourages cross-modal learning. We evaluate MoP across two domains spanning eight tasks and four modalities under a comprehensive evaluation protocol tailored to the privileged modality setting, where each modality is independently treated as the sole input at inference time. MoP consistently outperforms strong MLLM baselines, achieving up to 65% relative improvement, demonstrating that auxiliary modalities, even when unavailable at inference, can provide substantial gains when effectively leveraged during training.

1 Introduction

Stemming from the success of large language models (LLMs) [39, 4, 38], recent multimodal large language models (MLLMs) have enabled language-based interaction with diverse sensory inputs such as vision, audio, and depth [48, 15, 28, 42]. Yet real-world scenarios often exhibit an asymmetric structure: the sensors that are available for training are not always available for deployment. A model may be trained with rich auxiliary streams, such as egocentric video, depth, or audio, but later be required to operate from a single cheap, reliable, or non-invasive modality. This creates a *privileged modality setting*, where auxiliary modalities provide supervision during training but are unavailable at inference. For example, in activities of daily living (ADL), training data may include egocentric video from wearable devices [12, 11], depth from specialized sensors [16], and exocentric video from static cameras [10, 6], while deployment is typically restricted to exocentric video alone.

[*] Work done during internship at Sony Group Corporation
[†] Corresponding Author: qiyu.wu@sony.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

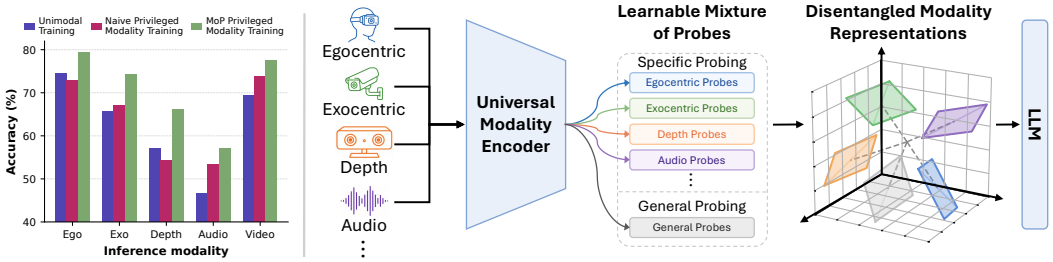


Figure 1: **(left)** MoP improves single-modality inference across five inference settings, outperforming both unimodal MLLMs and naive multimodal training. **(right)** MoP achieves this through a structured probing mechanism with modality-specific and modality-general probes, which disentangle modality representations before integration with the LLM.

Existing MLLMs [49, 20, 36, 2] are not designed to address this setting. They primarily view them as interchangeable inputs rather than treating modalities as sources of complementary supervision, exclusively aiming only to align each modality into a shared representation space [30, 51, 26, 14]. However, in the privileged modality setting, it is not sufficient for a model to merely align each modality; it must leverage information from modalities available only during training and transfer this knowledge to improve performance on the modality available at inference. Most methods achieve such alignment with modality-specific encoders [30, 51, 26], or through the more scalable unified modality encoder paradigm [14]. We identify two key limitations of these MLLMs in the privileged modality setting: (i) **sub-optimal modality representations** - new modalities are typically incorporated via shared encoders with limited modality-specific adaptation at the final layers, neglecting intermediate features that capture low-level modality-specific signals [33]; (ii) **unstructured cross-modal representations** - existing approaches lack explicit mechanisms to disentangle modality-general and modality-specific information, hindering effective cross-modal knowledge transfer [43].

To this end, we propose **Mixture of Probes (MoP)**, an MLLM architecture for the privileged modality setting that improves cross-modal knowledge transfer by encouraging disentangled *modality-specific* and *modality-general* representations to be learned through a structured probing mechanism. MoP treats all modalities symmetrically and designates none as privileged; privilege arises solely from which modality is available at deployment. Inspired by evidence from neuroscience that the human brain separates modality-specific maps from modality-general abstractions across sensory streams [8], MoP is built around the idea that effective cross-modal transfer requires preserving what is unique to each modality while isolating what can be shared across them. This intuition is further supported by the Modality Focusing Hypothesis [43], which finds that cross-modal transfer depends primarily on modality-general decisive features rather than modality-specific signals that do not transfer across modalities. Concretely, MoP introduces a structured probing mechanism that leverages intermediate representations throughout a shared universal encoder, enabling richer extraction of both modality-dependent and shared signals across the model depth. Modality-specific probes are conditioned only on individual modalities to capture modality-specific structure, while modality-general probes are conditioned on all modalities to capture general cross-modal structure, illustrated in Figure 1. A key component of our design is **MoP Cross-modal Training (MoP-X)**, which consists of two components: (i) a *probe disentanglement loss* that plays a central role by explicitly encouraging disentanglement between the modality-specific and modality-general probes and prevents probe collapse; (ii) *modality-interleaved batching* that ensures consistent cross-modal interaction during training, unlike prior methods that train each modality sequentially or in isolation [30, 51].

We evaluate MoP on **two domains** spanning **8 tasks** and **4 modalities**. In the *Activities of Daily Living* (ADL) domain, MoP is trained using egocentric video, exocentric video, and depth. In the *Music Understanding* domain, MoP is trained on paired audio-visual data. In each domain, we perform a comprehensive evaluation under the privileged modality setting, where each modality is independently treated as the sole input at inference time. Across both domains and all inference conditions, MoP consistently outperforms state-of-the-art MLLM baselines, demonstrating that auxiliary modalities, when properly leveraged, can substantially improve single-modality performance.

A summary of our contributions follows:

- We formalize the **privileged modality setting** for MLLMs, where auxiliary modalities enrich training but are absent at inference, and demonstrate that standard multimodal training fails to effectively leverage this asymmetry.
- We propose **Mixture of Probes (MoP)**, a novel probing-based architecture that learns disentangled modality-specific and modality-general representations through learnable probing across the depth of a modality encoder.
- We introduce **MoP Cross-modal Training (MoP-X)**, a training strategy complementing MoP via a probe disentanglement loss to separate modality-specific from modality-general representations, and promote cross-modal interaction through modality-interleaved batching.
- We demonstrate the effectiveness of MoP across two domains, eight tasks, and four modalities under the privileged modality setting, consistently outperforming strong MLLM baselines and highlighting the benefits of leveraging privileged modalities.

