

Gold Points Sniper: Self-guided Visual Reasoning in VLM for Fine-grained Action Understanding

Haodi Liu¹, Xinhang Yang¹, Kunda Yan¹, Sen Cui¹, Zeyu Zhang^{2†}, Changshui Zhang^{1†}

Abstract—Robots operating in everyday environments must understand fine-grained human actions, intentions, and contextual cues from broad views where people occupy only small regions, a capability unmet by current systems. While open-vocabulary action recognition methods remain limited to assigning predefined labels, and vision-language models (VLMs) face an inherent trade-off between informational richness and factual fidelity in their outputs, neither approach achieves the deep semantic interpretation required for reliable human-robot interaction. We propose Gold Points Sniper (GPS), a novel framework that empowers lightweight VLMs with self-guided multimodal reasoning capabilities for fine-grained human action understanding. Our approach comprises three key modules: Gold Points Extractor trains VLMs to identify critical action-relevant details, Selective Socratic Questioner validates and refines these details through selective self-questioning, and Semantic Entailment Evaluator quantitatively assesses factual consistency using semantic entailment classification. Extensive experiments on our curated instruction-tuning dataset based on the CAP benchmark demonstrate that GPS-enhanced lightweight VLMs achieve substantial performance improvements, with some models reaching performance comparable to proprietary GPT-4o while maintaining superior factual accuracy. Our work establishes a reliable foundation for fine-grained action understanding in domestic robotics, enabling robots to safely interpret human behavior through information-dense yet factually grounded descriptions. Source code, training configurations, annotation prompts, and dataset details are released at <https://github.com/Haodi-Liu/GPS-Gold-Point-Sniper>.

I. INTRODUCTION

The increasing adoption of robots in domestic environments necessitates a deeper understanding of human behavior. For robots to effectively collaborate, assist, and socialize with people, the ability to accurately recognize and interpret human actions is essential. This capability extends beyond merely identifying predefined activities from cropped, human-centric images; it requires sophisticated understanding of human actions, intentions, and contextual cues in broader views where humans occupy only a small portion of the image (see Fig. 1 top). However, neither open-vocabulary action recognition [1–8] nor vision-language models (VLMs) [9–14] fully achieves this level of deep semantic interpretation, as each approach suffers from inherent shortcomings.

¹ Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing, P.R.China

² State Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI).

† Corresponding authors. Zeyu Zhang <zhangzeyu@bigai.ai>, Changshui Zhang <zcs@mail.tsinghua.edu.cn>

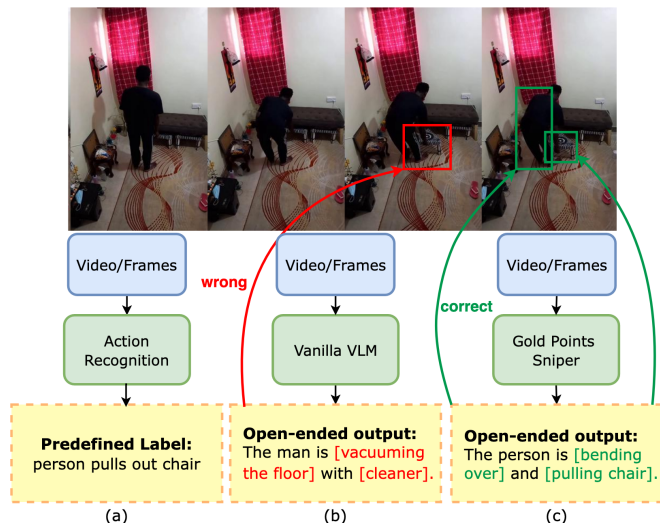


Fig. 1: The image on top is an example video frames from CAP [28] with label "person pulls out chair". (a). Action recognition gives precise but limited labels. (b). VLMs' outputs are informative but error-prone (shown in red). (c). Our GPS framework enables VLMs to precisely capture multiple fine-grained action clues (in green).

Despite technical advancements in open-vocabulary action recognition, including cross-modal alignment [1–3], knowledge augmentation [4, 5], and generalizable learning [6–8], current methods remain fundamentally limited to assigning predefined action labels [15–18] to visual inputs, as illustrated in Fig. 1 (a). These approaches fail to capture details about individuals, their environment, and the nuances of their interactions, thereby precluding deeper semantic interpretation and reasoning. While VLMs offer a promising alternative with their strong expressive capability for open-ended generation, they face an inherent trade-off between informational richness and factual fidelity: succinct descriptions often omit critical details, while verbose ones become prone to redundancy and hallucinations, as demonstrated in Fig. 1 (b). Although recent work has attempted to address these issues through structured reasoning [19–23] and self-questioning [24–27], this fundamental trade-off remains a significant limitation.

In this work, we aim to empower VLMs with self-guided generation capabilities for fine-grained human action understanding. This approach ensures outputs that distill core, task-relevant information into summaries that are both information-dense and factually accurate, as illustrated in Fig. 1 (c). Our objective is to enable VLMs to generate grounded and insightful summaries, capturing key attributes of people and scenes while elucidating the details of their

interactions. However, achieving this goal is hindered by three major challenges: (i) Lightweight VLMs often exhibit limited visual grounding capabilities, making it challenging to accurately identify and extract key details in response to direct prompts. (ii) VLMs’ inherent tendency toward hallucination poses significant risks during self-guided reasoning. Self-generated intermediate steps become contaminated, causing errors to accumulate through the reasoning process and result in fundamentally flawed outputs. (iii) The lack of appropriate metrics for assessing semantic accuracy and systematic frameworks to distinguish facts from hallucinations creates significant evaluation challenges for open-ended VLM outputs, hindering reliable assessment essential for trustworthy deployment.

