VISCOP: Visual Probing for Video Domain Adaptation of Vision Language Models

Dominick Reilly[1] **Manish Kumar Govind**[1] **Le Xue**[2] **Srijan Das**[1]
[1] University of North Carolina at Charlotte [2] Salesforce AI Research

Abstract

*Large Vision Language Models (VLMs) excel at general visual reasoning tasks, but their performance degrades sharply when deployed in novel domains with substantial distribution shifts compared to what was seen during pretraining. Existing approaches to adapt VLMs to novel target domains rely on finetuning standard VLM components. Depending on which components are finetuned, these approaches either limit the VLMs ability to learn domain-specific features, or lead to catastrophic forgetting of pre-existing capabilities. To address this, we introduce **Vis**ion **Co**ntextualized **P**robing (**VISCOP**), which augments the VLM's vision encoder with a compact set of learnable visual probes, enabling domain-specific features to be learned with only minimal updates to the pretrained VLM components. We evaluate* VISCOP *across three challenging domain adaptation scenarios: cross-view (exocentric → egocentric), cross-modal (RGB → depth), and cross-task (human understanding → robot control). Our experiments demonstrate that* VISCOP *consistently outperforms existing domain adaptation strategies, achieving superior performance on the target domain, while better retaining capabilities from the source domain. Code, models, and evaluation protocols are released at* https://github.com/dominickrei/VisCoP.

1. Introduction

Large Vision-Language Models (VLMs) [3, 32, 44, 51] have demonstrated strong performance across diverse multimodal tasks, including open-ended video question answering [26, 49] and spatial reasoning [19, 35]. These models couple pretrained vision encoders [34, 50] with Large Language Models (LLMs) [30, 33] and are trained on large-scale, web-curated image-video-text corpora containing broad but generic visual concepts [6, 28, 37, 52]. However, when deployed in domains that exhibit a pronounced *visual gap* arising from differences in viewpoint, sensing modality, or task structure, their performance degrades substantially under distribution shift. This challenge is prevalent in video domain adaptation settings such as ex-

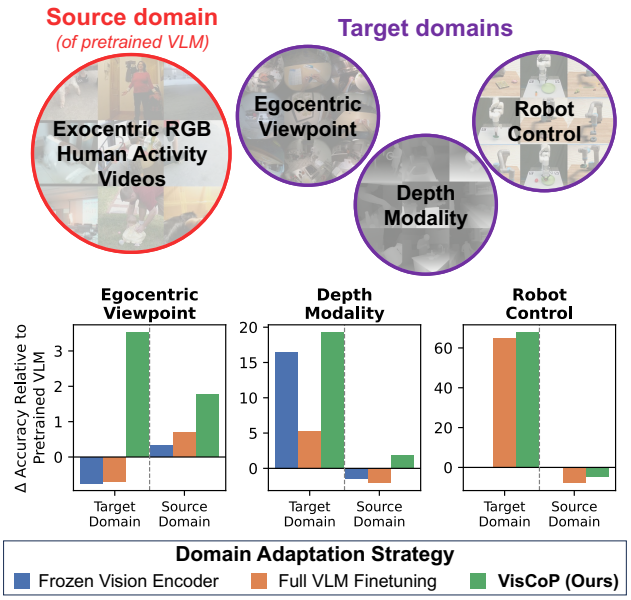


Figure 1. **Domain adaptation performance of different adaptation strategies.** VISCOP achieves superior target domain performance while better retaining source domain knowledge compared to other strategies.

ocentric → egocentric viewpoint, RGB → depth modality, and visual understanding → robotic control, where learning domain-specific *visual* representations is essential. Crucially, VLMs must *adapt while preserving* the broad world knowledge acquired during pretraining: a model adapted to egocentric data, for example, should retain competence on exocentric benchmarks to maintain cross-domain generalization.

A common strategy to address such distributional shift is partial fine-tuning of a pretrained VLM on domain-specific video instruction data. Freezing the vision encoder and updating only lightweight components (e.g., the vision–language connector) preserves pretrained knowledge but restricts visual specialization. In contrast, finetuning the vision encoder enables specialized visual understanding, albeit at the cost of catastrophic forgetting of pretrained knowledge [22, 46, 48]. This trade-off is particularly se-

vere when the dominant shift is visual. Moreover, existing approaches rely primarily on the final-layer representations of a frozen vision encoder, discarding intermediate features that encode multi-level visual structure [20]. We argue that such abstracted representations are insufficient for visually prominent domain shifts and pose the following question: *how can a VLM leverage domain-specific visual signals across the depth of a pretrained vision encoder while avoiding catastrophic forgetting?*

To this end, we introduce **Vis**ion **Co**ntextualized **P**robing (**VISCOP**), a lightweight adaptation mechanism that enables pretrained VLMs to specialize to a target domain while preserving their general-purpose visual knowledge. Motivated by the progressive emergence of semantic representations across transformer depths [4, 27, 42], VISCOP performs two way probing: it first extracts domain relevant signals from intermediate layers of a frozen vision encoder, and subsequently conditions the language model embedding space to enhance domain specific visual reasoning. This is achieved through a compact set of learnable probe tokens that interact layer wise with intermediate visual representations, forming an alternative parameterized pathway for adaptation. By leveraging multi-level abstractions across the encoder depth, VISCOP learns a parallel representation pathway that extracts domain-specific signals from intermediate encoder features while preserving pretrained representations. This enables the model to learn domain-specific visual patterns while mitigating catastrophic forgetting. In contrast, existing approaches [2, 14, 17, 21, 40] typically operate on the final-layer representations of the vision encoder or introduce prompt tokens that condition downstream reasoning without explicitly learning new visual cues *within* the vision encoder. This design provides a simple yet effective approach to video domain adaptation under prominent visual gaps without modifying the pretrained visual encoder parameters, which is not enabled by existing approaches. Empirically, we find that the representations learned via the VISCOP adaptation pathway enable effective cross-view, cross-modal, and cross-task adaptation of VLMs, while retaining their broad capabilities learned during pretraining. Metaphorically, the name VISCOP reflects its role as a "*traffic cop*", directing gradient flows away from the visual encoder and towards an alternative pathway for learning domain-specific visual features, avoiding the "*crash*" (catastrophic forgetting) that would otherwise occur if gradients flowed directly through the vision encoder.

