Visual Search Augmented Chain-of-Thought Reasoning for Attribute Value Extraction from Product Videos

Tong Wu[1], **Ming Cheng**[1], **Jiazhen Hu**[1], **Jiaying Gong**[2*], **Hoda Eldardiry**[1]
[1]Virginia Tech, [2]Amazon
{tongw,ming98,hjiazhen,hdardiry}@vt.edu, gojiayin@amazon.com

Abstract

Existing approaches to visual attribute value extraction (AVE) primarily rely on static product images, failing to capture temporal cues, multi-angle views and fine-grained visual details. Directly applying video vision-language models (VLMs) to product AVE results in limited performance due to the lack of domain knowledge, and fine-tuning them requires extensive high-quality data and substantial computational resources. Thus, we propose visual search augmented chain-of-thought reasoning (ViS-CoT), a training-free, plug-and-play pipeline that can be easily applied to any open-source video VLM for video-to-text AVE in e-Commerce. Specifically, ViS-CoT employs visual clustering to identify representative frames, followed by visual search to retrieve semantically similar product knowledge that can enrich attribute cues. Next, an interleaved CoT reasoning module iteratively refines reasoning through visually-aligned auxiliary texts derived from captioning and automatic speech recognition. Finally, the integrated information guides the model toward accurate and fine-grained attribute predictions. Extensive experiments across 14 product categories on the VideoAVE dataset show that ViS-CoT consistently enhances multiple state-of-the-art video VLMs, achieving an average improvement of **17.91** percentage points in micro-F1.

1 Introduction

Visual attribute value extraction (AVE) is a core problem in computer vision and e-Commerce that seeks to automatically identify and structure fine-grained product information such as brand, material, types, and item forms from multimodal content. Reliable extraction of such attributes is crucial not only for improving product retrieval, recommendation, and filtering, but also for enabling downstream tasks such as large-scale catalog management and standardization across sellers and marketplaces (Gong et al., 2025b).

Existing research on product visual AVE has primarily relied on static product image inputs (Gong et al., 2025b,a; Zou et al., 2024b,a; Khandelwal et al., 2023). However, many attributes cannot be reliably inferred from the product images alone, because the static nature of single or even multiple product images often fails to capture fine-grained details, temporal cues, and multi-angle views. Figure 1 shows that relying solely on multi-image sets may overlook fine-grained usage attributes. While static images display general characteristics, they often fail to capture key functional details such as closure type and protection type, which are more effectively revealed through dynamic video frames. In contrast, product videos can provide richer visual information as they cover all sides of an item through continuous rotation, revealing subtle product properties (i.e. color, material, etc.) under changing environments (i.e. lighting condition, usage scenarios, etc.). These dynamic views help reduce ambiguity, lower the chance of missing attributes, and improve the alignment with how consumers perceive products in real e-Commerce platforms, leading to more accurate and complete attribute value extraction.

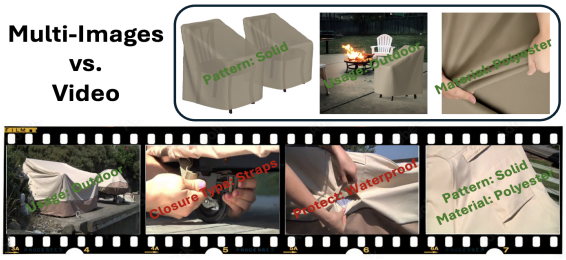


Figure 1: Comparison of product attribute value generation from multi-images vs. from videos.

Recent work (Cheng et al., 2025) introduced video-to-text AVE benchmarks to address the lim-

*This work was completed prior to joining Amazon, and does not relate to the author's position at Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48238–48254
October 24-29, 2026 ©2026 Association for Computational Linguistics

itations of product image inputs. However, current video VLMs typically rely on uniform frame sampling, which often captures product-irrelevant, noisy or redundant content which results in hallucinations. While advances in general video understanding (i.e. adaptive trees (Wang et al., 2025d), differential sliding-window captioning (Chen et al., 2024a), etc.) have improved semantic modeling of video dynamics, these methods differ fundamentally from product AVE. General video understanding emphasizes what happens in the video (events, narratives, storytelling), whereas product video understanding focuses on what the item is and its properties, leveraging motion primarily to reveal attributes rather than to construct a narrative.

To address the limitations, we propose **ViS-CoT**, a visual search augmented chain-of-thought framework for product video understanding. The key idea is to combine retrieval-based grounding with staged multimodal reasoning. Unlike general video understanding for storytelling, product video understanding aims at extracting structured, fine-grained product attributes by solving visual ambiguity between attributes. ViS-CoT consists of four stages: (1) **key frame extraction**, where visual clustering selects representative frames from the video; (2) **visual search**, which retrieves similar products to provide attribute priors; (3) **interleaved CoT reasoning**, which iteratively refines attribute hypotheses using captions and ASR transcripts; and (4) **integration and generation**, where multimodal evidence is combined to produce final attribute predictions. We evaluate ViS-CoT across 14 categories on a video-to-text e-Commerce dataset VideoAVE (Cheng et al., 2025). By applying ViS-CoT to five SOTA open-source video VLMs, we achieve an average improvement of **17.91** percentage points in micro-F1 on VideoAVE, while achieving competitive performance compared with fine-tuned video VLMs.

Our main contributions:

- We introduce **ViS-CoT**, the first plug-and-play visual search-augmented interleaved CoT pipeline designed to enhance any video VLM. ViS-CoT is training-free and broadly applicable, offering a resource-efficient solution for video-to-text AVE in e-Commerce.
- We observed a key technical insight: the **two-stage interleaved design** (vision-only hypothesis → retrieval-grounded refinement) is more effective than using visual search or CoT individually, improving attribute prediction by grounding reasoning in retrieved product knowledge.
- We conduct extensive experiments across 14 product categories using five state-of-the-art video VLMs, demonstrating significant performance improvements and validating the contribution of each component.

2 Related Work

Attribute Value Extraction Early research on AVE has mainly focused on unimodal approaches, where product attributes are derived solely from textual inputs such as titles or descriptions (Gong and Eldardiry, 2024; Fang et al., 2024a; Li et al., 2024b; Gong et al., 2023; Deng et al., 2023; Shinzato et al., 2023; Yang et al., 2023; Blume et al., 2023; Brinkmann and Bizer, 2025; Su et al., 2025). However, by relying only on text, these models often fail to capture the rich visual information and multimodal correlations in product images. Subsequent work has shifted toward vision language models (VLMs) (Bai et al., 2025; Grattafiori et al., 2024; Chen et al., 2024c; Hong et al., 2025; Comanici et al., 2025; Li et al., 2024a; Achiam et al., 2023; Abdin et al., 2024), to capture the rich visual information and multimodal correlations in product images (Gong et al., 2025b; Trabelsi et al., 2025; Gong et al., 2025a; Hu et al., 2025a; Hongwimol et al., 2025; Zou et al., 2024a,b; Khandelwal et al., 2023; Zhang et al., 2023b). While VLMs enable the inference of some implicit attributes that are not explicitly stated in product descriptions (i.e., material, pattern, etc.), they are still limited by the static nature of single-frame inputs, which fail to capture temporal or multi-angle product views. To enhance the understanding of both spatial and temporal cues inherent in video data, we further explore the video contents including audio transcripts, object locations, and scene-level captions on VideoAVE (Cheng et al., 2025).