To address these challenges, we propose **Gold Points Sniper (GPS)**, a framework that trains VLMs to self-guide multimodal reasoning and evaluates open-ended outputs through semantic entailment. Gold Points, our core concept, represent fine-grained key information highly relevant to human actions. Our framework comprises three main modules: (i) The **Gold Points Extractor** trains VLMs to accurately capture critical, task-relevant details from visual inputs, enhancing visual grounding capabilities. By focusing on grounded core details, VLMs can better guide their subsequent reasoning. (ii) The **Selective Socratic Questioner** trains VLMs to verify and refine extracted gold points through selective self-questioning. This internal validation ensures reasoning step integrity, mitigating error compounding and leading to reliable final summaries. (iii) The **Semantic Entailment Evaluator** employs semantic entailment classification to quantitatively assess factual consistency between VLM outputs and ground truth details. To facilitate framework training, we introduce an instruction-tuning dataset based on Consented Activity of People (CAP) [28] that benchmarks semantic alignment of VLM outputs against fine-grained action understanding. Our experiments demonstrate that our assessment module aligns closely with human intuition, and the overall framework enables multiple lightweight VLMs to generate more detailed and precise descriptions of human actions.

In summary, our contribution is threefold:

- 1) We propose the Gold Points Sniper (GPS) framework that enables lightweight VLMs to perform self-guided multimodal reasoning for fine-grained human action understanding, achieving information-dense and factually accurate outputs.
- 2) We curate a comprehensive instruction-tuning dataset based on the CAP [28] that facilitates training and evaluation of semantic alignment between VLM outputs and fine-grained action understanding in vision-language reasoning tasks.
- 3) Through extensive experiments, we demonstrate that our framework enables multiple lightweight VLMs to generate significantly more detailed and precise descriptions of human actions while maintaining factual accuracy.

II. RELATED WORK

A. Open-Vocabulary Action Recognition

Recent advances in open-vocabulary action recognition (OVAR) have been driven by CLIP-series models [29–32], with methodologies categorized into three paradigms. **Cross-modal alignment** approaches [1–3] have evolved from coarse video-label matching to fine-grained strategies that establish detailed correspondences between visual and textual modalities, while addressing issues like scene bias and noisy labels. **Knowledge distillation** methods [6,7] transfer semantic understanding from powerful VLMs to efficient architectures through teacher-student frameworks and offline knowledge distillation. **Prompt engineering** techniques [4,5] inject domain knowledge by leveraging LLMs to generate enriched textual representations and coordinate multiple model descriptions. Despite leveraging VLMs’ multimodal capabilities, these approaches remain limited to predefined action labels, hindering fine-grained understanding of human activities and their surroundings. In our work, we propose the GPS framework that exploits VLMs to generate detailed descriptions of actions, intentions, and contextual cues, establishing a reliable and informative perceptual foundation for domestic robots.

B. Structured Reasoning and Self-Questioning

Extensive research has focused on enabling LLMs and VLMs to generate high-quality responses aligned with user demands through two primary approaches. **Structured reasoning**, pioneered by Chain-of-Thought (CoT) [19], demonstrates that step-by-step reasoning processes enhance LLM performance [20,22]. **Self-questioning** techniques guide models toward focusing on salient details for more meaningful responses by providing higher-quality questions. For example, SQ-LLaVA [26] and Socratic Questioning [27] generate content-relevant questions about images, while other works [24] refine this by selectively applying self-questioning only to problems requiring in-depth reasoning. However, these methods face a trade-off between informational richness and factual fidelity. Our GPS framework integrates reliable knowledge accumulation, intermediate result verification, and self-questioning refinement for fine-grained human action understanding. Such factual grounding is essential for safe and reliable domestic robots, as inaccurate perception of human behavior can lead to unsafe actions and erode user trust in physical world interactions.

III. METHOD

This section details each module of the GPS framework shown in Fig. 2. While training and evaluating our framework requires extensive data annotation, a process made costly by manual labor or proprietary APIs, we bypass these bottlenecks by leveraging powerful open-source models to automate data annotation and outputs evaluation. Specifically, we employ LLaVA-OneVision-Qwen2-72B-ov-chat [14] for annotation and LLaMA-3.3-70B-Instruct [33] for evaluation.