To summarize, our key contributions:

1. We propose VISCOP (**Vis**ion **Co**ntextualized **P**robing), a novel video domain adaptation strategy for VLMs that learns domain-specific visual representations through layer-wise probing of a frozen vision encoder, enabling effective domain transfer and preventing catastrophic forgetting of multi-modal capabilities acquired during pretraining.
2. We establish a comprehensive evaluation setting for domain adaptation in VLMs, spanning three challenging target domains: cross-view (exocentric $\rightarrow$ egocentric), cross-modality (RGB $\rightarrow$ depth), and cross-task (action understanding $\rightarrow$ robotic control), along with standardized metrics to evaluate performance.
3. Our experiments demonstrate that VLMs trained with VISCOP outperform alternative domain adaptation strategies across diverse target domains, while retaining more knowledge of the source domain illustrated in Figure 1.

2. Related Works

Domain adaptation in vision-language encoders. Domain adaptation of contrastively trained vision-language encoders, such as CLIP [34, 36], is typically achieved through prompt tuning or adapter-based approaches. Both strategies aim to learn domain-specific features while keeping the pretrained vision and text encoders frozen. To accomplish this, prompt tuning approaches [47, 53, 54] introduce learnable prompt vectors as additional input to the text encoder, steering the model toward target domain. Adapter-based approaches [11, 46] insert lightweight trainable modules directly into the encoder space, thus updating their pretrained representations. In contrast to these approaches, VISCOP addresses the setting of domain adaptation in generative VLMs, enabling them to learn domain-specific features without requiring updates to the pretrained encoder representations.

Domain adaptation in VLMs. Domain adaptation in VLMs has largely been achieved through data-centric strategies rather than through architectural changes [7]. Existing approaches typically leverage automated pipelines [31, 38] or closed-source VLMs [5, 20] to curate visual-instruction pairs from existing datasets in the target domain. Their adaptation strategy usually follows a multi-stage training scheme similar to LLaVA [25], where different VLM components are selectively trained at each stage. However, the choice of trainable components creates a trade-off between extracting domain-specific features and retaining pretrained knowledge. Training only lightweight connectors retains pretrained knowledge but limits domain-specific visual understanding, while training the vision encoder enables specialized visual understanding at the cost of catastrophic forgetting. VISCOP avoids this trade-off through the introduction of visual probes that extract domain-specific features from a frozen vision encoder, enabling adaptation without disrupting the pretrained visual representations.

Visual probing vs. visual compression and prompt tuning. Several approaches employ learnable tokens to

bridge vision and language modalities [14, 40, 56] through architectures leveraging the Q-Former, Perceiver Resampler, or prompt tuning mechanisms. Q-Former [21] leverages learnable queries that cross-attend to representations from the final layer of the vision encoder, aggregating visual information into a reduced set of tokens for computational efficiency. Perceiver Resampler [2] operates similarly, aiming to compress the visual representations into a fixed number of learnable tokens. Prompt tuning methods [17, 24, 53] introduce learnable tokens to steer downstream reasoning, but operate solely on fixed encoder representations without enabling learning of new visual features. The visual probes proposed in VISCOP differ fundamentally, as they are designed to *extract and learn* domain-specific visual representations rather than to simply *compress* or *condition on* pretrained ones. This is enabled by their interaction with intermediate representations of the vision encoder, allowing the probes to extract domain-specific representations that are not propagated to the final representation of the pretrained vision encoder [34, 50].

3. Problem Formulation

Let $\mathcal{S}$ denote the *source domain*, on which the vision-language model f_{θ^0} has been pretrained, and let $\mathcal{T}$ denote the *target domain*, the domain of interest for adaptation. The two domains differ in their underlying distributions (e.g., viewpoint, modality, or task), which causes f_{θ^0} to perform poorly when directly applied to $\mathcal{T}$.

Training supervision in these domains is provided as video-QA pairs (v, q, a), where v is a video, q is an instruction or question, and a is the corresponding response. While f_{θ^0} has been pretrained on samples $(v, q, a) \sim \mathcal{S}$, at adaptation time we only assume availability of target domain samples $(v, q, a) \sim \mathcal{T}$. The objective of domain adaptation is to update the pretrained parameters θ^0 to obtain $\theta^\star$ that improves performance on domain $\mathcal{T}$, while retaining performance on domain $\mathcal{S}$. Formally,

$$R_{\mathcal{T}}(\theta^\star) < R_{\mathcal{T}}(\theta^0) \quad \text{and} \quad R_{\mathcal{S}}(\theta^\star) \approx R_{\mathcal{S}}(\theta^0)$$

where $R_{\mathcal{D}}$ denotes the VLM's expected autoregressive next-token prediction loss under domain $\mathcal{D}$. In summary, our problem statement considers adaptation of a pretrained VLM to a novel domain using only video-QA pairs from that domain. The objective is to improve target-domain performance while minimizing catastrophic forgetting of source-domain capabilities. In the next section, we introduce our proposed method, which enables balanced domain adaptation under these constraints.

4. Method: Video Domain-adaptive VLM

Given a video input $\mathbf{V} = \{\boldsymbol{I}_t\}_{t=1}^{T}$ consisting of T frames, the goal of the VLM is to generate the response corresponding to the input instruction in an autoregressive manner.