Video Understanding Though recent VLMs have extended their capabilities to video data (Wang et al., 2024e; Bai et al., 2025; Lin et al., 2024; Zhang et al., 2023a), these video VLMs still struggle to achieve fine-grained task-specific understanding (Tang et al., 2025). Task-specific video understanding models could be roughly categorized into classification (Rehman and Belhaouari, 2021; Tan et al., 2024b; Han et al.,

2024), captioning (Qasim et al., 2025; Lian et al., 2025; Chen et al., 2024a; Islam et al., 2024; Xu et al., 2024; Kim et al., 2024), and localization approaches (Khosla et al., 2025; Hu et al., 2025b; Feng et al., 2025; Pei et al., 2025; Kumar et al., 2024; Wang et al., 2024a) in diverse areas such as automatic driving (Fang et al., 2024b; Wang et al., 2024c), robotics (McCarthy et al., 2025; Sermanet et al., 2024), science and healthcare (Zhao et al., 2025; Shikino et al., 2025), etc. Building on recent advances in retrieval-augmented video comprehension (Jeong et al., 2025; Zheng et al., 2025; Ren et al., 2025; Mao et al., 2025; Arefeen et al., 2024a,b; Luo et al., 2026), which incorporate contextually relevant knowledge into the interpretation of video content, we leverage visual search to enhance interleaved chain-of-thought (CoT) reasoning for AVE in e-Commerce.

Chain-of-Thought Reasoning Chain-of-thought (CoT) reasoning enables models to decompose a problem into intermediate steps, forming a structured reasoning path to facilitate multi-step inference (Wang et al., 2025b). With the rapid development of VLMs, the integration of CoT reasoning into multimodal domain (MCoT) has gained increasing attention (Zhou et al., 2025; Chen et al., 2024b; Tan et al., 2024a; Zhang et al., 2023c; Zheng et al., 2023). MCoT strategies typically fall into three categories: prompt-based methods (Fei et al., 2024; Wu et al., 2024; Gao et al., 2024; Hu et al., 2025c; Wang et al., 2024d), plan-based methods (Zhang et al., 2025b; Yang et al., 2024b; Zheng et al., 2024), and learning-based methods (Yao et al., 2024; Wang et al., 2024b; Tan et al., 2024a; Zhong et al., 2024). These models often adopt structured pipeline and learn reasoning patterns from training data to guide multimodal reasoning. To enable more comprehensive reasoning from multimodal inputs, recent studies explore the integration of expert tools (i.e. LLM expertise, auxiliary visualization, etc.) (Hu et al., 2025c; Wu et al., 2024; Gao et al., 2024; Li et al., 2024c), and internal or external knowledge (i.e. knowledge graphs, object localization, etc.) (Mitra et al., 2024; Mondal et al., 2024; Khaliq et al., 2024; Jia et al., 2024). However, these MCoT approaches are limited by the fixed general-purpose pipelines designed for visual question answering, which causes inefficiencies and the inclusion of redundant or noisy information for e-Commerce product video understanding.

3 Methodology

We propose ViS-CoT, a training-free visual search augmented reasoning pipeline for large VLMs, which can be integrated into any video-language models. As shown in Figure 2, ViS-CoT consists of four major components: (1) key frame extraction, (2) visual search, (3) interleaved CoT reasoning, and (4) integration and generation.

3.1 Problem Formulation

We consider the task of multimodal product AVE, where the raw input is the product video $\mathcal{V}_p$. Following the two settings described in (Cheng et al., 2025), we consider both the attribute-conditioned and the general formulation. In the attribute-conditioned setting, the model is provided with a predefined list of attributes and is tasked with extracting their corresponding values from the video:

$$f_\theta(\mathcal{V}_p, \{a_1, a_2, \cdots, a_m\}) \rightarrow \{v_1, v_2, \cdots, v_m\} \quad (1)$$

where V_p is the product video, a_i is the pre-defined attribute, and v_i is the extracted value, m denotes the total number of attributes. The general setting is formulated as an open attribute-value pair extraction problem, where no attributes are specified in advance. The model infers both relevant attributes and their associated values:

$$f_\theta(\mathcal{V}_p) = \{(a_1, v_1), (a_2, v_2), \cdots, (a_m, v_m)\} \quad (2)$$

The raw input video $\mathcal{V}_p$ is then refined and enriched with auxiliary information, such as transcripts, metadata retrieved through visual search, and other contextual cues, as described in Sec. 3.2, Sec. 3.3, and Sec. 3.4.

3.2 Key Frame Extraction

Frame Quality Filtering. The first step of our framework automatically filters low-quality frames from product videos. Specifically, we remove frames that are blurry, overly black or white, or exhibit low contrast. The cleaned set of video frames for a product F_p^c is defined as:

$$\begin{aligned} F_p^c = \{ I \in F_p \mid & v_{\text{Lap}}(I) \geq \tau_{\text{blur}} \\ & \wedge\ \tau_{\text{black}} \leq \mu(I) \leq \tau_{\text{white}} \\ & \wedge\ \sigma(I) \geq \tau_{\text{contrast}} \} \end{aligned} \quad (3)$$

where $v_{\text{Lap}}(I)$ is the Laplacian variance measuring the sharpness of the video frame image I, $\mu(I)$ is the mean intensity capturing brightness, and

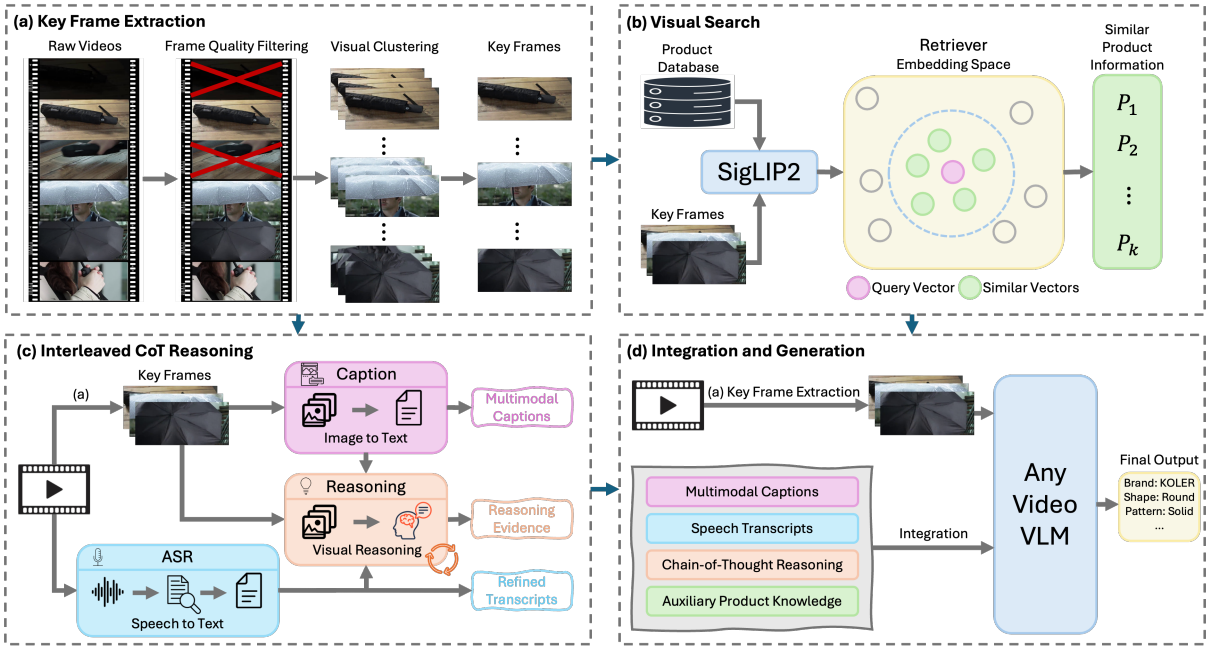


Figure 2: The framework of ViS-CoT, which includes four stages. In stage (a), key frames are extracted from the raw product video through filtering and clustering. In stages (b) and (c), video and the key frames are processed to generate four types of textual information: (i) similar product information via visual search, (ii) image captions, (iii) refined transcripts from ASR, and (iv) interleaved CoT reasoning. In stage (d), all auxiliary textual information from stages (b) and (c) are integrated with the key frames to produce the final answer.

$\sigma(I)$ represents contrast. τ_{blur}, τ_{white}, τ_{black}, and τ_{contrast} are thresholds for these attributes. This step ensures that only visually informative frames are retained, which is critical for preserving the quality and usability of the video data for downstream tasks.