2 Related Works

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) extend LLMs beyond text by mapping sensory inputs (e.g., vision, video) into a shared language representation space [21, 3, 7, 25, 45]. To support multiple modalities, many approaches introduce specific modality encoders and alignment modules to connect each modality to the LLM [48, 5, 51, 26, 37, 15, 30, 34], while others such as OneLLM use a shared universal encoder across modalities [14], motivated by previous works demonstrating the robustness of transformer representations [23, 50]. However, these methods are primarily designed for *modality alignment*, treating each modality as an input to be independently mapped into the LLM space. In contrast, we study the *privileged-modality* setting, where auxiliary modalities are available only during training and must improve a single-modality at inference, requiring explicit cross-modal knowledge transfer rather than alignment alone.

Learnable Token-based Adaptation Mechanisms. A related line of work adapts pretrained vision-language models using small sets of learnable tokens that interface with frozen backbones. Methods such as BLIP-2 [19] and InstructBLIP [7] introduce query tokens to extract task-relevant information from a vision encoder and align it with an LLM, while prompt tuning approaches adapt transformers through learnable tokens in the input without modifying the backbone [17]. However, these approaches treat learned tokens as a shallow interface to the model, typically interacting only with final encoder representations and capturing a single undifferentiated signal. In contrast, our approach uses probing as a mechanism for *structured representation learning*, inserting probes throughout the encoder to extract information at different levels of abstraction and explicitly disentangling modality-general and modality-specific signals. This enables the model to not only adapt, but to organize information to support effective cross-modal transfer in the privileged-modality setting.

Learning with Missing or Privileged Modalities. Learning with missing or privileged modalities has been widely studied across multimodal learning, domain adaptation, and knowledge distillation. The classical Learning Using Privileged Information (LUPI) framework [40] introduces auxiliary modalities available only at training time, later extended to deep learning via generalized distillation [22]. In multimodal settings, related ideas appear in cross-modal distillation and supervision, where one modality guides another (e.g., RGB $\rightarrow$ depth, audio $\rightarrow$ vision) [13, 1, 29]. More recent work addresses the *missing modality* setting, focusing on robustness when inputs are incomplete through modality dropout, reconstruction, or invariant fusion strategies [27, 24, 41], with emerging extensions to multimodal LLMs that compensate for missing inputs via retrieval or tool use [31].

Despite these advances, existing approaches primarily emphasize robustness to missing modalities or reconstruction of absent inputs, rather than explicitly leveraging auxiliary modalities to improve a target modality at inference. Moreover, existing methods are developed for task-specific multimodal models, making them non-trivial to adapt to MLLMs, where modalities must be integrated through a shared language-centered representation space. In contrast, we focus on *privileged-modality transfer* for MLLMs, where auxiliary modalities are available only during training and are used to directly improve single-modality performance at inference.

3 Problem Formulation

Let $\mathcal{M} = \{1, \ldots, M\}$ denote the set of modalities available during training. Each sample in our training dataset consists of multi-modal observations of the same scene

$$\mathbf{X} = \{X^m\}_{m \in \mathcal{M}}, \tag{1}$$

where X^m denotes the input from modality m (e.g., egocentric video, exocentric video, depth, or audio), together with a language query q and a target sequence $y = (y_1, \ldots, y_T)$.

We consider the *privileged modality* setting, in which all modalities are accessible during training, but only a *single* modality $X^{(m^\star)}$, where $m^\star \in \mathcal{M}$, is available at test time. Privilege is a property of the training and deployment protocol rather than of any particular modality: $m*$ is unknown during training, and whichever modalities are absent at inference act as privileged relative to $m*$. We aim to learn a predictor

$$f : (X^{(m^\star)}, q) \mapsto y, \tag{2}$$

that benefits from additional modalities during training despite their absence at inference. Formally, we aim to leverage the full multi-modal training distribution to learn representations or inductive biases that improve generalization under single-modality inference. This differs from standard multi-modal learning, which assumes $\{X^m\}_{m\in\mathcal{M}}$ are available at both training and test time, and instead aligns with the paradigm of learning with privileged information, where auxiliary modalities act as training-time-only supervision.

4 Method

4.1 Preliminaries

Recent MLLMs increasingly adopt a *unified architecture*, in which a universal encoder maps heterogeneous input modalities into a shared representation space, rather than relying on separate encoders for each modality. In this paradigm, lightweight modality-specific tokenizers first convert raw inputs into token sequences, which are processed by a shared universal modality encoder. The resulting features are projected into the language embedding space and fed to an LLM together with a text instruction. Below, we describe each component of this unified MLLM framework in more detail.

Modality-specific tokenizers. For an input observation X^m from modality $m \in \mathcal{M}$, a modality-specific tokenizer $\tau^m(\cdot)$ first converts the raw input into a sequence of modality tokens

$$\tilde{Z}^m = \tau^m(X^m) \in \mathbb{R}^{N^m \times d_v} \tag{3}$$

where N^m is the number of input tokens and d_v is the hidden dimension of the Universal modality encoder. These tokenizers account for modality-dependent input structure while mapping all modalities into a universal token embedding space.

Universal modality encoder. All tokenized modality inputs are processed by a shared universal encoder E composed of L Transformer layers. Given the tokenized input $\tilde{Z}^m$ for modality m, the encoder produces a sequence of hidden representations

$$H_l^m = E_l(H_{l-1}^m), \qquad l = 1, \ldots, L, \tag{4}$$

with $H_0^m = \tilde{Z}^m$ and $H_l^m \in \mathbb{R}^{N^m \times d_v}$. Encoder parameters are shared across all modalities, enabling inputs from heterogeneous sensor streams to be mapped into a common latent representation space.

Universal modality connector. The final-layer encoder representation H_L^m is projected into the LLM embedding space through a universal modality connector $g(\cdot)$:

$$Z^m = g(H_L^m) \in \mathbb{R}^{N^m \times d_{\text{lm}}} \tag{5}$$

where d_{lm} is the hidden dimension of the LLM. The connector aligns encoder features with the token embedding space of the language model, enabling inputs from different modalities to be consumed through a unified interface.