4.1. Preliminary

Existing VLMs for video representation learning [3, 51] consist of three standard components: **(i)** a vision encoder that maps visual inputs into a sequence of spatio-temporal tokens, **(ii)** a vision-language connector that projects the visual tokens to the embedding space of a language model, and **(iii)** an LLM that processes the projected visual tokens jointly with language tokens to enable multi-modal reasoning. For the input video $\mathbf{V}$, each frame $\boldsymbol{I}_t$ is processed independently by the vision encoder through a stack of L transformer layers. The visual tokens after the ℓ-th layer are denoted as

$$\mathbf{X}_t^\ell \in \mathbb{R}^{N \times d_v}, \quad \ell = 1, \ldots, L$$

where N is the number of spatial patch tokens per frame and d_v is the embedding dimension of the vision encoder. Concatenating these tokens over time yields $\mathbf{X}^\ell \in \mathbb{R}^{(TN) \times d_v}$ which represents the sequence of spatio-temporal visual tokens at the ℓ-th layer of the vision encoder. The final layer outputs $\mathbf{X}^L$ are then projected to the language embedding space via a vision-language connector $\mathcal{C}$ to obtain the visual embeddings used as input to the LLM

$$\mathbf{E} = \mathcal{C}(\mathbf{X}^L) \in \mathbb{R}^{(T\tilde{N}) \times d_{\text{lm}}}$$

where $\tilde{N}$ is the number of visual tokens input to the LLM after spatial downsampling [51]. and d_{lm} is the embedding dimension of the LLM.

The VLM is then trained to optimize a standard autoregressive next token prediction loss. Specifically, given the visual embeddings $\mathbf{E}$ and the tokenized QA pair $(\mathbf{Q}, \mathbf{A})$, we optimize the likelihood of predicting $\mathbf{A}$ conditioned on the visual embeddings and the question

$$P(\mathbf{A} \mid \mathbf{E}, \mathbf{Q}) = \prod_{j=1}^{\text{Len}} P_{\boldsymbol{\theta}}(\mathbf{a}_j \mid \mathbf{E}, \mathbf{Q}, \mathbf{A}_{<j})$$

where $\boldsymbol{\theta}$ are the trainable parameters of the VLM, Len indicates the token length of $\mathbf{A}$, and $\mathbf{A}_{<j}$ represents the subsequence of answer tokens preceding position j.

For domain-adaptive post training of VLMs, finetuning the vision encoder of a pretrained VLM for a target domain $\mathcal{T}$ often leads to overfitting on $\mathcal{T}$ and catastrophic forgetting of the source domain [22, 46, 48]. To mitigate this trade-off, a domain-adaptive pathway is required that adapts the VLM to $\mathcal{T}$ while retaining performance on $\mathcal{S}$.

4.2. VISCOP: Vision Contextualized Probing

To capture the relevant visual context that would otherwise be lost by freezing the vision encoder, we propose **Vis**ion **Co**ntextualized **P**robing (**VISCOP**), a mechanism that augments the vision encoder with a compact set of learnable tokens, called *visual probes*, and an interaction module that acts as a semantic interface between the probes and intermediate visual representations, as illustrated in Figure 2. Unlike prior token-based approaches that reinterpret fixed

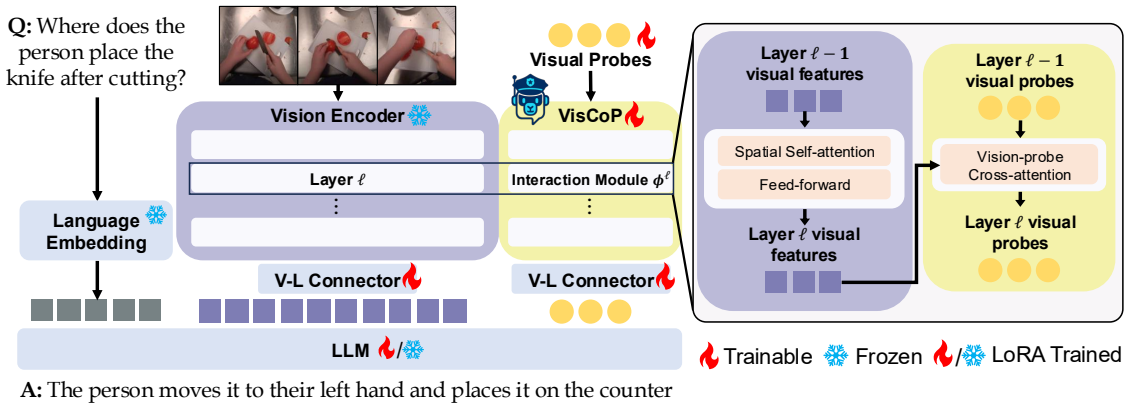


Figure 2. **Architecture of our proposed VisCoP.** Learnable visual probes are conditioned on intermediate representations of a frozen vision encoder through vision-probe cross-attention, which extracts domain-specific features that may have otherwise been discarded by the frozen encoder.

encoder features [2, 17, 21], VisCoP introduces a dedicated adaptation pathway that learns new visual representations from intermediate encoder layers while preserving the pretrained encoder parameters. In this section, we introduce how domain-adaptive VLMs are trained with VisCoP.