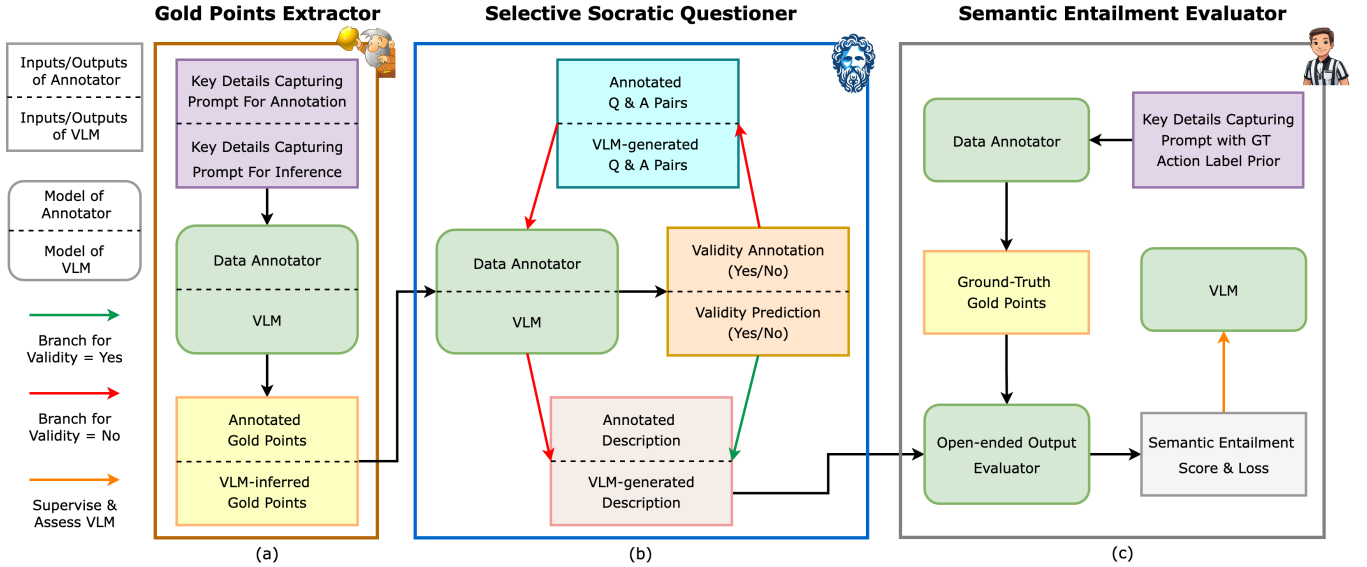


Fig. 2: Overview of the GPS framework. Dashed blocks are split into an upper section for data annotation and a lower section for VLM inference, with the annotated data used to fine-tune the VLM. Arrows denote data flow, and colored arrows highlight specific relationships. Video frames serve as the implicit input to all Data Annotator and VLM blocks. The Semantic Entailment Evaluator module has no dashed block because annotation and inference workflows diverge over there. The data annotator and output evaluator employed in our work are LLaVA-OneVision-Qwen2-72B-ov-chat [14] and LLaMA-3.3-70B-Instruct [33] respectively.

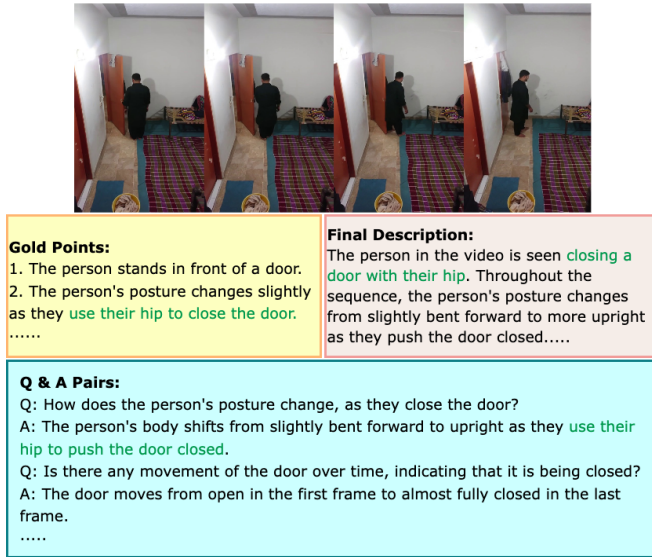


Fig. 3: Example video frames with GT label "person closes door with hip". Associated gold points, Socratic Q&A pairs, and final description are provided as demonstrations, with key action information highlighted in green.

A. Gold Points Extractor

As illustrated in Fig. 2 (a), the Gold Points Extractor module trains a VLM to identify critical action-relevant details from input video frames and formulate them as Gold Points (see yellow box in Fig. 3 for examples).

Let X_v denote the input video frames, \mathcal{G} the VLM serving as the Gold Points Extractor, X_{cap} the key details capturing prompt for inference, and $X_{gold} = \{X_{gold,i}\}_{i=1}^N$ a set of N **VLM-inferred Gold Points** (where N varies by case). The gold points extraction operation is:

$$X_{gold} = \mathcal{G}(X_{cap}, X_v) \quad (1)$$

as our gold points extraction operation.

The data pairs used to train \mathcal{G} are $\{X_v, X_{gold}^{ann}\}$, where the **Annotated Gold Points** X_{gold}^{ann} are generated by:

$$X_{gold}^{ann} = \mathcal{A}(X_{cap}^{ann}, X_v) \quad (2)$$

Here, \mathcal{A} represents the data annotator operating on video frames X_v with annotation prompt X_{cap}^{ann} . The training objective \mathcal{O}_{GPE} maximizes the likelihood of generating each annotated gold point $X_{gold,i}^{ann}$ given the visual input X_v and all previously produced points $X_{gold,<i}^{ann}$:

$$\mathcal{O}_{GPE} = \sum_{i=1}^N \log P(X_{gold,i}^{ann} | X_v, X_{cap}^{ann}, X_{gold,<i}^{ann}) \quad (3)$$

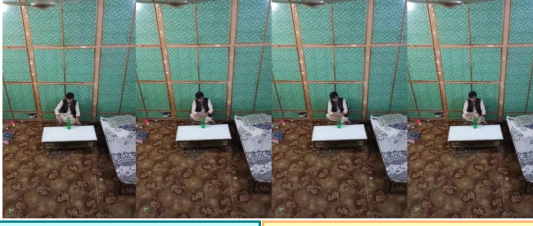
Inference prompts (e.g., X_{cap}) are concise versions of their annotation counterparts (e.g., X_{cap}^{ann}), abbreviated to meet fine-tuning token limits. For simplicity, we henceforth use data notation (e.g., X_{gold}) as shorthand for the complete prompt containing it when displayed as model input.

The inferred gold points X_{gold} subsequently guide the VLM's reasoning throughout downstream tasks, serving as foundational reference. In optimal cases, these points capture all critical information about actions, intentions, and context. Even when incomplete, they still provide substantial visual cues that establish a solid foundation for extracting and integrating finer-grained details in following processing stages.

B. Selective Socratic Questioner

Since X_{gold} may be incomplete or flawed, Selective Socratic Questioner (Fig. 2 (b)) trains the VLM to validate and selectively refine these points through a two-stage process.

Stage 1: Validating Gold Points. The VLM first assesses whether the gold points are accurate and sufficient for understanding the subject's action, intent, and context. Let



Open-ended VLM Output:

The frames are showing a person seated at a small white table with a green bottle placed on it and he is either preparing to drink from the bottle or has just finished doing so....

Ground Truth Gold Points:

1. The person is seated at a table with a green bottle in front of them.
2. The sequence of frames shows a progression that the person is screwing a lid onto a bottle
-

Entailment Relations & Reasons:

1. [2] The summary description directly supports the proposition that the person is seated at a table with a green bottle in front of them.
2. [0] The description says the person is about to drink from the bottle, indicating an action of screwing the lid off the bottle, contradicting the progression of screwing a lid onto a bottle
-

Fig. 4: Example video frames with GT label “person screws lid to bottle”. Associated GT gold points, entailment relations (Contradiction 0/Neutrality 1/Entailment 2) with reasons, and VLM output are shown, where entailments of key action information are highlighted in green and contradictions are highlighted in red.