LLM. The projected modality tokens Z^m are concatenated with the text tokens q and provided to the LLM for autoregressive generation. In this way, the LLM operates over a unified sequence consisting of modality-conditioned visual tokens and textual context, producing the output response in the standard causal decoding manner.

However, such unified MLLMs exhibit several limitations. **First,** the universal modality encoder is typically initialized from a pretrained unimodal model, kept frozen, and adapted to new modalities using only its final-layer representations, ignoring useful modality-dependent signals present in lower-level features. **Second,** they do not explicitly disentangle modality-general and modality-specific information, which restricts their ability to capture complementary signals and transfer knowledge across modalities [43]. **Third,** existing training procedures do not explicitly encourage the learning of complementary information across modalities, as modalities are often treated independently or without structured interactions during training.

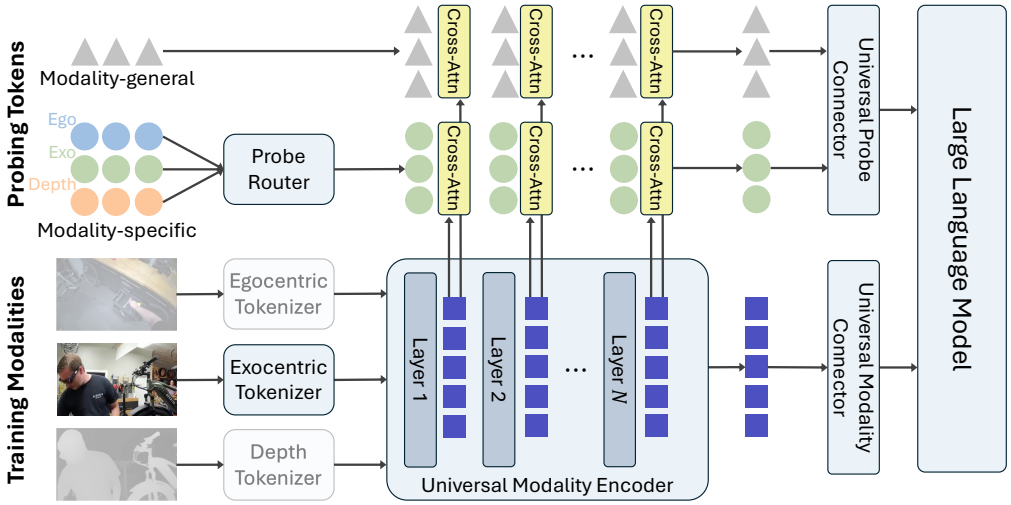


Figure 2: **MoP Architecture.** In this example, three modalities are shown: egocentric, exocentric, and depth. Each sample in a training batch consists of a single modality, in this example that modality is the exocentric video.

4.2 Mixture of Probes (MoP)

We introduce *Mixture of Probes (MoP)*, a probing-based architecture that augments a universal modality encoder with learnable tokens operating on intermediate representations across layers. By probing the encoder beyond its final-layer features, MoP extracts salient cues that may not be preserved in the final representation. Motivated by the modality focusing hypothesis [43], MoP learns two complementary probe families: modality-specific probes, which capture signals unique to each modality, and modality-general probes, which capture shared structure across modalities. This design allows the model to represent both shared and modality-specific information within a unified encoder to facilitate effective cross-modal representation learning. We illustrate the MoP architecture in Figure 2 and present each component in more detail below.

Modality-specific and modality-general probes. For each modality $m \in \mathcal{M}$, we define a set of modality-specific probes $P_s^m \in \mathbb{R}^{K_s \times d_v}$ where K_s is the number of modality-specific probes. In addition, we define a shared set of modality-general probes $P_g \in \mathbb{R}^{K_g \times d_v}$ where K_g is the number of modality-general probes. Both probe sets lie in the same hidden space as the universal modality encoder and are implemented as a set of learnable tokens.

Layer-wise probe interaction. Given tokenized modality input $\tilde{Z}^m$, a deterministic probe router selects the corresponding modality-specific probes for modality m, which are passed along with the modality-general probes through every layer of the universal modality encoder and updated through cross-attention with the encoder hidden states. Let $P_{s,l}^m$ and $P_{g,l}$ denote the modality-specific and modality-general probes after interaction with encoder layer l, respectively. We initialize $P_{s,0}^m = P_s^m$ and $P_{g,0} = P_g$. Then, for each layer $l = 1, \ldots, L$, the probes are updated as

$$P_{s,l}^m = \mathrm{CrossAttn}_s^{m,l}(P_{s,l-1}^m, H_l^m); \qquad P_{g,l} = \mathrm{CrossAttn}_g^l(P_{g,l-1}, H_l^m) \tag{6}$$

where the probe tokens act as queries and the encoder hidden states H_l^m act as keys and values. Here, $\mathrm{CrossAttn}_s^{m,l}$ denotes the modality-specific cross-attention module for modality m at layer l, while $\mathrm{CrossAttn}_g^l$ denotes the corresponding cross-attention module used to update the shared modality-general probes. In this way, the probes iteratively extract modality-relevant information from the universal encoder's intermediate representations across depth.

Importantly, modality-general probes are shared across modalities but interact with one modality per sample, so cross-modal transfer is achieved indirectly through shared parameter updates across modality-interleaved training iterations. This design is intentional, as performing interaction between jointly would substantially increase the token budget and exceed the context length of the MLLM.

Probe-conditioned language tokens. After the final encoder layer L, we obtain the final modality-specific and modality-general probe representations $P_{s,L}^m \in \mathbb{R}^{K_s \times d_v}$ and $P_{g,L} \in \mathbb{R}^{K_g \times d_v}$. The probe

tokens are concatenated to form the full probe representation P^m, and then projected into the LLM embedding space through a shared probe connector $g_p(\cdot)$:

$$Z_p^m = g_p(P^m) \in \mathbb{R}^{(K_g+K_s)\times d_{\text{lm}}} \tag{7}$$

The resulting probe tokens are then concatenated with the projected universal encoder tokens to obtain $\hat{Z}^m = [Z^m; Z_p^m]$, and fed to the LLM. Thus, the LLM receives both the backbone modality tokens and the additional probe tokens that encode modality-specific and modality-general signals.