Visual Clustering. After filtering out uninformative frames of a product video, there are still large quantities of redundant or irrelevant content. To reduce redundancy, we introduce a visual clustering step that groups the cleaned video frames F_p^c based on semantic similarity. This enables the model to focus on representative product-related frames from each cluster while discarding repetitive or irrelevant content. Given a set of cleaned video frames $F_p^c = \{f_1^c, f_2^c, \cdots, f_n^c\}$, we first extract visual features using a pre-trained encoder $f_i = E(f_i^c)$. Next, we use K-Means clustering to group into m clusters:

$$\min_{\{C_j\}_{j=1}^m} \sum_{j=1}^{m} \sum_{f_i \in C_j} \|f_i - \mu_j\|^2 \tag{4}$$

where μ_j is the centroid of cluster C_j. From each cluster, we select the frame closest to its centroid as the key frame f_i^{key}. This clustering process partitions the original n frames into m clusters ($m \ll n$), resulting in a non-redundant and diverse set of frames F_p^{key} that efficiently and effectively represent the product.

3.3 Visual Search

We first construct a database from the product images available in the training data. To enable efficient and accurate retrieval during the visual search phase, each training image is encoded into a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ using the visual encoder of SigLIP2 (Tschannen et al., 2025), which is stored alongside its product identifier. During inference, query key video frames F_p^{key} are encoded by the same pretrained SigLIP2 to obtain query embeddings $\{q_1, \cdots, q_m\}$. The cosine similarity between query and database embeddings is then computed, and to ensure robustness across multiple query frames, the similarity scores are averaged. The Top-k nearest neighbors are then retrieved as:

$$\{P_1, \cdots, P_k\} = \arg\operatorname*{top}_k \left(\{\frac{1}{m}\sum_{j=1}^{m} \frac{\mathbf{q}_j^\top \mathbf{x}_i}{\|\mathbf{q}_j\|\,\|\mathbf{x}_i\|} \mid i = 1, \ldots, N\}\right) \tag{5}$$

where P_i denote the retrieved top-k products and their associated attribute values. Finally, the top-k retrieved products form the candidate aspect set for the query product.

3.4 Interleaved Chain-of-Thought Reasoning

We generate the auxiliary texts from the video, followed by an interleaved chain-of-thought reasoning

step, which iteratively refines understanding by alternating between the generated reasoning and the auxiliary context evidence.

Auxiliary Information Generation. To provide richer semantics, context, and evidence for the following reasoning step, we generate two different forms of auxiliary information that are fed into the interleaved CoT reasoning step. (1) *Captions* D_p. To capture semantic details from the key frames F_p^{key} in Sec. 3.2, we leverage a VLM-based captioner to convert visual content into textual descriptions: $D_p = Caption(F_p^{key})$. These captions serve as detailed descriptions of the semantic content from the corresponding key frames. (2) *Transcripts* T_{asr}. Audio signals provide complementary information for video understanding, often providing additional context that may not be available through visual cues alone. For each product video $\mathcal{V}_p$, we first extract the raw audio U and transcribe it into text by an Automatic Speech Recognition (ASR) module, specifically Whisper (Radford et al., 2023). To further enhance conciseness and informativeness, we apply a VLM-based summarizer to the transcribed text, resulting in refined transcripts: $T_{asr} = Summarizer(Whisper(U))$.

Interleaved Chain-of-Thought. Vanilla CoT paradigm typically performs reasoning in one pass using only the input and a fixed instruction (i.e., 'Let's think step by step'). In contrast, our ViS-CoT framework adopts an interleaved reasoning procedure, where the model first performs a visual-only reasoning pass and then refines this reasoning by injecting auxiliary textual knowledge. We begin by generating an initial reasoning path based solely on the video key frames F_p^{key}, the attribute definition A_{def}, and the base prompt P_v:

$$\mathcal{R}'_{ini} = \underset{\mathcal{R}'}{argmax} P(\mathcal{R}' | F_p^{key}, A_{def}, P_v) \quad (6)$$

This initial reasoning $\mathcal{R}'_{ini}$ reflects what the model can infer solely from visual evidence, without external context. Next, we refine this visual-only reasoning by incorporating captions D_p, refined speech transcripts T_{asr}, and retrieved product knowledge $\{P_i\}_{i=1}^k$. Instead of concatenating everything at once, the model revisits and updates its earlier reasoning in light of the new textual cues:

$$\mathcal{R}_p = \arg\max_{\mathcal{R}} P\big(\mathcal{R} \mid \mathcal{R}'_{ini}, A_{def}, P_v, F_p^{key} \oplus D_p \oplus T_{asr} \oplus \{P_i\}_{i=1}^k\big). \quad (7)$$

where $\mathcal{R}_p$ represents the refined reasoning path, in which auxiliary information is incrementally incorporated to resolve ambiguity, enrich attribute interpretation, and ground the reasoning more reliably across modalities.

3.5 Integration and Generation

After collecting various forms of auxiliary textual information, we align and integrate them into a unified input. Specifically, we combine captions D_p, refined transcripts T_{asr}, similar product data $\{P_1, \cdots, P_k\}$, and the interleaved reasoning outcome R_p. These enriched textual inputs are then fused in chronological order, combined with the visual evidence from video key frames F_p^{key}, to jointly encode semantic and relational knowledge. This input is then fed into the video vision language models to generate the final result. The overall process can be formulated as:

$$y = VLM(F_p^{key}, D_p \oplus \{P_1, \cdots, P_k\} \oplus T_{asr} \oplus R_p) \quad (8)$$

where y denotes the final output. This integration preserves complementary multimodal cues within a unified representation, enabling the video VLM to ground its reasoning in richer and more diverse evidence while simultaneously reducing ambiguity.

4 Experiments

4.1 Experimental Setup

Dataset and Baselines. We evaluate ViS-CoT on a public dataset *VideoAVE* (Cheng et al., 2025), which is a video-to-text multi-attribute AVE dataset across 14 different categories with 172 unique attributes in e-Commerce. We evaluate our ViS-CoT in five open-source Video-Language Models, including Video-LLaVA (Video-LLaVA-7B) (Lin et al., 2024), VideoLLaMA3 (VideoLLaMA3-7B) (Zhang et al., 2025a), InternVideo2.5 (InternVideo2_5_Chat_8B) (Wang et al., 2025c), InternVL3.5 (InternVL3_5-8B-Instruct) (Wang et al., 2025a), and Qwen2.5-VL (Qwen2.5-VL-7B) (Bai et al., 2025).

Evaluation Metrics. We follow the evaluation criteria of (Cheng et al., 2025) to evaluate

Table 1: Category-level fuzzy F1 (%) under the attribute-conditioned and generalized settings. Best results are highlighted in bold, and models with a light blue background denote the base model augmented with the ViS-CoT.