VisCoP augments the frozen vision encoder of a VLM with a compact set of M learnable *visual probes* $\mathbf{P} \in \mathbb{R}^{M \times d_v}$. The probes are trained to extract domain-specific spatio-temporal cues from intermediate representations of the vision encoder. To enable this extraction, a learnable *interaction module* Φ^ℓ inserted at each layer of the vision encoder conditions the probes on the hierarchical representations of the vision encoder at layer ℓ:

$$\mathbf{P}^{\ell+1} = \Phi^\ell(\mathbf{P}^\ell, \mathbf{X}^\ell).$$

Concretely, Φ^ℓ is implemented as a vision-probe cross-attention between the visual embeddings and the probes at layer ℓ. Let $(\boldsymbol{W_q}, \boldsymbol{W_k}, \boldsymbol{W_v})$ be the projection matrices in Φ^ℓ, then the probe update is

$$\mathbf{P}^\ell = \text{softmax}\left(\frac{\mathbf{P}^\ell \boldsymbol{W_q}^\ell (\mathbf{X}^\ell \boldsymbol{W_k}^\ell)^\top}{\sqrt{d_v}}\right)(\mathbf{X}^\ell \boldsymbol{W_v}^\ell),$$

Each Φ^ℓ is parameterized independently, enabling layer-specific aggregation of low- to high-level visual semantics. In contrast to the vision encoder of existing VLMs, which only learns spatial relationships through intra-frame self-attention, the visual probes attend to *all* frames in the video, enabling them to learn complex spatio-temporal relationships specific to video understanding. In some settings, such as robotic control, vision-probe cross-attentions are restricted to spatial tokens only.

After the final layer, the updated probes $\mathbf{P}^L$ are projected to the language space via a dedicated connector $\mathcal{C}_{\text{probe}}$,

$$\mathbf{Z} = \mathcal{C}_{\text{probe}}(\mathbf{P}^L) \in \mathbb{R}^{M \times d_{\text{lm}}},$$

and the VLM is trained with the standard autoregressive objective additionally conditioned on $\mathbf{Z}$:

$$P(\mathbf{A} \mid \mathbf{E}, \mathbf{Q}, \mathbf{Z}) = \prod_{j=1}^{\text{Len}} P_{\boldsymbol{\theta}}(\mathbf{a}_j \mid \mathbf{E}, \mathbf{Q}, \mathbf{Z}, \mathbf{A}_{<j}).$$

Thus, the probes act as low-dimensional control knobs that bias learning toward domain-relevant structure and away from spurious artifacts. This is reinforced by applying updates through the probe connector, and through LoRA [15] updates in the LLM embedding space, which confine parameter changes to a low-rank, probe-defined visual subspace that preserves generalizable behavior while enabling targeted specialization.

5. Experiments

We evaluate VisCoP for effective domain adaptation and minimal forgetting. Section 5.1 details the setup (architecture, training, metrics); Section 5.2 reports results on egocentric, depth, and robotic-control targets; Section 5.3 presents ablations and representation analyses of the probes and interaction modules.

5.1. Experimental Setting

VLM Architecture. We consider a VLM architecture consisting of a SigLIP [50] vision encoder, Qwen 2.5 [33] LLM, and a 2-layer MLP vision-language connector, with all modules initialized from the pretrained weights of VideoLLaMA3 [51]. The embedding dimension of the vision encoder is $d_v = 1152$, and the embedding dimension of the LLM is $d_{\text{lm}} = 3584$. We refer to this pretrained model as the *base* VLM, and to models adapted to a target domain as *expert* VLMs. To adapt the base VLM to a target domain, we perform finetuning on the target domain with a learning rate of 1×10^{-5} for the LLM and vision-language connector, and a learning rate of 2×10^{-6} for the vision encoder (when trainable). The model is finetuned on 4 NVIDIA H200 GPUs for 3 epochs when adapting to video domains,

Table 1. **Egocentric Video Understanding Experts.** Performance on the egocentric target domain and exocentric source domain. 🔥 denotes trainable and ❄ denotes frozen components 🔥/❄ denotes LoRA updates. Δ_{target} and Δ_{source} denote relative gains over the Base VLM.

Adaptation Strategy		Egocentric (Target)						Exocentric (Source)					Metrics	
VE	LLM	Action Und.	Task Regions	HOI	Hand Ident.	EgoSchema	Avg	NeXTQA	Video MME	ADL MCQ	ADL Desc	Avg	Δ_{target}	Δ_{source}
Base VLM		75.37	74.88	75.56	65.38	60.98	70.43	**84.32**	**65.37**	77.36	70.65	74.42	–	–
Finetuning Strategies (🔥 Vision Language Connector)														
❄	❄	73.00	76.71	72.85	65.51	60.43	69.70	84.21	62.67	76.56	75.51	74.74	-0.74	+0.31
🔥	❄	76.13	**82.93**	73.32	64.86	61.14	71.68	83.87	61.41	77.05	76.09	74.61	+1.24	+0.18
🔥	🔥	73.28	82.68	72.96	**65.77**	60.31	71.00	82.34	64.26	78.21	70.89	73.93	+0.57	-0.50
❄	❄/🔥	73.49	74.27	74.50	64.99	61.52	69.75	84.24	64.41	77.42	74.36	75.11	-0.68	+0.68
VISCOP		**81.28**	82.80	**78.75**	64.86	**62.11**	**73.96**	84.31	64.70	**78.97**	**76.78**	**76.19**	**+3.53**	**+1.77**

or 2 epochs when adapting to robotic control domains.

VISCOP Details. By default, VISCOP operates at every layer of the vision encoder and employs $M = 16$ visual probes unless otherwise stated. The visual probes are initialized from the normal distribution $\mathcal{N}(0, 0.02)$. Each interaction module Φ^{ℓ} is implemented as a multi-head cross-attention [42], and its weights are initialized from the self-attention weights of the vision encoder at layer ℓ. During domain adaptation, we freeze the vision encoder and update only the visual probes, interaction modules, vision–language connectors, and the LLM's LoRA parameters. For adaptation to video understanding domains, we update the LLM using LoRA ($r = 16$), while the entire LLM is updated when adapting to the robotic control domain.