4.3 MoP Cross-modal Training (MoP-X)

We train MoP to encourage cross-modal representation transfer in the privileged modality setting. While the probing architecture of MoP exposes both modality-specific and modality-general signals, effective transfer requires these representations to remain distinct. MoP-X enforces this through a probe disentanglement loss, along with a simple modality-interleaved batching strategy to ensure consistent cross-modal interaction during training. During training, we optimize the modality-specific tokenizers, the modality-language connector, the MoP components, and the LLM through LoRA, while keeping the universal modality encoder frozen.

Probe Disentanglement Loss. Although the modality-specific and modality-general probes are designed to capture complementary information, this separation is not guaranteed by architecture alone and in practice, probes may collapse toward similar representations. To mitigate this, we introduce a probe disentanglement loss that penalizes inter and intra-probe similarity within each encoder layer. Formally, let

$$P_l^m = [P_{g,l}\,;\, P_{s,l}^m] \in \mathbb{R}^{K\times d_v}, \tag{8}$$

denote the concatenated modality-general and modality-specific probes at layer l, where $K = K_g + K_s$. We perform normalization along the feature dimension and compute the cosine similarity:

$$S_l^m = \bar{P}_l^m (\bar{P}_l^m)^\top \in \mathbb{R}^{K\times K}, \tag{9}$$

where $\bar{P}_l^m$ denotes the row-wise ℓ_2 normalized probe matrix. We then encourage *only* the off-diagonal entries of S_l^m. Let $D \in \mathbb{R}^{K\times K}$ denote the matrix with zeros on the diagonal and ones off-diagonal, the probe probe disentanglement loss is computed by averaging the layer-wise loss across layers:

$$\mathcal{L}_{\text{dis}} = \frac{1}{LK(K-1)} \sum_{l=1}^{L} \|D \odot S_l^m\|_1, \tag{10}$$

Modality-interleaved Batching Strategy. We construct mini-batches by sampling examples across multiple modalities, as opposed to previous methods that process each modality sequentially or in isolation [51, 30]. This exposes shared model components to heterogeneous modality signals within a single optimization step. While each training sample contains a single modality, interaction between modality-specific and modality-general components is facilitated through shared representations, rather than explicit cross-modal coupling at the sample level. We favor indirect transfer through shared components over explicit paired objectives for two reasons: jointly aligning time-synchronized modalities induces conflicting gradients [34], and single-modality samples match the inference distribution while avoiding a multi-modality forward pass per sample.

Autoregressive Loss. Given an input from modality $m \in \mathcal{M}$, let $\hat{Z}^m$ denote the final sequence of projected encoder and probe tokens produced by MoP, as defined in Section 4.2. Conditioned on $\hat{Z}^m$ and q, the model is trained with the standard autoregressive objective

$$\mathcal{L}_{\text{AR}} = -\sum_{t=1}^{T} \log p_\theta\Big(y_t \mid \hat{Z}^m, q, y_{<t}\Big) \tag{11}$$

where θ denotes the trainable parameters of the model and $y_{<t}$ represents the subsequence of tokens preceding position t. The final loss used for training is the λ-weighted combination of the probe disentanglement loss and autoregressive loss:

$$\mathcal{L} = \mathcal{L}_{\text{AR}} + \lambda \mathcal{L}_{\text{dis}} \tag{12}$$

Table 1: **Results on Ego-in-Exo Perception.** Results are reported on the four tasks of Ego-in-Exo Perception under single modality inference with egocentric, exocentric, or depth modalities. Tasks are abbreviated as: Action Und. = *Action Understanding*, Task Regions = *Task Relevant Regions*, HOI = *Hand Object Interactions*, and Hand Ident. = *Hand Identification*.

Method	Egocentric Inference					Exocentric Inference					Depth Inference				
	Action Und.	Task Regions	HOI	Hand Ident.	Avg	Action Und.	Task Regions	HOI	Hand Ident.	Avg	Action Und.	Task Regions	HOI	Hand Ident.	Avg
Trained without privileged modalities															
Unimodal MLLM [47]	77.04	80.49	76.03	64.74	74.57	59.50	66.60	71.60	65.00	65.68	44.89	59.88	59.74	63.71	57.06
Trained with privileged modalities (Ego+Exo+Depth)															
X-InstructBLIP [30]	77.31	76.22	76.15	65.38	73.77	61.31	68.54	71.31	65.12	66.57	46.90	50.85	62.34	65.12	56.30
OneLLM [14]	35.14	35.73	47.93	63.71	45.63	23.17	35.73	46.28	62.16	41.84	21.02	32.44	44.27	62.29	40.01
Naive MLLM [47]	75.30	75.98	75.44	65.12	72.96	60.68	69.39	73.08	64.99	67.04	41.06	52.20	59.74	64.48	54.37
MoP (Ours)	84.48	86.95	80.76	65.64	**79.46**	82.68	71.68	76.86	65.51	**74.18**	56.23	76.10	67.53	64.99	**66.21**

5 Experiments

5.1 Experimental Setup

Benchmarks We evaluate our approach on two distinct task domains: Activities of Daily Living (ADL) understanding and Music Understanding. For ADL understanding, we evaluate on **Ego-in-Exo Perception** [35], a benchmark containing 3,881 questions spanning four task categories, where each question corresponds to a time-synchronzied ego-exo video pair. We evaluate models using a single input modality per question, allowing us to assess performance under different inference sensors. Specifically, we consider egocentric video, exocentric video, and depth (obtained from exocentric video using DepthAnythingV2 [44]) as separate evaluation settings.

For Music Understanding, we evaluate on **Music-AVQA** [18], which contains 9,185 questions across five music understanding tasks. Although each question is paired with audio and video modalities, we evaluate each question using a single input modality per sample to stay consistent with our problem setting. We report results under both audio and video inference, allowing controlled analysis of how information from privileged modalities at training time transfers to different inference modalities.

Compared Models. We compare MoP against three baseline settings. (i) **Unimodal MLLM**, trained using only the target inference modality without access to privileged modalities. (ii) **Naive MLLM**, trained on the same modality-interleaved data stream as MoP, where each sample contains a single modality with its instruction, but without probing or the disentanglement loss. (iii) **Existing MLLMs**, including X-InstructBLIP [30] and OneLLM [14], which represent state-of-the-art approaches to multimodal alignment. All methods leveraging privileged modalities share the same training data, differing only in how cross-modal information is modeled.