\mathcal{V} denote the validator VLM. The validation process is:

Let \mathcal{V} be the VLM playing the role of validator, then the gold points validation process can be formulated as

$$y = \mathcal{V}(X_v, X_{gold}) \quad (4)$$

where y is a boolean indicator (1 for valid, 0 for invalid).

To better align with real inference, we initially annotate X_{gold}^{ann} without access to the GT action label B . For training \mathcal{V} , we then provide the annotator \mathcal{A} with B to determine whether X_{gold}^{ann} alone suffices to robustly predict B from X_v :

$$y^{ann} = \mathcal{A}(X_v, X_{gold}^{ann}, B) \quad (5)$$

where y^{ann} is the annotated boolean validity. Thereafter, $\{(X_v, X_{gold}^{ann}), y^{ann}\}$ serve as training data for \mathcal{V} and the training objective \mathcal{O}_{valid} is the negation of BCE loss:

$$\mathcal{O}_{valid} = -BCE(y, y^{ann}) \quad (6)$$

Stage 2: Selective Questioning Refinement. The module’s workflow branches based on the validity assessment. When X_{gold} are deemed valid ($y = 1$), the describer VLM \mathcal{D} generates the final description X_{des} directly:

$$X_{des} = \mathcal{D}(X_v, X_{gold} | y = 1) \quad (7)$$

The corresponding training data for \mathcal{D} consists of $\{(X_v, X_{gold}^{ann}), X_{des}^{ann}\}$ where $y^{ann} = 1$, with annotated description X_{des}^{ann} obtained via:

$$X_{des}^{ann} = \mathcal{A}(X_v, X_{gold}^{ann} | y^{ann} = 1) \quad (8)$$

When X_{gold} are deemed invalid ($y = 0$), Selective Socratic Questioner initiates a refinement process featuring self-questioning (Socratic Questioning). Specifically, a VLM \mathcal{Q} , acting as a *socratic questioner*, queries itself about omissions

or inaccuracies within X_{gold} in order to direct its attention to relevant visual regions. As a result, the discovered valuable cues are formatted into Q&A pairs X_{socra} .

$$X_{socra} = \{(X_{q,j}, X_{a,j})\}_{j=1}^M = \mathcal{Q}(X_v, X_{gold} | y = 0) \quad (9)$$

where the number of Q&A pairs M varies by case. The training data $\{(X_v, X_{gold}^{ann}), X_{socra}^{ann}\}$, with associated $y^{ann} = 0$, for \mathcal{Q} is obtained by

$$X_{socra}^{ann} = \mathcal{A}(X_v, X_{gold}^{ann} | y^{ann} = 0) \quad (10)$$

where X_{socra}^{ann} comprises annotated Q&A pairs. And the training objective \mathcal{O}_{socra} maximizes the likelihood of producing each annotated Q&A pair $X_{socra,j}^{ann} = (X_{q,j}^{ann}, X_{a,j}^{ann})$ given X_v and all previous rounds $X_{socra,<j}^{ann} = (X_{q,<j}^{ann}, X_{a,<j}^{ann})$:

$$\mathcal{O}_{socra} = \sum_{j=1}^M \left[\log P(X_{a,j}^{ann} | X_v, X_{gold}^{ann}, X_{q,j}^{ann}) + \log P(X_{q,j}^{ann} | X_v, X_{gold}^{ann}, X_{q,<j}^{ann}, X_{a,<j}^{ann}) \right] \quad (11)$$

Next, X_{socra} augments X_{gold} to generate the final, improved description X_{des} : (Example Q&A pairs and final description are shown in Fig. 3).

$$X_{des} = \mathcal{D}(X_v, X_{gold}, X_{socra} | y = 0) \quad (12)$$

The training data for \mathcal{D} is $\{(X_v, X_{gold}^{ann}, X_{socra}^{ann}), X_{des}^{ann}\}$ where $y^{ann} = 0$ and the annotated description X_{des}^{ann} is obtained via,

$$X_{des}^{ann} = \mathcal{A}(X_v, X_{gold}^{ann}, X_{socra}^{ann} | y^{ann} = 0) \quad (13)$$

Considering both possibilities of validity, the training objective \mathcal{O}_{des} maximizes the likelihood of generating annotated description X_{des}^{ann} given verified intermediate results:

$$\mathcal{O}_{des} = \log P(X_{des}^{ann} | X_v, X_{gold}^{ann}) y^{ann} + \log P(X_{des}^{ann} | X_v, X_{gold}^{ann}, X_{socra}^{ann}) (1 - y^{ann}) \quad (14)$$

C. Semantic Entailment Evaluator

To quantitatively assess how precisely a VLM’s open-ended description X_{des} portrays the actions, intentions, and context of individuals in visual inputs, the Output Evaluator \mathcal{E} within our Semantic Entailment Evaluator module Fig. 2 (b) ascertains the factual consistency between X_{des} and a set of ground truth key details $X_{gold}^{GT} = \{X_{gold,i}\}_{i=1}^K$ (termed **Ground Truth Gold Points**). Unlike the annotation of X_{gold}^{ann} , we provide \mathcal{A} with GT action label B as a strong prior when generating X_{gold}^{GT} to ensure fidelity ($X_{gold,K}$ explicitly contains B). For each ground truth point in X_{gold}^{GT} , \mathcal{E} classifies its semantic relationship with X_{des} as entailment [2], neutral [1], or contradiction [0]:

$$R = \mathcal{E}(X_{des}, X_{gold}^{GT}) \quad (15)$$

where $R = \{R_i\}_{i=1}^K \in \{0, 1, 2\}^K$ represents the determined entailment relations. A demonstration of semantic entailment evaluation outcomes is given in Fig. 4.

From these relations, we compute the Semantic Entailment

Score $S_{\text{sem_ent}} \in [0, 1]$:

$$S_{\text{sem_ent}} = \frac{1}{2} \left(\frac{N_e - N_c}{K} + 1 \right) \quad (16)$$

where N_e and N_c denote the counts of entailments and contradictions in R , respectively. The initial score $S_{\text{ent_init}} \in [-1, 1]$ is normalized to obtain $S_{\text{sem_ent}}$. This score serves as an objective metric for semantic consistency and will measure VLMs’ action understanding performance in **Sec. IV**.