Model	Appl.	Arts	Auto	Baby	Beauty	Clothes	Groc.	Indus.	Music	Patio	Pet	Phones	Sports	Toys	Gains
Attribute-Conditioned Setting															
VideoLLaVA	15.96	11.68	13.26	12.53	10.97	19.73	10.16	14.04	15.16	12.93	15.09	13.02	14.56	15.26	-
VideoLLaVA + ViS-CoT	29.27	23.58	21.30	33.97	25.54	42.31	31.84	29.56	32.42	28.97	23.17	25.16	29.64	29.93	↑15.17
VideoLLaMA3	33.33	28.72	29.22	33.44	29.20	35.93	25.97	31.35	31.84	29.56	29.49	30.34	31.89	33.10	-
VideoLLaMA3 + ViS-CoT	41.80	33.62	36.22	48.19	40.98	55.11	44.25	37.19	42.16	42.66	41.92	35.63	36.06	37.94	↑10.03
InternVideo2.5	39.22	31.67	32.22	35.22	32.01	34.93	26.87	35.24	37.46	32.18	30.82	35.17	33.61	33.94	-
InternVideo2.5 + ViS-CoT	57.49	43.81	49.43	58.12	47.00	55.16	38.37	54.23	58.21	56.41	46.77	46.58	57.91	52.73	↑17.98
InternVL3.5	26.79	27.41	28.18	33.91	29.02	26.88	23.03	30.37	36.66	28.11	28.96	32.89	31.69	32.14	-
InternVL3.5 + ViS-CoT	58.49	**51.12**	**52.40**	62.50	**55.56**	62.73	55.40	**57.79**	**63.75**	58.60	**53.80**	**54.11**	63.15	**59.06**	**↑28.03**
Qwen2.5-VL	39.18	31.58	32.40	37.63	33.23	42.51	31.01	34.27	37.75	33.08	30.88	35.19	36.40	36.86	-
Qwen2.5-VL + ViS-CoT	**63.24**	49.07	52.10	**63.64**	52.88	**67.90**	**56.54**	51.85	62.12	**61.32**	52.99	50.26	**66.87**	57.89	↑22.62
Generalized Setting															
VideoLLaVA	5.65	5.36	6.47	6.95	4.69	9.13	3.27	6.48	6.09	6.27	4.70	5.81	7.21	5.76	-
VideoLLaVA + ViS-CoT	20.02	18.79	17.88	21.06	14.52	32.70	11.52	23.62	25.27	22.42	15.36	22.79	25.57	18.86	↑14.75
VideoLLaMA3	9.23	8.70	8.14	11.04	8.03	12.35	4.75	8.96	8.33	9.18	7.84	10.06	10.59	8.30	-
VideoLLaMA3 + ViS-CoT	27.36	24.18	**23.95**	**32.45**	**27.09**	**37.58**	**27.41**	29.13	29.75	**28.25**	**24.35**	**26.15**	28.80	24.55	**↑18.96**
InternVideo2.5	11.95	12.32	10.16	14.76	11.37	16.40	8.22	11.89	9.93	10.73	9.07	14.51	13.29	9.73	-
InternVideo2.5 + ViS-CoT	**28.59**	**24.52**	23.49	30.71	25.72	36.55	23.66	**29.37**	**32.64**	27.90	23.74	25.85	28.68	**25.85**	↑15.92
InternVL3.5	7.05	8.43	6.25	9.58	6.49	11.35	3.59	7.71	8.02	7.05	5.20	9.70	8.54	6.94	-
InternVL3.5 + ViS-CoT	23.99	20.66	20.35	27.51	22.30	30.16	21.11	24.99	30.02	24.60	22.14	22.28	28.36	21.70	↑16.73
Qwen2.5-VL	4.66	7.24	5.61	9.27	6.47	10.79	6.15	5.64	8.33	6.21	5.05	10.06	7.82	5.65	-
Qwen2.5-VL + ViS-CoT	26.68	21.50	20.76	30.06	25.29	33.23	24.97	24.66	30.55	24.76	22.66	25.32	**29.27**	24.22	↑18.93

VideoAVE from two perspectives: (1) attribute-conditioned value prediction, where a predefined list of attributes is provided, and (2) open attribute-value pair extraction, where no attributes are given in advance. Consistent with prior works (Zhang et al., 2023b; Gong et al., 2025a; Cheng et al., 2025), we adopt the fuzzy F1 score as our evaluation metric. We report results at two levels: category-level, where fuzzy F1 is aggregated across all attribute-value predictions within each product category, and attribute-level, where fuzzy F1 is computed separately for each attribute.

Implementation Details. We performed all experiments on NVIDIA A100 GPUs with 80GB memory. All the base models take the default configurations without additional modification. For ViS-CoT, frame quality filtering is applied with thresholds set to $\tau_{\text{blur}} = 100$, $\tau_{\text{black}} = 10$, $\tau_{\text{white}} = 245$, and $\tau_{\text{contrast}} = 15$. In visual clustering, the number of clusters is fixed to $m = 5$, while the top-k retrieved items in the visual search stage is set to $k = 10$. The influence of both m and k is analyzed in Sec. 4.2.4. For the interleaved CoT, the captioning module ($Caption$) uses Qwen2.5-VL while the summarization module ($Summarizer$) and the attribute definition generator (A_{def}) are based on GPT-4.1. The final prompt provided to any video VLM in the attribute-conditioned setting follows the template: *...Extract the values for these attributes:* *attributes* *in the images given the following additional context:* *Captions* D_p, *Knowledge* $P_1, \cdots, P_k$, *Transcripts* T_{asr}, *Reasoning* R_p. *Choose one specific value for each attribute based on your analysis....* For the general setting, the instruction "*Extract the values...*" is replaced with "*Given the following rules:* *rules*, *generate the most relevant attributes based on what you can observe.*"

Additional analyses are provided in the **Appendix**, including latency evaluation (Appendix A), ablations (Appendix B), single-pass vs. interleaved comparison (Appendix C), label-leakage robustness (Appendix D), and open-source reproducibility (Appendix E).

4.2 Results and Discussions

4.2.1 Main Results

Category-Level Results. Table 1 reports category-level performance. We observe that (1) Integrating ViS-CoT consistently improves all base VLMs across both attribute-conditioned and generalized settings, highlighting ViS-CoT's effectiveness and generality as a plug-and-play module that boosts SOTA video-to-text AVE without fine-tuning. (2) ViS-CoT is particularly effective in categories that are visually complex and semantically rich, suggesting that the integration of visual clustering, retrieved product context, and CoT reasoning is especially valuable when raw videos alone are insufficient for precise extraction. (3) Performance under the attribute-conditioned setting is generally higher where the attributes are provided. We notice that even in the more challenging generalized setting, ViS-CoT also demonstrates consistent improvements, highlighting the robustness of ViS-CoT in open-world scenarios where attributes should be discovered automatically.

Attribute-Level Results. Table 2 reports attribute-level performance. Due to space limitations, we present the 10 most frequent attribute results. We

Table 2: Attribute-level fuzzy F1 (%) results. Top 10 frequent attributes are reported. Best results are highlighted in bold, and models with a light green background denote the base model augmented with the plug-and-play ViS-CoT.

Model	Brand	Color	Material	Country of Origin	Item Form	Power Source	Shape	Suggested Users	Pattern	Finish Type
VideoLLaVA	6.49	19.62	15.12	15.61	13.02	2.31	22.89	2.28	21.85	10.68
VideoLLaVA + ViS-CoT	28.67 ↑22.18	22.44 ↑2.82	31.02 ↑15.90	21.34 ↑5.73	33.19 ↑20.17	41.64 ↑39.33	31.31 ↑8.42	42.55 ↑40.27	29.38 ↑7.53	16.65 ↑5.97
VideoLLaMA3	43.27	32.76	28.34	29.68	19.82	8.82	34.70	1.68	27.64	14.52
VideoLLaMA3 + ViS-CoT	34.90 ↓8.37	28.84 ↓3.92	37.08 ↑8.74	47.06 ↑17.38	44.03 ↑24.21	58.83 ↑50.01	56.45 ↑21.75	45.57 ↑43.89	56.25 ↑28.61	32.42 ↑17.90
InternVideo2.5	42.41	34.95	28.37	43.78	19.45	8.03	29.95	1.58	44.13	17.47
InternVideo2.5 + ViS-CoT	42.12 ↓0.29	39.71 ↑4.76	51.33 ↑22.96	**75.21** ↑31.43	47.90 ↑28.45	70.69 ↑62.66	63.08 ↑33.13	80.21 ↑78.63	68.94 ↑24.81	40.02 ↑22.55
InternVL3.5	46.91	33.13	27.77	36.87	14.63	13.87	37.87	2.10	48.25	14.43
InternVL3.5 + ViS-CoT	**50.40** ↑3.49	**43.43** ↑10.30	56.45 ↑28.68	74.60 ↑37.73	**57.24** ↑42.61	**74.29** ↑60.42	**69.46** ↑31.59	**86.64** ↑84.54	**74.40** ↑26.15	**47.27** ↑32.84
Qwen2.5-VL	46.89	34.42	33.69	34.90	25.34	11.96	35.71	1.87	46.16	16.27
Qwen2.5-VL + ViS-CoT	48.34 ↑1.45	39.54 ↑5.12	**59.37** ↑25.68	73.15 ↑38.25	53.50 ↑28.16	73.55 ↑61.59	68.77 ↑33.06	79.87 ↑78.00	66.56 ↑20.40	42.14 ↑25.87