MoP Implementation Details We instantiate the universal modality encoder as SigLIP [46] and the language model as Qwen 2.5 [32], both initialized from large-scale pretrained weights [47]. We train on instruction pairs from the EgoExo4D dataset [35] for ADL understanding, and the training split of Music-AVQA for music understanding. Models are trained for 1 epoch on 4 NVIDIA H200 GPUs.

During training, each sample is randomly assigned a single input modality, resulting in mixed-modality batches that expose the model to diverse modalities within each optimization step (Section 4.3). The universal modality encoder is kept frozen and we train the modality-specific tokenizers, the universal modality-language connector, and the MoP components, while adapting the language model with LoRA of rank $r = 16$. We use a learning rate of 10^{-5} for the universal connectors and LoRA parameters, and 2×10^{-6} for the modality-specific tokenizers and probe cross-attention modules. Unless otherwise specified, we use 8 modality-specific probes and 8 modality-general probes. We initialize the probes from $\mathcal{N}(0, 0.02)$, and the probe disentanglement loss weight is set to $\lambda = 100$.

5.2 Main Results

Ego-in-Exo Perception. Table 1 reports results on Ego-in-Exo Perception across egocentric, exocentric, and depth inference. We observe that unimodal baselines consistently outperform the Naive MLLM baseline on their corresponding inference modality (e.g., Unimodal MLLM achieves 74.57% vs. 72.96% for the Naive MLLM under egocentric inference). Despite being trained on additional modalities the naive MLLM fails to improve performance, suggesting that simply combining modalities without structured interaction does not result in effective knowledge transfer.

In contrast, MoP significantly improves performance across all inference modalities, achieving gains of +6.50%, +7.14%, and +11.84% over the Naive MLLM for egocentric, exocentric, and depth inference, respectively. Notably, MoP also outperforms the corresponding modality-only models

Table 2: **Results on Music-AVQA.** Results are reported under single-modality inference. Audio-answerable questions are evaluated using audio alone, and video-answerable questions are evaluated using video alone, following the task definitions of Li et al. [18].

Method	Audio Inference			Video Inference		
	Counting (A)	Comparative (A)	Avg	Counting (V)	Location (V)	Avg
Trained without privileged modalities						
Unimodal MLLM [47]	44.84	48.65	46.75	80.40	58.22	69.31
Trained with privileged modalities (Audio+Video)						
X-InstructBLIP [30]	42.18	52.05	47.12	45.12	59.84	52.48
OneLLM [14]	59.98	47.47	53.73	65.58	69.71	67.65
Naive MLLM [47]	53.13	53.74	53.44	83.49	64.25	73.87
MoP (Ours)	62.05	52.36	**57.21**	83.90	71.25	**77.58**

for all three inference settings. Since these baselines are trained directly on the target inference modality alone, this improvement shows that MoP successfully leverages privileged modalities during training to improve single-modality inference. MoP also outperforms both X-InstructBLIP [30] and OneLLM [14] across all three inference modalities, showing that its gains are not simply due to stronger multimodal alignment, but to more effective use of privileged modalities.

Music-AVQA. Table 2 reports results on Music-AVQA under audio and video inference. Unlike the ADL setting, naive multimodal training already improves over the unimodal baselines (53.44 vs. 46.75 under audio inference and 73.87 vs. 69.31 under video inference), indicating that paired audio-visual data is complementary even without structured interaction. MoP builds substantially on this, achieving further gains of +3.77 and +3.71 points over the Naive MLLM under audio and video inference, respectively. Notably, MoP also outperforms recent alignment-based MLLMs, including X-InstructBLIP and OneLLM, under both audio and video inference. While these methods are trained with privileged modalities, their objective is primarily to align each modality with the LLM space, rather than to structure cross-modal representations for single-modality inference. The consistent gains of MoP suggest that alignment alone is not sufficient in the privileged modality setting; explicitly separating modality-specific and modality-general information provides a more effective mechanism for transferring knowledge from auxiliary modalities.

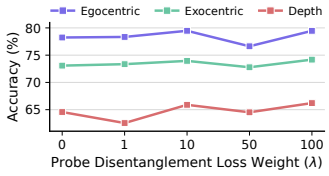


Figure 3: **Ablation on probe disentanglement loss weight.**

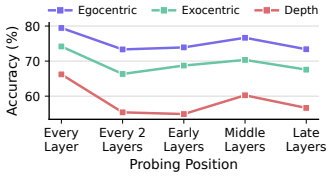


Figure 4: **Ablation on position of probing layers.**

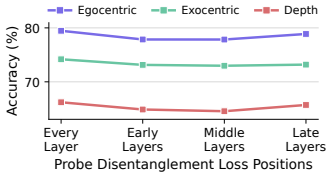


Figure 5: **Ablation on application of disentanglement loss.**

5.3 Ablation Study

We perform all ablations on the three inference settings of the Ego-in-Exo Perception benchmark. Additional ablations are presented in the Appendix.

Effect of probing types. Table 3 studies the contribution of the two probe families by comparing the naive MLLM, MoP with only modality-specific probes, and MoP with both modality-specific and modality-general probes. Adding modality-specific probes already yields clear gains across all three inference modalities, showing that intermediate-layer probing is effective even when restricted to modality-dependent information. Incorporating modality-general probes provides a further improvement in every case, leading to the best overall performance on egocentric, exocentric, and depth inference. These results support the core design of MoP: modality-specific probes capture information unique to each modality, while modality-general probes capture shared structure that is complementary.

Table 3: **Ablation of Probe Types.** Impact of modality-specific and modality-general probing across different inference modalities.