Since a single VLM performs all roles—gold points extractor \mathcal{G} , validator \mathcal{V} , questioner \mathcal{Q} , and describer \mathcal{D} in this work—we combine all objectives for unified fine-tuning:

$$\mathcal{O}_{\text{overall}} = \mathcal{O}_{\text{GPE}} + \mathcal{O}_{\text{valid}} + \mathcal{O}_{\text{socra}} + \mathcal{O}_{\text{des}} \quad (17)$$

This comprehensive objective enhances the VLM’s fine-grained reasoning capabilities across all module functions.

IV. EXPERIMENT

We conduct extensive experiments across a range of lightweight VLMs to demonstrate the advantages of our GPS framework. Our experiments aim to investigate two key aspects of the GPS framework: (i) the performance boost when incorporating the GPS framework on VLMs for fine-grained human action understanding tasks and the adaptability to unseen scenarios, and (ii) the necessity of the GPS framework (*e.g.*, whether it could be replaced by Chain-of-Thought prompt engineering). Additionally, we conduct an ablation study to analyze the contribution of each core module in the GPS framework, providing a thorough understanding of our proposed approach.

A. Experiment Setup

Benchmark: We construct an *instruction-tuning dataset* and an *assessment benchmark* using CAP [28], a corpus designed for fine-grained human action recognition. To be helpful for fine-grained action understanding, all selected action categories feature rich subject interactions. Specifically, the instruction-tuning dataset consists of 14,281 video frames from 80 action categories. For training, it provides video frames X_v , annotated gold points $X_{\text{gold}}^{\text{ann}}$, ground-truth gold points $X_{\text{gold}}^{\text{GT}}$, validation y^{ann} , optional socratic Q&A pairs $X_{\text{socra}}^{\text{ann}}$ and final description $X_{\text{des}}^{\text{ann}}$. In contrast, the assessment benchmark is designed purely for evaluation and contains 3,321 video frames from 40 categories, comprising only video frames X_v and their ground truth gold points $X_{\text{gold}}^{\text{GT}}$. To evaluate generalization, the assessment benchmark is further divided into a held-in set (20 categories, 1,117 frames) and a held-out set (20 categories, 2,204 frames). The action categories in the held-in set are seen during instruction tuning, but the video frames are from different scenarios. Conversely, the categories in the held-out set are absent from the tuning set, enabling zero-shot evaluation.

Models: We select four SOTA lightweight LVLMs: Qwen-VL-Chat [9], MiniGPT-v2 [11], LLaMA3-LLaVA-NeXT-8B (LLaVA-NeXT-8B) [13] and LLaVA-OneVision-Qwen2-7B (LLaVA-OV-7B) [14] for our experiments and VLMs’ fine-tunings are performed on four A800-SXM4-80GB. In addition

to these lightweight models, we also evaluate a SOTA open-source large VLM, LLaVA-OneVision-Qwen2-72B-ov-chat (LLaVA-OV-72B) [14], and the proprietary model, GPT-4o [12], on our benchmark. This serves a dual purpose: first, to assess the capabilities of leading large-scale VLMs on our specialized task of fine-grained human action understanding; and second, to compare their performances against our enhanced lightweight models to highlight the effects of improvements.

Finetuning & Prompting: The GPS framework, as detailed in **Sec. III**, is built upon a four-stage reasoning structure: gold points extraction, gold points validation, selective questioning refinement, and fine-grained description. The application of this framework consists of two distinct phases: In the **Finetuning phase**, the instruction-tuning dataset is used to explicitly train the VLMs on how to perform each individual stage of the reasoning structure. Subsequently, in the **Prompting phase**, the VLM is prompted to replicate this trained, multi-stage process in inference. Critically, during inference necessary inputs like gold points, validation outcome and refined Q&A pairs are generated by the VLM in real-time and fed into its own subsequent reasoning steps via appropriate prompts.

B. Performance Boost with GPS

In this experiment we evaluate to what extent, the GPS framework boost the performance of the lightweight VLMs and whether such improvements can generalize to new scenarios and new actions. To highlight the effectiveness of the GPS framework, we evaluate the performance of various vanilla VLMs against their GPS-enhanced counterparts on our benchmark, with performance measured by semantic entailment scores **16**. The term “GPS-Enhanced” indicates that lightweight VLMs adopt GPS fine-tuning and prompting phases, while for LLaVA-OV-72B and GPT-4o, where fine-tuning is infeasible, this refers solely to GPS prompting. To assess learning efficacy and zero-shot generalization, we report results on the whole benchmark as well as on held-in (seen) and held-out (unseen) sets. This evaluation protocol is maintained throughout all subsequent experimental sections.

As illustrated in **Fig. 5**, the evaluation results align closely with human intuition that more advanced models achieve higher ratings, supporting the credibility of our Semantic Entailment Evaluator. Compared to vanilla baselines, all lightweight VLMs achieve substantially better scores on held-in, held-out, and whole benchmark after GPS enhancement. Notably, GPS-enhanced LLaVA-NeXT-8B and LLaVA-OV-7B demonstrate performance comparable to GPT-4o on the held-in set, indicating that GPS effectively enhances lightweight VLMs’ fine-grained action understanding capabilities, producing information-dense and factually accurate outputs. For the already powerful GPT-4o and LLaVA-OV-72B, the effect of this multi-stage prompting is less pronounced. LLaVA-OV-72B’s minor performance dip can be attributed to instability from applying complex prompt engineering to a large model for such a complicated task.