Adaptation Metrics. We evaluate the domain adaptation of VLMs across two dimensions: **(i)** their "*improvement*" on the target domain $\mathcal{T}$, and **(ii)** their "*retention*" on the source domain $\mathcal{S}$. Improvement on the target domain is measured as the performance difference between the expert and base VLMs on target domain benchmarks; retention is the corresponding difference on source domain benchmarks. If $\text{Acc}_{\mathcal{D}}$ denotes the average accuracy over all benchmarks within the domain $\mathcal{D}$, then the metrics are computed by:

$$\Delta_{\text{target}} = \text{Acc}^{\text{expert}}_{\text{target}} - \text{Acc}^{\text{base}}_{\text{target}}$$
$$\Delta_{\text{source}} = \text{Acc}^{\text{expert}}_{\text{source}} - \text{Acc}^{\text{base}}_{\text{source}}$$

5.2. Source and Target Domains

The source domain $\mathcal{S}$ is fixed throughout this paper: exocentric RGB videos of human actions reflecting the samples used to train generic VLMs for video representation learning. Our target domains $\mathcal{T}$ deliberately shift the input distribution, and consist of (1) egocentric video understanding, (2) depth-modality video understanding, and (3) robotic control. All data (videos and instructions) in our chosen target domain benchmarks were not used in the pretraining of the base VLM [51]. We evaluate VISCOP's adaptation to each target while measuring retention of source domain competencies: (i) when adapting to egocentric video, exocentric understanding should be preserved; (ii) when adapting to depth video, RGB understanding should be preserved; and (iii) when adapting to robotic control, human-action understanding should be preserved.

Training datasets. For ego and depth video understanding domains, we adapt using EgoExo4D [13], a large-scale multi-view dataset containing time-synchronized egocentric and exocentric videos of skilled human activities. We utilize a total of 24,688 videos from the keystep recognition subset to generate 74,064 video instruction pairs. These instructions are recaptioned from the instruction pairs provided in [39]. For the **egocentric** target domain, we adapt on 45,888 egocentric video-instruction pairs. For the **depth** target domain, we convert all exocentric RGB videos to depth using DepthAnythingV2 [45] while keeping the language instructions unchanged, yielding 28,176 depth instruction pairs.

We perform adaptation to the **robotic control** domain in both simulated and real-world robot environments. In the *simulated environment*, we leverage the training set of VIMA-Bench [18]. VIMA-Bench contains 17 object manipulation tasks with an action space comprising two 2D coordinates (for pick and place positions) and two quaternions (for rotation). Since the training set of VIMA-Bench lacks natural language instructions by default, we leverage the instruction pairs generated in LLaRA [23], resulting in 13,922 instruction pairs across 7,995 action trajectories. In the *real-world environment*, we collect a dataset using a 6-DoF xArm 7 robot arm deployed in a tabletop manipulation setting. This dataset, which we refer to as xArm-Det, contains 1,007 instruction pairs depicting novel objects and spatial configurations not present in simulation. During adaptation, we train jointly on VIMA-Bench and xArm-Det, resulting in a total of 14,929 instruction pairs. The large-scale simulated data enables the model to learn manipulation skills, while xArm-Det exposes the model to the real environment.

5.2.1. Egocentric Video Understanding

Target and source benchmarks. For evaluation on the **target domain**, we evaluate on the Ego-in-Exo Perception [39] and EgoSchema [29] benchmarks. Ego-in-Exo Perception is derived from EgoExo4D and comprises 3,991 video question-answer (video-QA) pairs spanning four cat-

Table 2. **Depth Video Understanding Experts.** Performance on the depth target domain and RGB source domain. 🔥 denotes trainable components, ❄ denotes frozen components, and 🔥/❄ denotes LoRA updates. Δ_{target} and Δ_{source} denote relative gains over the Base VLM.

Adaptation Strategy		Depth (Target)					RGB (Source)						Metrics	
VE	LLM	Action Und.	Task Regions	HOI	Hand Ident.	Avg	Ego-in-Exo (Exo RGB)	NeXTQA	Video MME	ADL MCQ	ADL Desc	Avg	Δ_{target}	Δ_{source}
Base VLM		34.73	50.61	35.06	63.06	45.86	66.27	**84.32**	**65.37**	**77.36**	70.65	72.79	–	–
Finetuning Strategies (🔥 Vision Language Connector)														
❄	❄	55.67	66.59	62.46	64.49	62.30	71.36	83.15	62.41	70.90	69.05	71.37	+16.44	-1.42
🔥	❄	**57.20**	69.63	54.43	**64.48**	61.44	60.97	82.89	62.00	71.48	67.26	68.92	+15.57	-3.87
❄	🔥/❄	42.94	53.54	43.92	63.96	51.09	60.97	83.73	64.19	72.19	72.49	70.71	+5.23	-2.08
VISCOP		56.78	**73.17**	**66.23**	64.35	**65.13**	**71.89**	83.91	64.30	76.59	**76.47**	**74.63**	**+19.27**	**+1.84**

Table 3. **Robot Control Experts (Simulation).** Performance on the robotic control target domain and human understanding source domain. 🔥 denotes trainable components, ❄ denotes frozen components. Δ_{target} and Δ_{source} denote relative gains over the Base VLM.

Adaptation Strategy		Robotic Control (Target)				Human Understanding (Source)						Metrics	
VE	LLM	L1	L2	L3	Avg	Ego-in-Exo (Exo RGB)	NeXTQA	Video MME	ADL MCQ	ADL Desc	Avg	Δ_{target}	Δ_{source}
Base VLM		0	0	0	0	66.27	**84.32**	**65.37**	**77.36**	**70.65**	**72.79**	–	–
Finetuning Strategies (🔥 Vision Language Connector)													
🔥	🔥	**69.62**	60.77	65.00	65.13	56.92	83.24	62.74	52.21	64.50	63.92	+65.13	-8.87
🔥	❄	63.46	63.08	68.75	65.10	59.42	83.16	64.41	52.92	64.86	64.95	+65.10	-7.84
VISCOP		67.69	**65.77**	**70.00**	**67.82**	**71.19**	83.71	63.67	55.89	66.62	68.22	**+67.82**	**-4.58**

egories: action understanding (Action Und.), task-relevant region understanding (Task Regions), human-object interactions (HOI), and hand identification (Hand Ident.). Because it is derived from EgoExo4D, Ego-in-Exo Perception can be evaluated from either the egocentric or the exocentric viewpoint. For the ego target domain experiments, we report results using the egocentric videos, denoted as Ego-in-Exo Perception (Ego RGB). EgoSchema consists of 5,031 egocentric video-QA pairs derived from Ego4D [12].