Probing type	Inference Modality		
	Egocentric	Exocentric	Depth
Naive MLLM (No Probing)	72.96	67.04	54.37
+ Modality-specific	76.72 (+3.76)	72.87 (+5.83)	63.22 (+8.85)
+ Modality-general	**78.24 (+5.28)**	**73.08 (+6.04)**	**64.57 (+10.2)**

Effect of probe disentanglement loss. We study the effect of the probe disentanglement loss weight λ in Figure 3 by varying its contribution relative to the autoregressive objective. Overall, MoP is reasonably stable across a broad range of λ, indicating that performance does not depend on highly

Table 4: **Controlled comparisons on Ego-in-Exo Perception.** These baselines match or exceed MoP's parameters, compute, or training samples without its probe structure. Params column shows trainable parameters (total / active per sample).

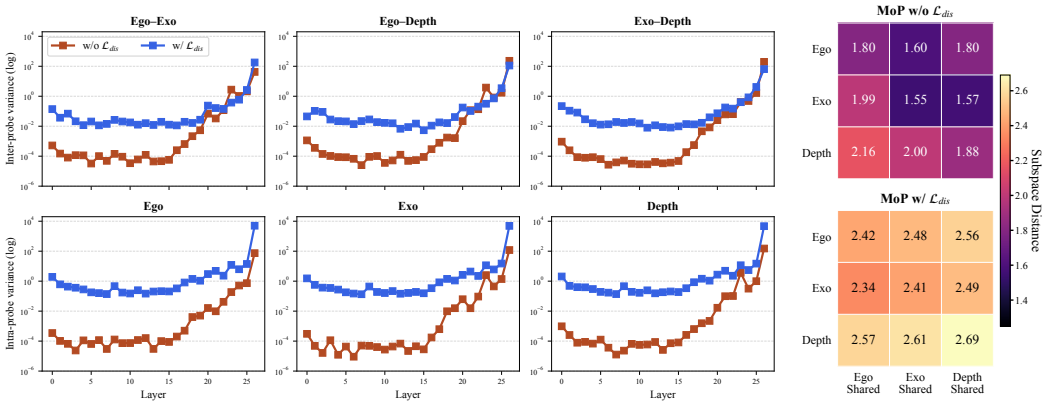
Method	Training Budget		Inference Modality		
	Params (B)	TFLOPs ($\times 10^6$)	Egocentric	Exocentric	Depth
Naive MLLM [47]	0.178 / 0.178	71.3	72.96	67.04	54.37
Capacity-matched					
Expanded Adapter	0.487 / 0.487	73.4	72.78	67.41	59.10
Q-Former (final layer)	0.767 / 0.767	75.5	73.10	66.77	55.08
Q-Former + Learnable Tokens	0.767 / 0.767	75.8	72.42	66.09	53.52
Cross-modal distillation and data-matched					
Cross-modal Distillation	0.178 / 0.178	73.6	72.68	67.05	54.44
Unimodal MLLM (3× data)	0.178 / 0.178	71.3	74.52	67.70	56.42
MoP (Ours)	0.769 / 0.482	75.3	**79.46**	**74.18**	**66.21**

sensitive tuning of the loss term. We observe that $\lambda = 100$ yields the strongest overall performance across all three inference modalities, and we therefore use this value in the remaining experiments.

Location of probing layers. In Figure 4, we ablate where probing is applied within the universal modality encoder. Specifically, we compare probing at every layer, every other layer, only early layers (1-5), only middle layers (5-15), and only late layers (15-27). Probing at every layer performs best overall, suggesting that MoP benefits from interacting with the encoder throughout its entire depth rather than only at a restricted subset of layers.

Placement of probe disentanglement loss. In Figure 5, we study where the probe disentanglement loss is applied within the universal modality encoder. We compare applying the loss at every layer against restricting it to early, middle, or late layers only. Applying the probe disentanglement loss at every layer yields the strongest overall results, indicating that encouraging probe specialization throughout the full encoder is more effective than constraining it only locally.

Controlled comparisons. In Table 4, we test whether MoP's added capacity (+3.7% active parameters per sample and +5.65% training compute) explains its gains by comparing against baselines that match or exceed its budget without the probing structure: an expanded projector, a Q-Former [19] on the final encoder layer, and a Q-Former fed by learnable tokens [9] traversing every encoder layer [17]. We further test explicit cross-modal distillation [13] and unimodal training on data duplicated to match MoP's sample count. Every alternative trails MoP by at least 4.9 points on each inference modality, showing that the gains stem from the probe structure rather than from added parameters, compute, tokens, or data.



(a) **Probe variance across universal modality encoder depth.** (b) **Subspace distances.**

Figure 6: **Effect of probe disentanglement loss on learned probe representations.**

5.4 Model Diagnosis & Analysis of MoP

Effect of probe disentanglement loss on learned probes. Figure 6a analyzes the variance of the probe representations with and without probe disentanglement loss. The top row reports *inter-probe variance* (across modality pairs), while the bottom row reports *intra-probe variance* (within each

modality). Without probe disentanglement loss, probe variance remains consistently low, indicating that probes collapse toward similar representations, particularly in early and mid-level layers. With the probe disentanglement loss, both inter- and intra-probe variance increase across layers and modalities, suggesting that probes learn more diverse and specialized representations.

To further understand this behavior, we analyze the geometry of the learned representations through pairwise subspace distances in Figure 6b. At a high level, we ask whether modality-specific and shared probes occupy distinct directions in representation space. We measure this via the Frobenius distance between the SVD-extracted projection matrices of each probe group, where larger values indicate less directional overlap. Without the disentanglement loss, these distances tend to be smaller, suggesting the two probe groups orient similarly. With it, distances are consistently larger, consistent with specific probes being pushed toward more orthogonal structure relative to the shared subspace.

MoP data scaling. Figure 7 studies how performance changes as the amount of training data increases from 25% to 100%. At each scale, both models are trained on the same subset, where each sample contains a single modality and its instruction, so the number of samples, tokens, and underlying events is matched. The naive MLLM baseline remains largely flat across all three inference modalities, indicating limited benefit from additional training data. In contrast, MoP improves consistently with scale, rising from 73.19 to 79.46 on egocentric inference, from 66.40 to 74.18 on exocentric inference, and from 55.48 to 66.21 on depth inference. These results suggest that MoP is able to exploit additional supervision more effectively than naive multimodal instruction tuning.