Performance Comparison: Vanilla VLMs vs Their GPS-Enhanced Counterparts

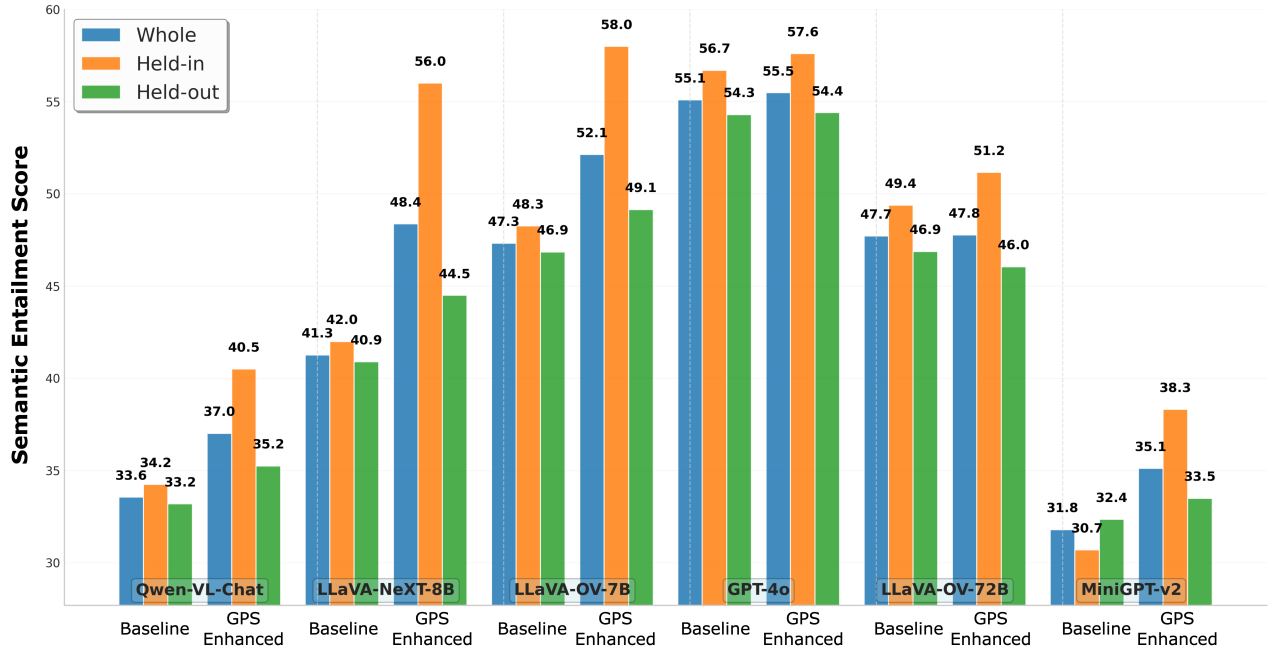


Fig. 5: Histogram showcasing the performances of various VLMs (vanilla & GPS-Enhanced version) on our benchmark.

The substantial improvements on the held-in set demonstrate that GPS effectively teaches VLMs to understand actions encountered during training. More importantly, despite smaller improvement margins, the consistent enhancements on the held-out set validate the zero-shot generalization capability of GPS-Enhanced VLMs across new scenarios and actions. This achievement is particularly significant given the held-out set’s inherent difficulty—vanilla models typically underperform here compared to the held-in set, and action categories are entirely unseen during training. Even modest improvements on this challenging set therefore represent substantial progress. GPS does not merely help models memorize training examples but teaches a generalizable analytical framework: observe key details, verify accuracy, and identify missing information. Since this analytical skill transcends specific tasks, it generalizes effectively to new scenarios and completely unseen actions.

Please offer a visually grounded summarizing description of the ongoing human action, as well as the associated intention and context. Reasoning supported by reliable external knowledge and visual details is permitted. Also be sure to leverage the temporal changes in the frames and state the recognized action (including the interacted objects if possible) with one sentence at last.

TABLE I: Generic prompt used as our baseline prompt.

C. Necessity of GPS Framework

In this experiment, we investigate the necessity of the GPS framework by analyzing whether its performance can be

Models	Semantic Entailment Score		
	Held-in	Held-out	Whole
Qwen-VL-Chat	34.25	33.20	33.56
+ Prompting	33.40 (↓2.48%)	31.40 (↓5.42%)	32.08 (↓4.41%)
+ Finetuning	34.90 (↑1.90%)	33.65 (↑1.36%)	34.07 (↑1.52%)
+ Combined	40.50 (↑18.25%)	35.25 (↑6.17%)	37.02 (↑10.31%)
MiniGPT-v2	30.70	32.35	31.80
+ Prompting	29.15 (↓5.05%)	31.00 (↓4.17%)	30.38 (↓4.47%)
+ Finetuning	37.00 (↑20.52%)	32.50 (↑0.46%)	34.02 (↑6.98%)
+ Combined	38.32 (↑24.82%)	33.50 (↑3.55%)	35.12 (↑10.44%)
LLaVA-NeXT-8B	42.00	40.90	41.27
+ Prompting	41.25 (↓1.79%)	37.10 (↓9.29%)	38.50 (↓6.71%)
+ Finetuning	54.40 (↑29.52%)	42.80 (↑4.65%)	46.70 (↑13.16%)
+ Combined	56.00 (↑33.33%)	44.50 (↑8.80%)	48.37 (↑17.20%)
LLaVA-OV-7B	48.26	46.85	47.32
+ Prompting	50.23 (↑4.08%)	46.30 (↓1.17%)	48.68 (↑2.87%)
+ Finetuning	54.90 (↑13.76%)	47.90 (↑2.24%)	49.20 (↑3.97%)
+ Combined	58.00 (↑20.18%)	49.15 (↑4.91%)	52.13 (↑10.16%)
LLaVA-OV-72B	49.38	46.88	47.72
+ Prompting	51.16 (↑3.60%)	46.05 (↓1.77%)	47.77 (↑0.10%)
GPT-4o	56.70	54.30	55.10
+ Prompting	57.60 (↑1.59%)	54.40 (↑0.18%)	55.48 (↑0.69%)

TABLE II: Assessing the independent and synergistic effects of GPS finetuning and prompting phases described in Sec. IV-A.

achieved through pure Chain-of-Thought prompt engineering. To validate this necessity, we isolate the prompting phase of GPS—itsself a form of CoT prompt engineering—from the overall framework to assess its individual effect relative to the finetuning phase. To this end, we designed the following four distinct experimental configurations: The **baseline configuration** employs a generic prompt Tab. I to guide a vanilla VLM in action description. The **+ Prompting** configuration involves only GPS prompting, directing the vanilla VLM to

follow GPS’s four-stage reasoning structure. The + **Fine-tuning** configuration consists solely of finetuning, where the VLM learns GPS reasoning stages but receives only generic prompting **Tab. I** at inference. Finally, the + **Combined** setting represents complete GPS Enhancement, first training the VLM on four-stage reasoning during finetuning, then prompting the trained model to replicate the full reasoning process during inference.