For evaluation on the **source domain**, we select benchmarks that measure exocentric video understanding capability. Specifically, we evaluate on the NeXTQA [43], VideoMME [10], and ADL-X [38] benchmarks. NeXTQA and VideoMME are general-purpose video-QA benchmarks built from web-scraped videos (e.g., from YouTube), with 8,564 QA pairs in NeXTQA and 2,700 QA pairs in VideoMME. ADL-X is a video-QA benchmark built from videos of activities of daily living, it contains a total of 10,561 multiple-choice questions (ADL-X MCQ) and 1,862 video description questions (ADL-X Desc) derived from various activities of daily living datasets [8, 9, 16, 41].

Results. Table 1 reports results of adaptation to the egocentric viewpoint. Training only the vision-language connector or the connector together with LLM LoRA adapters does not lead to effective adaptation to the target domain ($\Delta_{\text{target}} < 1$). Updating all three modules (connector, vision encoder, and LLM) improves performance on the target domain by $\Delta_{\text{target}} = +0.57$, but the large number of trainable parameters results in forgetting on the source benchmarks ($\Delta_{\text{source}} = -0.50$). In contrast, updating the connector and vision encoder alone slightly improves performance on the target domain and does not lead to forgetting on the source domain. These results highlight that the core difficulty of domain adaptation in existing VLMs arises from the necessity of updating the vision encoder to learn domain-specific visual representations, which inevitably leads to forgetting of pretrained knowledge. Our proposed VISCOP achieves the strongest adaptation performance, with the highest improvement on the target domain ($\Delta_{\text{target}} = +3.5$) while simultaneously maintaining retention on the source benchmarks ($\Delta_{\text{source}} = +1.8$). Interestingly, VISCOP not only avoids catastrophic forgetting but also improves performance on some source benchmarks (e.g., ADL-X). We attribute this positive transfer to a multi-axis domain shift: although source and target differ in viewpoint (exocentric vs. egocentric), their action distributions overlap. ADL-X, while exocentric, encapsulates activities of daily living that closely aligns with the EgoExo4D action distribution, enabling cross-domain generalization.

5.2.2. Depth Video Understanding

Target and source benchmarks. For evaluation on the **target domain**, we evaluate on the Ego-in-Exo Perception [39] benchmark. In the depth-adaptation setting, we train on depth maps of exocentric videos extracted with DepthAnythingV2 [45] and evaluate on exocentric depth videos following [39], denoted Ego-in-Exo Perception (Exo Depth). For the **source domain**, we use *RGB* benchmarks of exocentric understanding: Ego-in-Exo Perception (Exo RGB), NeXTQA, VideoMME, and ADL-X.

Results. We present the results for adaptation to the depth modality in Table 2. In contrast to the results on egocentric viewpoint adaptation, we find that all training strategies achieve improvements on the target domain, reflecting the disparity of the visual embedding space between the depth and RGB modalities. We find that this disparity leads to different behavior across training strategies. Jointly updating the vision encoder and the vision-language connector preserves source performance for egocentric adaptation but causes severe catastrophic forgetting under depth adaptation ($\Delta_{\text{source}} = -3.87$). This arises from the substantial encoder updates required to bridge RGB and depth, which overwrite RGB representations. In contrast, VISCOP preserves RGB features and source performance while achieving the largest target domain gains ($\Delta_{\text{target}} = +19.27$).

5.2.3. Robot Control

Target and source benchmarks. For evaluation on the **target domain**, we consider both simulated and real-world robotic environments. In simulation, we use the evaluation set of VIMA-Bench [18], which organizes tasks into three levels of difficulty: L1 (Object Placement), where all objects have been seen during training; L2 (Novel Combination), where objects seen during training appear in new pairings or contexts; and L3 (Novel Objects), where objects entirely unseen during training are introduced. Together, these levels measure generalization from familiar training conditions to progressively more challenging distributions. In the real-world setting, we evaluate on three tabletop manipulation tasks: T1) Place the {object} on the plate, T2) Pick up and rotate {object} by {angle}; and T3) Move all {color} objects onto the plate. On these robotic control benchmarks, the reported accuracy corresponds to the success rate across all robot manipulation tasks. For **source domain** evaluation of VLMs trained on both real and simulated robotic environments, we use the human-activity video benchmarks Ego-in-Exo (Exo RGB), NeXT-QA, VideoMME, and ADL-X.

Table 4. **Robot Control Experts (Real-world).** Performance on the robotic control target domain and human understanding source domain. 🔥 denotes trainable and ❄ denotes frozen components.