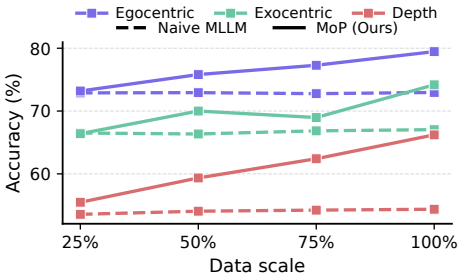


Figure 7: **Data scaling property of MoP.**

Table 5: **Computational overhead of MoP.**

Model	Training		Inference	
	Total Params (All / Active)	TFLOPs ($\times 10^6$)	Max VRAM (GB)	Latency (s)
Naive MLLM	8.21B / 8.21B	71.3	24.4	0.767
MoP (Ours)	+7.2% / +3.7%	75.3 (+5.65%)	32.5	1.078

Computational cost. In Table 5, we report MoP's training and inference overhead relative to the Naive MLLM. Training compute increases by 5.65%, and the layer-wise probe cross-attention raises inference latency from 0.767s to 1.078s per sample and peak VRAM from 24.4GB to 32.5GB. Because routing activates a single modality-specific branch per sample, MoP adds 7.2% total parameters but only 3.7% active parameters. MoP requires only one modality at inference, so its cost scales with the inference modality rather than the number of training modalities.

6 Conclusion and Limitations

We introduce **Mixture of Probes (MoP)** for multimodal large language models under the *privileged modality setting*. MoP a probing-based framework that improves cross-modal knowledge transfer by separating modality-specific information from modality-general structure. Our results demonstrate that this separation is the key to making auxiliary modalities useful: when properly leveraged, they can substantially improve single-modality performance. Despite these promising results, our study has two natural limitations. First, MoP assumes a universal modality encoder and does not directly extend to heterogeneous modality-specific encoders; this design is intentional, as the probing mechanism relies on a common representation space to extract and compare signals across modalities at intermediate layers. Second, we focus on the strict privileged modality setting with a single modality at inference and do not address multi-modal inference scenarios. This choice isolates the core challenge of transferring knowledge from auxiliary modalities. We leave extensions to heterogeneous encoders and multiple test-time modalities as important directions for future work.

Acknowledgments and Disclosure of Funding

We would like to thank Toshimitsu Uesaka for carefully reading the manuscript and providing us with valuable feedback. This work was supported in part by the National Science Foundation (IIS-2245652) and the NSF National AI Research Resource Pilot (NAIRR240338).

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 3

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[5] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 3

[6] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[8] Shilpa Dang, Jessica Emily Antono, Igor Kagan, and Arezoo Pooresmaeili. Modality-specific and modality-general representations of subjective value in frontal cortex. *Communications Biology*, 7:1550, 2024. 2

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. 9

[10] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 833–842, 2019. 1

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz,

Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, June 2022. 1

[12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 9

[14] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 7, 8

[15] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 1, 3

[16] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 43(5):1318–1334, 2013. 1

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, 2022. 3, 9

[18] Guangyao li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 3, 9

[20] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[22] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, 2016. 3

[23] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 3

[24] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 3

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 3

[26] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, and Anuj Kumar. Anymal: An efficient and scalable any-modality augmented language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 2, 3

[27] Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016. 3

[28] OpenAI. Gpt-4o system card, 2024. 1

[29] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[30] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 6, 7, 8

[31] Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[32] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. 7

[33] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[34] Dominick Reilly, Rajatsubhra Chakraborty, Arkaprava Sinha, Manish Kumar Govind, Pu Wang, Francois Bremond, Le Xue, and Srijan Das. Llavidal: A large language-vision model for daily activities of living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 6

[35] Dominick Reilly, Manish Kumar Govind, Le Xue, and Srijan Das. From my view to yours: Learning egocentric cues from exocentric video using privileged egocentric supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. 7

[36] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[37] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 3

[38] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. 1

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[40] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research (JMLR)*, 2015. 3

[41] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research (TMLR)*, 2026. 3

[42] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 1

[43] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 4, 5

[44] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024. 7

[45] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[46] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2023. 7

[47] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 7, 8, 9

[48] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 1, 3

[49] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 2

[50] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 3

[51] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 2, 3, 6

A Broader impacts

The broader impacts of MoP stem from its ability to improve single-modality inference by leveraging privileged modalities available only during training, which can benefit real-world applications where high-fidelity sensors are useful for supervision but impractical, expensive, invasive, or unavailable at deployment, such as assistive robotics, healthcare monitoring, and content understanding. At the same time, because MoP is built on top of pretrained language models and multimodal encoders, it may inherit or amplify biases, hallucinations, privacy risks, and safety limitations associated with these foundation models. Moreover, since MoP transfers information from auxiliary modalities, care must be taken when those modalities contain sensitive information, such as egocentric video, audio, depth, or other sensor streams that may reveal personal behavior or private environments.

B Subspace distances (additional details)

To further understand how MoP affects the representation spaces of each modality, we analyze the geometry of the learned probe representations through pairwise subspace distances in Figure 6b on a subset of 500 samples of the Ego-in-Exo Perception benchmark. For a given probe group, we collect its learned probe representations into

$$P \in \mathbb{R}^{N \times K \times d_v},$$

where N is the number of samples, K is the number of probes in the group, and d_v is the probe feature dimension. For modality-general probes, $K = K_g$, while for modality-specific probes, $K = K_s$. We then flatten this probe representation to obtain $X \in \mathbb{R}^{(NK) \times d_v}$, so that each row corresponds to one probe representation. We center X by subtracting its feature-wise mean, and then compute its top-k right singular vectors using singular value decomposition:

$$X = U \Sigma V^\top.$$

The first k right singular vectors define an orthonormal basis $B \in \mathbb{R}^{d_v \times k}$ for the dominant representation subspace of that probe group.

Given two probe groups with subspace bases B_1 and B_2, we compare their directions using the Frobenius distance between their projection matrices:

$$d(B_1, B_2) = \left\| B_1 B_1^\top - B_2 B_2^\top \right\|_F.$$

Smaller values indicate that two probe groups span similar directions in representation space, while larger values indicate less directional overlap. In our analysis, we use $k = 10$ and compute this distance pairwise across probe groups.