Results in **Tab. II** show that + **Prompting** caused performance degradation for nearly all lightweight VLMs, with only minor improvement in LLaVA-OV-7B. This reveals a key insight: complex, multi-stage prompts become cognitive burdens rather than helpful guidance for lightweight VLMs. Lacking inherent capability to follow four-stage reasoning, these models likely generate flawed Gold Points or fail at self-verification, causing the entire reasoning chain to collapse. This demonstrates that complex reasoning skills must be learned through targeted training rather than elicited by inference-time prompts. In contrast, + **Finetuning** led to substantial performance enhancements across all models, confirming that our training strategy is essential for equipping models with necessary skills. Peak performance was achieved exclusively under the + **Combined** setting, revealing crucial synergy between training and prompting in our framework. The finetuning phase endows models with GPS reasoning capabilities, while the prompting phase provides structured guidance during inference, ensuring these learned abilities are reliably invoked in correct sequence.

D. Ablation Study of Core Components

LLaVA-NeXT-8B	Held-in	Held-out	Whole
w/o GPE	53.35 (↓4.73%)	42.15 (↓5.28%)	45.92 (↓5.07%)
w/o GPV	53.75 (↓4.02%)	41.15 (↓7.53%)	45.39 (↓6.16%)
w/o SQR	56.80 (↑1.43%)	43.80 (↓1.57%)	48.17 (↓0.41%)
full GPS	56.00	44.50	48.37

LLaVA-OV-7B	Held-in	Held-out	Whole
w/o GPE	54.44 (↓6.14%)	45.52 (↓7.39%)	48.52 (↓6.93%)
w/o GPV	54.69 (↓5.71%)	45.55 (↓7.32%)	48.63 (↓6.71%)
w/o SQR	57.38 (↓1.07%)	48.60 (↓1.12%)	51.55 (↓1.11%)
full GPS	58.00	49.15	52.13

TABLE III: Ablations on prompting LLaVA-NeXT-8B and LLaVA-OV-7B (both GPS finetuned). Three ablated components are Gold Points Extraction (GPE), Gold Points Validation (GPV) and Selective Questioning Refinement (SQR).

To evaluate each module’s contribution in the GPS framework, we conduct a prompting ablation study on LLaVA-NeXT-8B and LLaVA-OV-7B models after GPS finetuning. During inference, we systematically ablate each of the three steps—GPE, GPV, and SQR (as shown in **Tab. III**)—from the prompting-based reasoning process while retaining the other two. We perform this ablation on finetuned VLMs because only models capable of properly executing each GPS reasoning stage can provide faithful assessment of component effectiveness. We examine three ablation settings: (i) **Without GPE**: The VLM uses only visual input to choose between direct description generation and self-questioning for additional visual clues. (ii) **Without GPV**: The VLM skips validation of generated gold points and proceeds directly to selective refinement, but lacks validation feedback to



GPT4o: The person is repeatedly lifting a dumbbell with both hands in front of their body.

LLaVA-OV-7B: The image is a sequence of five identical frames showing a person in a room. The person appears to be wearing dark clothing and is holding a black object that resembles a camera or a phone. They are standing near a bed with blue and beige bedding, and there is a small table with various items on it next to the bed. The room has a light-colored wall and a ceiling fan above. The floor has a patterned carpet. In each frame, the person seems to be in the same pose, suggesting minimal movement between the images.

GPS-Enhanced LLaVA-OV-7B: The person in the video is engaged in the action of pouring a liquid from a kettle into a cup. The sequence of frames shows the individual holding the kettle and the cup, raising the kettle above their head to initiate the pouring process, continuing the pour with the kettle still raised, lowering the kettle back towards the cup to complete the action, and finally placing the kettle back on the table. The recognized action is "pouring liquid from a kettle into a cup".

Fig. 6: A qualitative example, with GT label “person pours coffee into mug”, demonstrating the usage outcomes of vanilla GPT-4o, vanilla LLaVA-OV-7B and GPS-Enhanced LLaVA-OV-7B on a video frames from our benchmark.

guide this selection. (iii) **Without SQR**: The VLM generates the final description based solely on generated gold points and their validation results without further refinement.

The ablation results in **Tab. III** confirm that all components contribute meaningfully while revealing their distinct functional roles. Gold Points Extraction (GPE) and Gold Points Validation (GPV) serve as foundational components, with their removal causing the most substantial performance drops, demonstrating that the reasoning chain depends critically on accurately extracting and verifying core factual information. After GPE extracts facts and GPV validates them, SQR identifies missing details and resolves ambiguities, transforming factually correct descriptions into more informative outputs. The smaller performance drop from SQR removal indicates its primary role is improving output quality rather than establishing correctness. This ablation study demonstrates that GPS effectively balances informational richness and factual fidelity by first ensuring accurate fact extraction and validation, then systematically enhancing descriptive detail without compromising correctness.

E. Qualitative Evaluation

Fig. 6 presents qualitative results of GPS-Enhanced human action understanding, demonstrating the effectiveness of the GPS framework. The baseline models show critical failures in fine-grained action understanding despite adequate general scene description. GPT-4o hallucinates both object and action, misinterpreting pouring from a kettle as “lifting a dumbbell”, while vanilla LLaVA-OV-7B correctly perceives the static environment but fails to grasp temporal dynamics, incorrectly concluding the frames are “identical” with

“minimal movement”. In contrast, the GPS-Enhanced model accurately identifies and sequences the core action. Rather than providing static descriptions, it correctly decomposes the event into constituent steps: holding objects, raising the kettle, pouring, and lowering it back. This demonstrates that GPS enables VLMs to move beyond coarse scene-level perception to structured, step-by-step reasoning that accurately captures fine-grained action semantics.