observe: (1) ViS-CoT consistently improves performance across fine-grained visual and semantic attributes. For example, 'power source' may only be inferred from some specific usage scenes, while 'item form' may be revealed during product interactions rather than from static packaging. Such attributes require reasoning over complementary evidence, which ViS-CoT enables through interleaved visual-textual cues. (2) Some attributes, such as 'country of origin' and 'suggested users' are often ambiguous or absent from visual appearance. In these cases, retrieved similar product information significantly improves the performance, confirming that the retrieved knowledge is essential for filling gaps not inferable from video alone. (3) For some relatively 'easy' attributes such as 'brand' and 'color', ViS-CoT may cause a negative effect. We think that these attributes are readily identifiable from almost any frame. The added complexity of reasoning can introduce noise or hallucinations. Simplified strategies can be explored in future work to balance reasoning depth with attribute-specific difficulty.

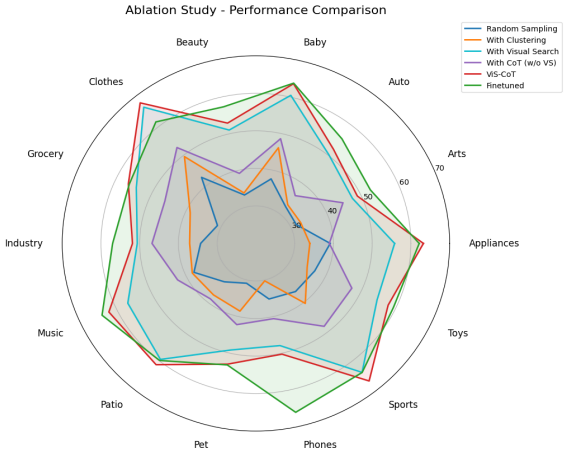


Figure 3: Performance comparison (micro-F1%) of frames with random sampling, key frame clustering, visual search augmentation, CoT reasoning, ViS-CoT with search-enhanced interleaved CoT, and a full-size finetuned base model using training data.

4.2.2 Ablation Study

Our ablation study isolates the effect of each key module and quantifies its impact on overall end-to-end performance. We also compare the performance against a full-size finetuned VLM trained on the whole training data. All experiments are conducted using the same Qwen2.5-VL-7B-Instruct base model for fair comparison. From Figure 3, we observe: (1) Clustering proves more effective than random sampling for frame selection. Selling videos often include background elements or non-relevant items, and visual clustering emphasizes the most representative product-centric frames, reducing redundant and noisy information. (2) Visual search augmentation in ViS-CoT outperforms purely text-based reasoning. By retrieving visually similar products, the model gains access to items that often share key attributes with the query product. This retrieved information can further correct errors that may arise in subsequent interleaved CoT step. (3) By leveraging visual cues to guide and refine the CoT, ViS-CoT achieves the best performance across almost all categories. This training-free framework achieves competitive performance compared with the full-size finetuned VLM, highlighting the promise of retrieval-augmented reasoning as a scalable and efficient paradigm for video-to-text AVE in e-Commerce. (4) A single-pass variant concatenating all auxiliary signals reaches only 46.5% F1, 11.9 pp below full ViS-CoT and even below visual search alone, confirming the staged design is essential (Appendix C). (5) On a clean test subset removing all training-overlapping products, visual search retains a +13.3 pp gain (vs. +14.9 pp on the full set), confirming gains stem from genuine visual similarity, not label leakage (Appendix D).

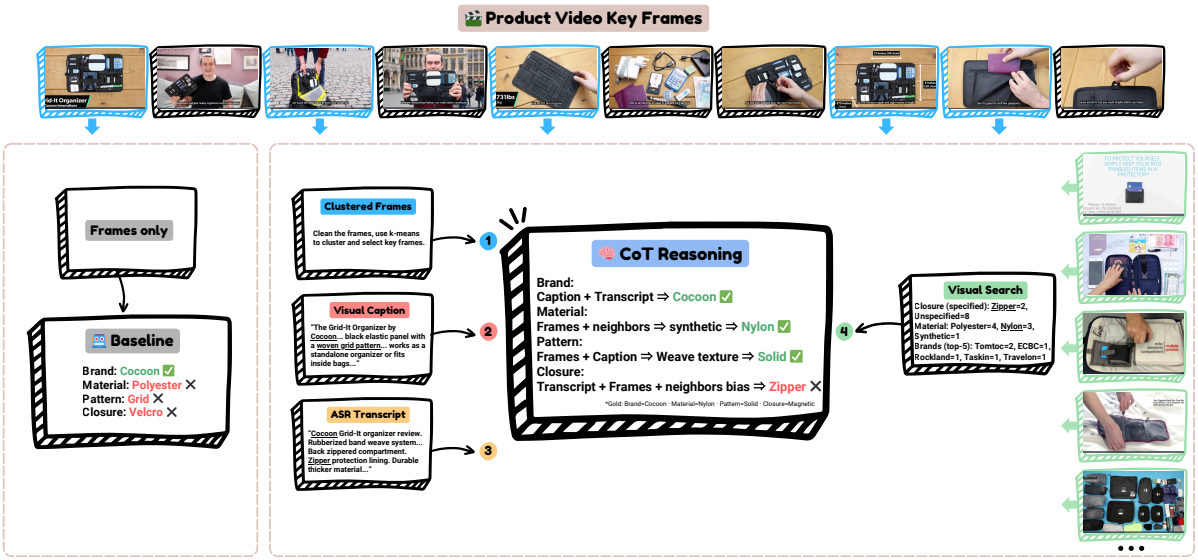


Figure 4: **Case study on a Cocoon organizer sleeve.** Top: product video key frames. **black** border: original extracted frames; **blue** border: clustered/selected key frames. **Left panel (Baseline):** uniformly sampled video frames, without key-frame selection. **Right panel (ViS-CoT):** (1) clustered & cleaned key frames, (2) visual caption from the selected frames, (3) ASR transcript (speaker mentions brand, functions, and details), (4) visual-search neighbors providing attribute priors; Fusing (1)–(4) corrects Material→Nylon and Pattern→Solid; Closure remains biased to Zipper (gold: Magnetic).

4.2.3 Case Study

To illustrate how ViS-CoT improves AVE, we analyze a Cocoon organizer sleeve (Figure 4). The *baseline* consumes only video frames and gets only 'Brand: Cocoon' correct, while misclassifying Material as Polyester, Pattern as Grid, and Closure as Velcro. In contrast, *ViS-CoT* augments the clustered key frames with caption, ASR transcript, and top-k visual-search products, then reasons over these signals. Close-ups of the weave with retrieved similar information corrects it to *Nylon*; the 'textured' cue in captions disambiguates weave texture from a printed pattern, yielding *Solid*. Transcript phrases ('back *zippered* compartment,' '*zipper* protection lining') still pull the chain toward *Zipper*, whereas the video's main flap is magnetic, an informative, auditable residual error.

4.2.4 Parameter Sensitivity

Effect of different cluster number m. To analyze the effect of the number of clusters m, we select four representative categories that vary in dataset size, domain characteristics, and clustering sensitivity. We evaluate $m \in \{5, 10, 15, 20\}$ in the attribute-conditioned setting. As shown in Figure 6, increasing m generally leads to a slight decline in performance. This suggests that using fewer clusters can yield better results, as selecting a moderate number (i.e. $m = 5$) helps the model focus on semantically coherent groups of attributes, balancing granularity and stability. While finer clustering can capture more detailed information, excessive cluster numbers may introduce irrelevant product information, causing over-fragmentation and reducing overall performance.