Adaptation Strategy		Robotic Control (Target)				Metrics	
VE	LLM	T1	T2	T3	Avg	Δ_{target}	Δ_{source}
Training data: VIMA-Bench							
🔥	🔥	**45.00**	60.00	15.00	40.00	+40.00	-8.87
VISCOP		40.00	**70.00**	**20.00**	**43.33**	**+43.33**	**-4.58**
Training data: VIMA-Bench + xArm-Det							
🔥	🔥	85.00	85.00	70.00	80.00	+80.00	-11.04
VISCOP		**100.00**	**100.00**	**90.00**	**96.67**	**+96.67**	**-11.00**

Results. The results of adaptation to the robotic control domain are presented in Table 3. The base VLM demonstrates weak performance on all robot control tasks, as its pretraining distribution lacks action trajectories (i.e., instruction data mapping from visual observations to robot actions). This lack of pretraining results in 0% accuracy across all levels of VIMA-Bench, and is consistent with prior works [23]. This highlights the extreme domain gap both in the visual space (robot observations vs. human videos) and in the language space (control actions vs. linguistic outputs) between the source and target domains. Similarly to the depth adaptation setting, we find that training the vision encoder improves performance on the target domain, but results in the worst source domain retention ($\Delta_{\text{source}} = -8.87$) of all robot control experts. In contrast, our proposed VISCOP achieves the best performance on the target domain ($\Delta_{\text{target}} = +67.82$) while retaining the most source domain knowledge ($\Delta_{\text{source}} = -4.58$) compared to other experts, demonstrating the effectiveness of our method even when the gap between the source and target domains is very large. Also note that VISCOP operates on per-timestep images in these experiments; thus the visual probes consume the same visual tokens as the vision encoder, suggesting they extract domain-specific representations more effectively than the base encoder.

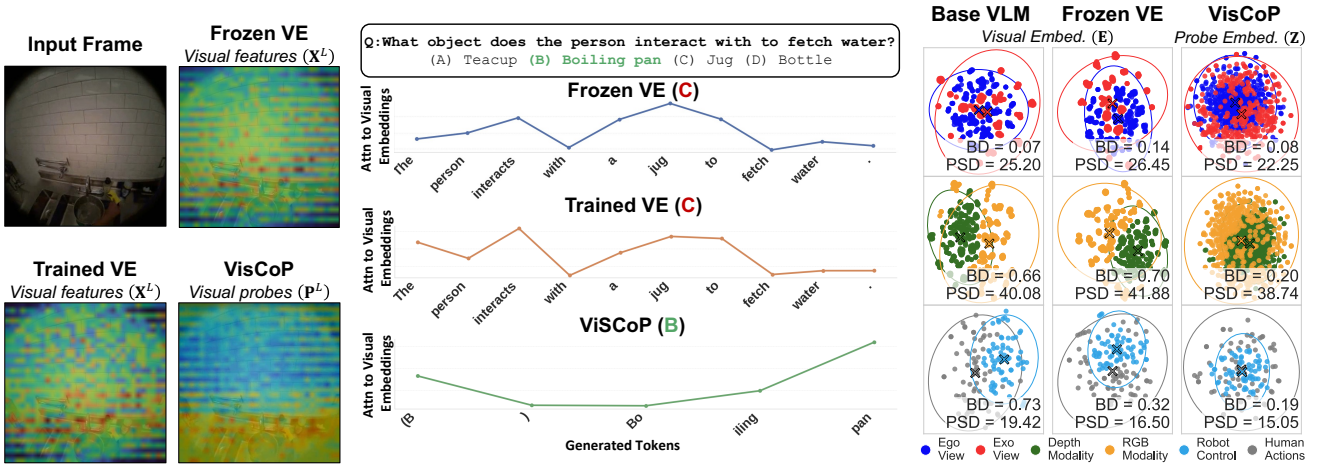
We further evaluate adaptation in the real-world setting using the xArm-Det dataset in Table 4. We consider a *transfer setting*, where the experts are trained only on VIMA-Bench and directly evaluated on xArm-Det, and the setting where the experts are jointly trained on both VIMA-Bench and xArm-Det. In both cases, our proposed VISCOP outperforms the vision encoder trained experts on target domain adaptation as well as source domain retention.

Table 5. **Alternative adaptation strategies.** Legend: *VP* (visual probing with no interaction modules), *Last-4* (train only the last 4 vision encoder layers), *QFormer Style* (interaction modules placed after the last VE layer).

Method	Target	Source	Adaptation Metrics	
	Avg	Avg	Δ_{target}	Δ_{source}
Base VLM	70.43	74.42	–	–
VP	65.57	75.05	-4.86	+0.62
LoRA (Full)	69.85	75.35	-0.59	+0.92
Last-4	70.46	72.62	+0.02	-1.80
VPT [17]	71.36	74.97	+0.93	+0.55
QFormer Style	70.99	75.03	+0.56	+0.61
Model Tailor [55]	70.27	75.29	-0.16	+0.86
VISCOP	**73.96**	**76.19**	**+3.53**	**+1.77**

5.3. Model Diagnosis and Analysis

Alternatives to learnable queries. Table 5 compares VISCOP against alternative adaptation strategies. *Visual Probes Only (VP)* trains only visual probes with their vision-language connector ($\mathcal{C}_{probe}$) without any interaction modules. *Partial Encoder Training (Last-4)* makes the final four layers of the vision encoder trainable. *QFormer-Style Compression* uses visual probes with interaction modules only at the vision encoder's final layer, mimicking Q-



(a) **Attention of visual features and probes.** (b) **Attentions of generated language tokens to visual embeddings.** (c) **t-SNE of source and target domain features.**

Figure 3. **Analysis of VISCOP. (a)** Attentions between visual features and visual probes. **(b)** Attention of generated language tokens to visual embeddings. **(c)** t-SNE of visual and probe embeddings. Ellipses denote 95% confidence regions of a fitted 2D Gaussian, and cross markers indicate the Gaussian means. Bhattacharyya distance (BD) and per-sample distance (PSD) are shown.

Former's compression approach [21]. *Model Tailor* [55] performs post-hoc domain adaptation by fusing parameter updates from a fine-tuned VLM back into the base VLM, modifying only the LLM parameters and leaving the vision encoder untouched. Training with QFormer-Style compression or only training with visual probes (*VP*) underperforms compared to VISCOP, indicating the importance of probe interactions at intermediate layers of the vision encoder to learn domain-specific features across multiple levels of abstraction. Similarly, training only the last four layers of the vision encoder, or training it with LoRA, also underperforms, highlighting that even partial parameter updates fail to capture domain-specific signals as effectively as VISCOP's visual probes. Model Tailor also falls short in this setting, suggesting that approaches which do not leverage intermediate vision encoder representations struggle to learn domain-specific visual features.