V. CONCLUSION

We introduced the GPS framework to enable fine-grained human action understanding for domestic robots through self-guided multimodal reasoning in lightweight VLMs. Our three-module approach—Gold Points Extraction, Selective Socratic Questioning, and Semantic Entailment Evaluation—resolves the trade-off between informational richness and factual fidelity in vision-language models. Experiments demonstrate that GPS-enhanced lightweight VLMs achieve substantial improvements across seen and unseen action categories, with some reaching performance comparable to larger proprietary systems while maintaining superior factual accuracy. The framework’s generalization to new scenarios validates its real-world applicability, establishing a systematic approach for information-dense yet factually grounded action descriptions that enable trustworthy human-robot interaction in domestic environments.

VI. ACKNOWLEDGEMENT

This work was supported by National Science and Technology Major Project (No. 2022ZD0114903) and Beijing Natural Science Foundation (funding Number L258013).

REFERENCES

- [1] Y. Chen, D. Chen, R. Liu, H. Li, and W. Peng, “Video action recognition with attentive semantic units,” in *International Conference on Computer Vision (ICCV)*, pp. 10170–10180, 2023.
- [2] R. Zhang, J. Xue, F. Lin, Q. Zhang, P. Smirnov, X. Ma, and X. Yan, “Enhancing human action recognition with fine-grained body movement attention,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2024.
- [3] H. Cheng, C. Ju, H. Wang, J. Liu, M. Chen, Q. Hu, X. Zhang, and Y. Wang, “Denoiser: Rethinking the robustness for open-vocabulary action recognition,” *arXiv preprint arXiv:2404.14890*, 2024.
- [4] C. Jia, M. Luo, X. Chang, Z. Dang, M. Han, M. Wang, G. Dai, S. Dang, and J. Wang, “Generating action-conditioned prompts for open-vocabulary video action recognition,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4640–4649, 2024.
- [5] L. Chen, B. Li, S. Shen, J. Yang, C. Li, K. Keutzer, T. Darrell, and Z. Liu, “Large language models are visual reasoning coordinators,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 70115–70140, 2023.
- [6] X. Huang, H. Zhou, K. Yao, and K. Han, “Froster: Frozen clip is a strong teacher for open-vocabulary action recognition,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [7] H. Lu, H. Jian, R. Poppe, and A. A. Salah, “Enhancing video transformers for action understanding with vlm-aided training,” *arXiv preprint arXiv:2403.16128*, 2024.
- [8] Y. Yu, C. Cao, Y. Zhang, and Y. Zhang, “Learning to generalize without bias for open-vocabulary action recognition,” *arXiv preprint arXiv:2502.20158*, 2025.
- [9] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” *arXiv preprint arXiv:2308.12966*, 2023.
- [10] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.
- [11] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, “Minigt-v2: large language model as a unified interface for vision-language multi-task learning,” *arXiv preprint arXiv:2310.09478*, 2023.
- [12] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [13] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li, “Llava-next: Stronger llms supercharge multimodal capabilities in the wild,” May 2024.
- [14] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, “Llava-onevision: Easy visual task transfer,” *Transactions on Machine Learning Research*, 2024.
- [15] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [16] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015.
- [17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “Hmdb: a large video database for human motion recognition,” in *International Conference on Computer Vision (ICCV)*, pp. 2556–2563, IEEE, 2011.
- [18] K. Soomro, A. R. Zamir, and M. Shah, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24824–24837, 2022.
- [20] Z. Zhang, A. Zhang, M. Li, hai zhao, G. Karypis, and A. Smola, “Multimodal chain-of-thought reasoning in language models,” *Transactions on Machine Learning Research*, 2024.
- [21] L. He, Z. Li, X. Cai, and P. Wang, “Multi-modal latent space learning for chain-of-thought reasoning in language models,” in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 18180–18187, 2024.
- [22] D. Mondal, S. Modi, S. Panda, R. Singh, and G. S. Rao, “Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning,” in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, pp. 18798–18806, 2024.
- [23] Y. Zhang, J. Yang, Y. Yuan, and A. C.-C. Yao, “Cumulative reasoning with large language models,” *Transactions on Machine Learning Research*, 2023.
- [24] H. Zhang, J. Liu, Z. Han, S. Chen, B. He, V. Tresp, Z. Xu, and J. Gu, “Visual question decomposition on multimodal large language models,” in *EMNLP (Findings)*, 2024.
- [25] J. Chen, B. Wang, Z. Jiang, and Y. Nakashima, “Putting people in llms’ shoes: Generating better answers via question rewriter,” in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 39, pp. 23577–23585, 2025.
- [26] G. Sun, C. Qin, J. Wang, Z. Chen, R. Xu, and Z. Tao, “Sq-llava: Self-questioning for large vision-language assistant,” in *European Conference on Computer Vision (ECCV)*, pp. 156–172, Springer, 2024.
- [27] W. Hu, H. Liu, L. Chen, F. Zhou, C. Xiao, Q. Yang, and C. Zhang, “Socratic questioning: Learn to self-guide multimodal reasoning in the wild,” *arXiv preprint arXiv:2501.02964*, 2025.
- [28] J. Byrne, G. Castañón, Z. Li, and G. Ettinger, “Fine-grained activities of people worldwide,” in *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, pp. 3308–3319, 2023.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, pp. 8748–8763, PMLR, 2021.
- [30] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, “X-clip: End-to-end multi-grained contrastive learning for video-text retrieval,” in *Proceedings of the 30th ACM international conference on multimedia*, pp. 638–647, 2022.
- [31] H. Rasheed, M. U. Khattak, M. Maaz, S. Khan, and F. S. Khan, “Fine-tuned clip models are efficient video learners,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6545–6554, 2023.
- [32] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, W. HongFa, Y. Pang, W. Jiang, J. Zhang, Z. Li, *et al.*, “Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [33] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.