C Additional Ablations

Table 6: **Ablation on the number of probes.**

K_s	K_g	Inference Modality		
		Egocentric	**Exocentric**	**Depth**
4	4	77.27	71.03	65.27
8	4	78.74	71.52	64.77
4	8	78.48	73.56	64.00
8	8	**79.46**	**74.18**	**66.21**
16	16	76.70	70.95	62.11

Table 7: **Ablation on privileged modality combinations.** – denotes a modality absent from training.

Training Modalities	Inference Modality		
	Egocentric	**Exocentric**	**Depth**
Ego + Exo	76.13	71.45	–
Ego + Depth	76.71	–	62.65
Exo + Depth	–	70.15	60.93
Ego + Exo + Depth	**79.46**	**74.18**	**66.21**

Number of probes. Table 6 varies the number of modality-specific (K_s) and modality-general (K_g) probes. The 8/8 configuration performs best on all three inference modalities, and performance stays within about four points across configurations, so MoP does not depend on precise tuning of the probe count. Doubling to 16/16 yields the weakest results, further indicating that MoP's gains do not come from added probe capacity.

Privileged modality combinations. Table 7 trains MoP on each pair of modalities, holding the architecture, recipe, and total number of training samples fixed. Each model is evaluated only on the modalities it was trained on, since a modality absent from training has no probes to route to. The full ego+exo+depth set outperforms every pairwise combination on each evaluated inference modality; under exocentric inference, for example, adding egocentric video to exo+depth training

raises accuracy from 70.15 to 74.18. MoP's gains therefore accumulate across privileged modalities rather than being driven by any single one.

D t-SNE analysis on learned probes

Figure 8 visualizes the learned modality-specific and modality-general probe representations using t-SNE. We observe that the modality-specific probes form clearly separated clusters for egocentric, exocentric, and depth inputs, suggesting that they capture distinct modality-dependent structure. The modality-general probes form a cluster of their own, but we find that they additionally consist of smaller modality-conditioned sub-clusters. This behavior is consistent with the intended role of the two probe families: specific probes specialize to each modality, while general probes learn a shared representation space that is distinct from the specific probes. Importantly, the general probes do not collapse into identical representations across modalities, nor do they separate as strongly as the modality-specific probes, suggesting that they preserve transferable structure while retaining enough modality context to support cross-modal knowledge transfer.

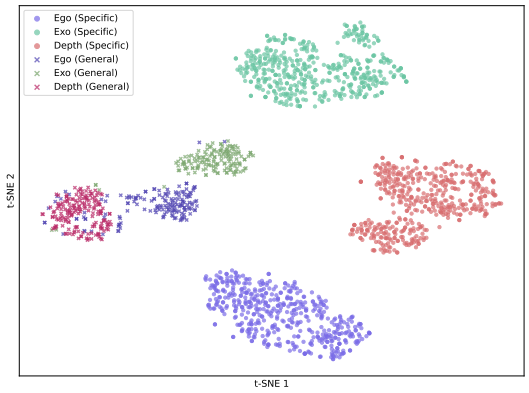


Figure 8: **t-SNE visualization of specific and general probes.**

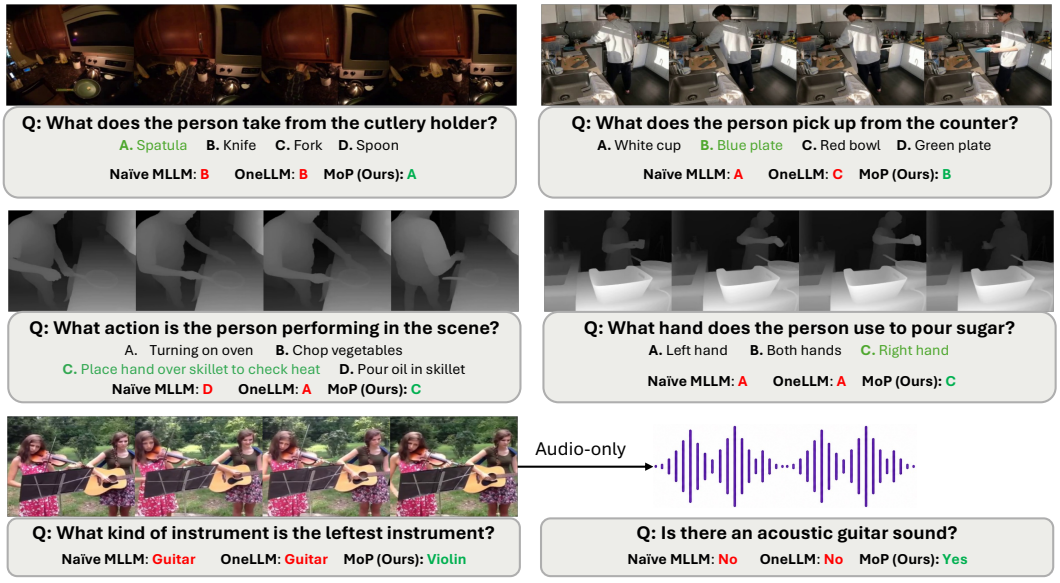


Figure 9: **Qualitative results of MoP on single modality inference.**

E Qualitative Results

Figure 9 presents qualitative comparisons between MoP and baseline MLLMs under single-modality inference. MoP produces more accurate responses, demonstrating a stronger ability to leverage the privileged modalities available during training. These examples qualitatively support the improvements observed in our quantitative evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract discusses the privileged modality setting and our approach using learnable probing, our methods and experiments reflect this setting and method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss two limitations of MoP in the Conclusion and Future Work section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We do not provide theoretical assumptions of proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The datasets we use are fully public and our methodology and implementation details are discussed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets we use are publicly available and our code will be released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimentatl details for our method are discuessed, dataset specific details are provided in the corresponding dataset papers.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Unfortunately due to computational intensity of MLLM training and a lack of resources, we are unable to provide error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details of the computational resources used by our model in the experimental section, as well as a runtime analysis in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We do not have any ethical concerns in our paper, all datasets used are public.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the introduction we discuss applications and scenarios where MoP can be used, we provide a dedicated section for broader impacts in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is not risk for our work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Datasets and code we use are freely available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No IRB necessary.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We use LLMs for automated review and refinement for the writing (as disclosed in the submission form), but we do not use it to generate a significant portion of our idea.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.