Visualizing attention in domain-adapted VLMs In Figure 3a, we analyze attention maps of various VLM adaptation strategies to assess how different components capture domain-specific visual features. For both the frozen and trainable vision encoders, we visualize attention using attention rollout [1], for VISCOP we visualize the attentions of the visual probes, averaged across all probes. The frozen vision encoder fails to focus consistently on relevant regions under the experimented domains, reflecting its limited ability to capture domain-specific features. The trained vision encoder yields sharper attention on the relevant regions, indicating its ability to learn domain-specific features, albeit at the cost of catastrophic forgetting of the source domain as shown in Section 5.2. In contrast, the visual probes of VISCOP have a sharp focus on the task-relevant regions, despite the vision encoder being frozen. This indicates that the probes alone are able to extract the domain-specific visual features necessary for adaptation. In Figure 3b, we visualize the attention of generated language tokens to visual embeddings. We find that VISCOP correctly responds to the query, with more focus given to tokens corresponding to relevant objects.

Learning domain-specific representations. Figure 3c compares t-SNE embeddings of source and target domains across different VLMs. Circles represent individual samples, and ellipses denote 95% confidence regions of fitted 2D Gaussians. For the egocentric and depth target domains, each source-target pair corresponds to time-synchronized videos of the same action. For the robot target domain, pairs correspond to pick-and-place actions performed by humans. Ideally, the embeddings of paired samples should lie closer together in the embedding space, reflecting alignment across the source and target domains. We quantify this using two metrics: the *Bhattacharyya distance (BD)* computed between the Gaussians fitted to each domain, and the *per-sample distance (PSD)*, defined as the mean Euclidean distance between paired embeddings across domains. We observe that the visual probes of VISCOP learn stronger alignment between the source and target domains.

6. Conclusion

We introduced VISCOP, a mechanism that extracts domain-specific visual features through probing of a frozen vision encoder to enable effective domain adaptation in VLMs and prevent catastrophic forgetting. VLMs equipped with VISCOP achieve superior target domain performance, while maintaining strong source domain capabilities across cross-view, cross-modal, and cross-task adaptation scenarios. We will release all code, models, and evaluation protocols to facilitate future research.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 2, 3, 4

[3] 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. 1, 3

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[5] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. 2

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, 2024. 1

[7] Daixuan Cheng, Shaohan Huang, Ziyu Zhu, Xintong Zhang, Wayne Xin Zhao, Zhongzhi Luan, Bo Dai, and Zhenliang Zhang. On domain-adaptive post-training for multimodal large language models. In *Conference on Empirical Methods in Natural Language Processing Findings*, 2025. 2

[8] 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. 6

[9] 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. 6

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal large language models in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 6

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. In *International Journal of Computer Vision*, 2023. 2

[12] 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 Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, 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 Puentes, 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 Arbelaez, 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*, pages 18995–19012, 2022. 6

[13] 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*, 2025. 5

[14] Cuong Nhat Ha, Shima Asaadi, Sanjeev Kumar Karn, Oladimeji Farri, Tobias Heimann, and Thomas Runkler. Fusion of domain-adapted vision and language models for medical visual question answering. In *Proceedings of the Clinical Natural Language Processing Workshop at the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. 2, 3

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

[16] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *Proceedings of the European Conference on Computer Vision*, pages 767–783, 2020. 6

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2, 3, 4, 7

[18] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. In *International Conference on Machine Learning*, 2023. 5, 7

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1

[20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, 2023. 2

[21] 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. 2, 3, 4, 8

[22] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Conference on Empirical Methods in Natural Language Processing*, 2024. 1, 3

[23] Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Sung Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, Mu Cai, Yong Jae Lee, and Michael S. Ryoo. Llara: Supercharging robot learning data for vision-language policy. In *International Conference on Learning Representations*, 2025. 5, 7

[24] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[26] Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024. 1

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021. 2

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding, 2024. 1

[29] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023. 5

[30] Meta. The llama 3 herd of models, 2024. 1

[31] Fnu Mohbat and Mohammed J. Zaki. Llava-chef: A multi-modal generative model for food recipes. In *ACM International Conference on Information and Knowledge Management*, 2024. 2

[32] OpenAI. Thinking with images, 2025. Accessed: May 27, 2026. 1

[33] 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, 2025. 1, 4

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3

[35] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S. Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12977–12987, 2024. 1

[36] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[37] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. 1

[38] 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. 2, 6

[39] Dominick Reilly, Manish Kumar Govind, Le Xue, and Srijan Das. From my view to yours: Ego-augmented learning in large vision language models for understanding exocentric daily living activities, 2025. 5, 6

[40] Michael S. Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Jongwoo Park, Kanchana Ranasinghe, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms, 2025. 2, 3

[41] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision*, pages 510–526, 2016. 6

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2, 5

[43] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 6

[44] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Shaoyen Tseng, Gustavo A Lujan-Moreno, Matthew L Olson, Musashi Hinck, David Cobbley, Vasudev Lal, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models, 2025. 1

[45] 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. 5, 6

[46] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video understanding. In *International Conference on Learning Representations*, 2023. 1, 2, 3

[47] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[48] Yuhang Zang, Hanlin Goh, Josh Susskind, and Chen Huang. Overcoming the pitfalls of vision-language model finetuning for ood generalization. In *International Conference on Learning Representations*, 2024. 1, 3

[49] Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[50] 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. 1, 3, 4

[51] 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. 1, 3, 4, 5

[52] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 1

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[54] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[55] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *International Conference on Machine Learning*, 2024. 7, 8

[56] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3