Effect of different number of top-k. To explore top-k retrieval effect in the visual search component, we evaluate $k \in \{1, 5, 10, 20\}$ in the attribute-conditioned setting, with results shown in Figure 5. As illustrated, $k = 10$ provides the best overall performance, achieving the highest micro-F1 in most categories. Though $k = 20$ shows slightly better performance in categories of Arts, Grocery, and Pet, it often introduces redundant or noisy information that negatively impacts other categories and increases latency during inference. Thus, we adopt $k = 10$ in our final configuration, which best balances effectiveness and efficiency.

4.2.5 Error Analysis

To better understand the limitations of ViS-CoT, we analyze prediction errors produced by the full system across seven representative categories.

Failure Mode Taxonomy. Based on manual inspection of incorrect predictions, we group errors into three categories:

- **Type I: Visually Ambiguous** (28.6%) — At-

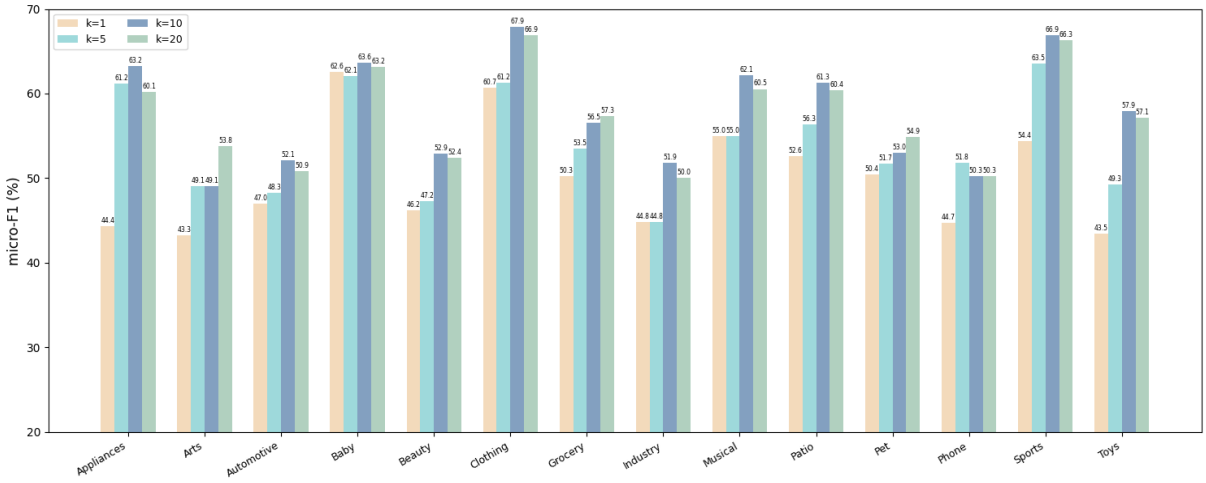


Figure 5: Performance with different number of top-k in visual search when using Qwen2.5-VL-7B as the base.

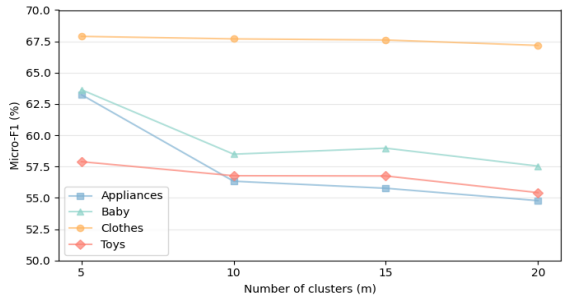


Figure 6: Performance with different cluster number m in visual clustering on four categories of VideoAVE.

tributes that are difficult to determine from the available video frames (e.g., *material*, *country of origin*, *item form*). In these cases, the visual evidence is insufficient to uniquely identify the correct value.

- **Type II: Knowledge Dependent** (28.7%) — Attributes that require external or domain-specific knowledge not present in the video or among retrieved neighbors (e.g., niche *brand* identities or *compatible devices*).
- **Type III: Hallucination** (42.7%) — The model generates a plausible but incorrect value that is not supported by the available visual or textual evidence. These errors typically occur when multiple candidate values are possible.

The distribution of failure modes varies across categories. Hallucination errors are the most frequent type overall, accounting for approximately 37–50% of failures across the seven evaluated categories. Knowledge-dependent errors occur more frequently in Pet (57%) and Toys (48%), which contain attributes requiring specialized product knowledge (e.g., *compatible device type*, *assembly required*). Visually ambiguous errors appear more often in Patio (30%) and Grocery (29%), where attributes such as *material grade* or *item form* are difficult to determine from visual cues alone.

These observations suggest several possible directions for improving performance, including incorporating additional structured product knowledge sources, adapting reasoning depth to attribute difficulty, and improving confidence calibration to reduce unsupported predictions.

5 Conclusion

We present ViS-CoT for video-to-text e-Commerce product AVE. ViS-CoT leverages visual clustering to extract key frames from the video and integrates the retrieved similar products data, captions, transcripts, and interleaved CoT reasoning to enhance video understanding. Unlike fine-tuning VLMs, which is resource-intensive and time-consuming, ViS-CoT provides a resource-efficient, plug-and-play solution that can be seamlessly applied to any open-source video VLMs. ViS-CoT shows strong performance gains across five SOTA video VLMs under both attribute-conditioned and open AVE settings.

6 Limitations

While ViS-CoT demonstrates substantial improvements, several limitations remain. First, although the multi-stage ViS-CoT pipeline yields an average micro-F1 gain of 17.91 percentage points, its iterative search and reasoning procedure introduces additional inference latency. This may hinder scalability in real-time or large-scale production en-

vironments. The inference latency is described and analyzed in the Appendix (Appendix A). Second, ViS-CoT is specifically designed for product-focused video attribute value extraction, where structured aspect accuracy is prioritized. Consequently, we do not include comparisons with chain-of-thought approaches developed for broader video understanding tasks (e.g., storytelling, captioning, summarization). Such evaluations would require qualitatively different objectives and metrics and thus fall outside the scope of this work. Third, our empirical evaluation is focused on video-to-text e-Commerce AVE, and VideoAVE (Cheng et al., 2025), to the best of our knowledge, is the only publicly available benchmark directly aligned with this task. While this scope matches the problem targeted by ViS-CoT, future work should evaluate the framework on additional product-video datasets once comparable public resources become available. We mainly evaluate open-source video VLMs to support reproducibility, controllability, and cost-efficient large-scale experimentation. Although commercial multimodal systems such as GPT and Gemini are important future baselines, systematic evaluation with them is constrained by API cost, limited transparency, and changing model versions. Moreover, our Appendix E shows that replacing GPT-4.1-based helper modules with open-source Qwen2.5-7B yields comparable performance, further supporting the practicality of the open-source setting.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024a. irag: Advancing rag for videos with an incremental approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4341–4348.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024b. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2266–2274. IEEE.

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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ansel Blume, Nasser Zalmout, Heng Ji, and Xian Li. 2023. Generative models for product attribute extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 575–585.

Alexander Brinkmann and Christian Bizer. 2025. Self-refinement strategies for llm-based product attribute value extraction. *arXiv preprint arXiv:2501.01237*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, and 1 others. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Ming Cheng, Tong Wu, Jiazhen Hu, Jiaying Gong, and Hoda Eldardiry. 2025. Videoave: A multi-attribute video-to-text attribute value extraction dataset and benchmark models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6340–6345, New York, NY, USA. Association for Computing Machinery.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yang Deng, Wenxuan Zhang, Qian Yu, and Wai Lam. 2023. Product question answering in E-commerce: A survey. In *Proceedings of the 61st Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 11951–11964, Toronto, Canada. Association for Computational Linguistics.

Chenhao Fang, Xiaohan Li, Zezhong Fan, Jianpeng Xu, Kaushiki Nag, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2024a. Llm-ensemble: Optimal large language model ensemble method for e-commerce product attribute value extraction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2910–2914.

Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. 2024b. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22030–22040.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Yisen Feng, Haoyu Zhang, Meng Liu, Weili Guan, and Liqiang Nie. 2025. Object-shot enhanced grounding network for egocentric video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24190–24200.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and 1 others. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

Jiaying Gong, Wei-Te Chen, and Hoda Eldardiry. 2023. Knowledge-enhanced multi-label few-shot product attribute-value extraction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 3902–3907, New York, NY, USA. Association for Computing Machinery.

Jiaying Gong, Ming Cheng, Hongda Shen, Pierre-Yves Vandenbussche, Janet Jenq, and Hoda Eldardiry. 2025a. Visual zero-shot E-commerce product attribute value extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 460–469, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiaying Gong and Hoda Eldardiry. 2024. Multi-label zero-shot product attribute-value extraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2259–2270, New York, NY, USA. Association for Computing Machinery.

Jiaying Gong, Hongda Shen, and Janet Jenq. 2025b. MICE: Mixture of image captioning experts augmented e-commerce product attribute value extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1151–1160, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. 2024. Video recognition in portrait mode. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21831–21841.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Pollawat Hongwimol, Dong Sheng, Li Zhang, Kai Liu, and Xiufei Wang. 2025. Gavel: Generative attribute-value extraction using llms on llm-augmented datasets. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 81–90.

Jiazhen Hu, Jiaying Gong, Hongda Shen, and Hoda Eldardiry. 2025a. Hypergraph-based zero-shot multimodal product attribute value extraction. In *Proceedings of the ACM on Web Conference 2025*, pages 4853–4862.

Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. 2025b. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3581–3589.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2025c. Visual sketchpad: sketching as a visual chain of thought for multimodal language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics:*

ACL 2025, pages 21278–21298, Vienna, Austria. Association for Computational Linguistics.

Zixi Jia, Jiqiang Liu, Hexiao Li, Qinghua Liu, and Hongbin Gao. 2024. Dcot: Dual chain-of-thought prompting for large multimodal models. In *The 16th Asian Conference on Machine Learning (Conference Track)*.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296, Miami, Florida, USA. Association for Computational Linguistics.

Anant Khandelwal, Happy Mittal, Shreyas Kulkarni, and Deepak Gupta. 2023. Large scale generative multimodal attribute extraction for E-commerce attributes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 305–312, Toronto, Canada. Association for Computational Linguistics.

Savya Khosla, Alexander Schwing, Derek Hoiem, and 1 others. 2025. Relocate: A simple training-free baseline for visual query localization using region-based representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3706.

Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. 2024. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904.

Yogesh Kumar, Saswat Mallick, Anand Mishra, Sowmya Rasipuram, Anutosh Maitra, and Roshni Ramnani. 2024. Qdetrv: query-guided detr for one-shot object localization in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2831–2839.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024b. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. 2024c. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1915–1929, Miami, Florida, USA. Association for Computational Linguistics.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, and 1 others. 2025. Describe anything: Detailed localized image and video captioning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21766–21777. IEEE.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and 1 others. 2026. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033.

Mingyang Mao, Mariela M Perez-Cabarcas, Utteja Kallakuri, Nicholas R Waytowich, Xiaomin Lin, and Tinoosh Mohsenin. 2025. Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding. *arXiv preprint arXiv:2505.23990*.

Robert McCarthy, Daniel CH Tan, Dominik Schmidt, Fernando Acero, Nathan Herr, Yilun Du, Thomas G Thuruthel, and Zhibin Li. 2025. Towards generalist robot learning from internet video: A survey. *Journal of Artificial Intelligence Research*, 83.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806.

Pengfei Pei, Yun Cao, Jinchuan Li, Zeyu Zhang, and Yuqi Pang. 2025. Object-based video tampering localization via trace consistency analysis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Iqra Qasim, Alexander Horsch, and Dilip Prasad. 2025. Dense video captioning: A survey of techniques, datasets and evaluation protocols. *ACM Computing Surveys*, 57(6):1–36.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023.

Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Atiq Rehman and Samir Brahim Belhaouari. 2021. Deep learning for video classification: A review. *TechRxiv preprint*.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, and 1 others. 2024. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE.

Kiyoshi Shikino, Kazuyo Yamauchi, Nobuyuki Araki, Ikuo Shimizu, Hajime Kasai, Tomoko Tsukamoto, Hiroshi Tajima, Yu Li, Misaki Onodera, and Shoichi Ito. 2025. Understanding community health care through problem-based learning with real-patient videos: Single-arm pre-post mixed methods study. *JMIR Medical Education*, 11:e68743.

Keiji Shinzato, Naoki Yoshinaga, Yandi Xia, and Wei-Te Chen. 2023. A unified generative approach to product attribute-value identification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6599–6612, Toronto, Canada. Association for Computational Linguistics.

Yindu Su, Huike Zou, Lin Sun, Ting Zhang, Haiyang Yang, Chen Li Yu, David Lo, Qingheng Zhang, Shuguang Han, and Jufeng Chen. 2025. Taclr: A scalable and efficient retrieval-based method for industrial product attribute value identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31526–31538.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. 2024a. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pages 305–322. Springer.

Weijun Tan, Qi Yao, and Jingfeng Liu. 2024b. Overlooked video classification in weakly supervised video anomaly detection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 212–220. IEEE.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Ameni Trabelsi, Maria Zontak, Yiming Qian, Brian Jackson, Suleiman Khan, and Umit Batur. 2025. What matters when building vision language models for product image analysis? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1282–1291. IEEE.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18209–18220.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024b. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. 2024c. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024d. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024e. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang.

2025c. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025d. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. 2024. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pages 164–182. Springer.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. 2024. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jun Cheng Yang, Zuchao Li, Shuai Xie, Wei Yu, Shijun Li, and Bo Du. 2024b. Soft-prompting with graph-of-thought for multi-modal representation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15024–15036, Torino, Italia. ELRA and ICCL.

Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. Mixpave: Mix-prompt tuning for few-shot product attribute value extraction. In *Findings of the association for computational linguistics: ACL 2023*, pages 9978–9991.

Junyi Yao, Yijiang Liu, Zhen Dong, Mingfei Guo, Helan Hu, Kurt Keutzer, Li Du, Daquan Zhou, and Shanghang Zhang. 2024. Promptcot: Align prompt distribution via adapted chain-of-thought. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7027–7037.

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. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Ziyu Guo, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Peng Gao, and Hongsheng Li. 2025b. Let's verify and reinforce image generation step by step. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 28662–28672.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023b. Pay attention to implicit attribute values: A multi-modal generative framework for AVE task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023c. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8475–8489.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. 2024. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 419–428.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

Xiongtao Zhou, Jie He, Lanyu Chen, Jingyu Li, Haojing Chen, Victor Gutierrez Basulto, Jeff Z. Pan, and Hanjie Chen. 2025. MiCEval: Unveiling multimodal chain of thought's quality via image description and reasoning steps. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10002–10039, Albuquerque, New Mexico. Association for Computational Linguistics.

Henry Peng Zou, Vinay Samuel, Yue Zhou, Weizhi Zhang, Liancheng Fang, Zihe Song, Philip S. Yu, and Cornelia Caragea. 2024a. ImplicitAVE: An open-source dataset and multimodal LLMs benchmark for implicit attribute value extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 338–354, Bangkok, Thailand. Association for Computational Linguistics.

Henry Peng Zou, Gavin Heqing Yu, Ziwei Fan, Dan Bu, Han Liu, Peng Dai, Dongmei Jia, and Cornelia Caragea. 2024b. EIVEN: Efficient implicit attribute value extraction using multimodal LLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 453–463, Mexico City, Mexico. Association for Computational Linguistics.

A Inference Latency Analysis

Component	Latency (s)	%
Keyframe clustering (subtotal)	2.40	48.1
SigLIP embedding	2.24	(44.9)
K-means	0.16	(3.2)
Visual search (FAISS)	0.11	2.2
CoT Stage 1 (vision hypothesis)	0.97	19.4
CoT Stage 2 (grounded extraction)	1.51	30.3
Full ViS-CoT total	**4.99**	**100.0**
Baseline VLM (reference)	*1.31*	—

Table 3: Inference latency breakdown of ViS-CoT on a representative toy product. Indented gray rows break down the keyframe-clustering subtotal. Keyframe clustering is a one-time preprocessing step per video. Excluding it, per-query inference requires only 2.59 s vs. 1.31 s for the baseline.

Condition	F1 (%)	ΔF1	s/product	F1 eff.
C0 Baseline	33.4	—	1.6	—
C1 +Visual Search	48.3	+14.9	2.8	12.3 pp/s
C2 +CoT (w/o VS)	33.5	+0.1	3.2	0.1 pp/s
C3 Full ViS-CoT	58.5	+25.1	3.3	**14.3 pp/s**

Table 4: Per-condition inference time vs. F1 gain. F1 efficiency = ΔF1 per additional second over C0. Full ViS-CoT delivers the best return on compute: each added second yields 14.3 pp of F1 improvement.

We measure per-sample inference latency on a single GPU using Qwen2.5-VL-7B. Video downloading and frame extraction are excluded from all measurements. As shown in Table 3, the full ViS-CoT pipeline processes one product in 4.99 s, with component latencies summing to 100%.

Keyframe clustering (SigLIP embedding followed by K-means) accounts for 2.40 s (48.2%) of the total runtime. However, this step is executed only once per product video regardless of the number of attribute queries, and therefore acts as a one-time preprocessing cost rather than a per-query inference overhead. Excluding this preprocessing stage, per-query inference requires 2.59 s, representing a +1.28 s increase over the 1.31 s baseline.

Visual search implemented with FAISS introduces negligible overhead (0.11 s, 2.2%). Table 4 further reports the trade-off between latency and performance across ablation settings. The full ViS-CoT system achieves the highest efficiency, yielding 14.3 percentage points of F1 improvement per additional second of inference. In contrast, CoT alone (C2) provides only 0.1 pp/s, indicating that retrieval-based visual grounding substantially improves the effectiveness of the reasoning process. At query time (i.e., after one-time keyframe clustering), end-to-end throughput on a single A100 GPU is approximately **0.30 products/sec/GPU**, acceptable for offline catalog enrichment and trivially scalable via data parallelism.

B Ablation Study: Component Contributions

Table 5 presents a systematic evaluation of each component of ViS-CoT using Qwen2.5-VL-7B as the base model across seven representative product categories. We consider four incremental conditions: **C0** (baseline: images + attributes only, without auxiliary context or CoT reasoning), **C1** (C0 + Visual Search), **C2** (C0 + interleaved CoT only, without retrieval), and **C3** (Full ViS-CoT with all components). To ensure a controlled and fair comparison, all conditions use the same inference framework with identical prompting and image preprocessing. Absolute F1 values are anchored to the baseline and full-system performance reported in Table 1; the relative improvements within each table isolate individual component contributions.

Category	C0	C1 (+VS)	C2 (+CoT)	Full
Arts	31.6	43.1$_{\uparrow 11.5}$	31.7$_{\uparrow 0.1}$	49.1$_{\uparrow 17.5}$
Baby	37.6	57.2$_{\uparrow 19.6}$	39.6$_{\uparrow 2.0}$	63.6$_{\uparrow 26.0}$
Clothing	32.4	47.3$_{\uparrow 14.9}$	33.3$_{\uparrow 0.9}$	67.9$_{\uparrow 35.5}$
Grocery	31.0	47.9$_{\uparrow 16.9}$	31.8$_{\uparrow 0.8}$	56.5$_{\uparrow 25.5}$
Patio	33.1	50.0$_{\uparrow 16.9}$	32.3$_{\downarrow 0.8}$	61.3$_{\uparrow 28.2}$
Pet	30.9	40.2$_{\uparrow 9.3}$	28.6$_{\downarrow 2.3}$	53.0$_{\uparrow 22.1}$
Toys	36.9	52.2$_{\uparrow 15.3}$	37.2$_{\uparrow 0.3}$	57.9$_{\uparrow 21.0}$
Avg	**33.4**	**48.3**$_{\uparrow 14.9}$	**33.5**$_{\uparrow 0.1}$	**58.5**$_{\uparrow 25.1}$

Table 5: Component ablation (micro-F1%, ↑/↓ vs. C0). C0: base VLM with images and attributes only; C1: C0 + Visual Search (RAG); C2: C0 + interleaved CoT only, no retrieval; Full (C3): complete ViS-CoT. Subscripts show absolute change from C0.

Table 6 shows the contribution of each modality

by removing it from the full system.

Category	Full	w/o Cap.	w/o Trans.	w/o RAG
Arts	49.1	39.5$_{\downarrow 9.6}$	39.4$_{\downarrow 9.7}$	37.9$_{\downarrow 11.2}$
Baby	63.6	57.5$_{\downarrow 6.1}$	58.0$_{\downarrow 5.6}$	54.5$_{\downarrow 9.1}$
Clothing	67.9	61.7$_{\downarrow 6.2}$	61.4$_{\downarrow 6.5}$	59.8$_{\downarrow 8.1}$
Grocery	56.5	48.1$_{\downarrow 8.4}$	49.2$_{\downarrow 7.3}$	46.3$_{\downarrow 10.2}$
Patio	61.3	54.4$_{\downarrow 6.9}$	54.4$_{\downarrow 6.9}$	51.8$_{\downarrow 9.5}$
Pet	53.0	47.8$_{\downarrow 5.2}$	48.0$_{\downarrow 5.0}$	46.5$_{\downarrow 6.5}$
Toys	57.9	51.0$_{\downarrow 6.9}$	50.6$_{\downarrow 7.3}$	48.4$_{\downarrow 9.5}$
Avg	**58.5**	**51.4**$_{\downarrow 7.1}$	**51.6**$_{\downarrow 6.9}$	**49.3**$_{\downarrow 9.2}$

Table 6: Modality knock-out ablation (micro-F1%, ↓ drop from Full). Cap.=visual captions; Trans.=ASR transcripts; RAG=visual search. All three modalities are necessary; removing any one degrades performance across all categories.

Key findings. *(1) Visual search provides the primary performance improvement.* Adding visual search alone (C1) increases average F1 from 33.4 to 48.3 (+14.9 pp). In contrast, CoT reasoning without retrieval (C2) produces only a marginal change (+0.1 pp on average). This suggests that retrieval-based context plays an important role in grounding the reasoning process.

(2) The full system provides additional gains beyond retrieval alone. Full ViS-CoT (C3) achieves an average F1 of 58.5, corresponding to a +25.1 pp improvement over the baseline and a further +10.2 pp gain beyond visual search alone. This indicates that interleaved CoT reasoning becomes more effective when combined with retrieved contextual evidence.

(3) Performance improvements are consistent across categories. Visual search improves performance for all seven categories, while CoT-only configurations show minimal or slightly negative changes in several cases (e.g., Patio and Pet). These results highlight the importance of retrieval-based grounding for stable reasoning across diverse product domains.

C Single-Pass vs. Interleaved Reasoning

A natural alternative to our two-stage interleaved design is to provide all retrieved and auxiliary signals (captions, ASR transcripts, retrieved neighbor attributes, and attribute definitions) directly in a single prompt, letting the VLM perform attribute extraction in one pass without an explicit Stage-1 visual hypothesis. We refer to this variant as **Single-pass**. It tests whether the staged design is necessary for the system to effectively use the same set of evidence.

We run Single-pass under the same evaluation framework as Appendix B, using Qwen2.5-VL-7B and the same seven representative categories. Table 7 reports the comparison. All conditions are reproduced from Table 5 except the Single-pass row, which is new.

Condition	Avg F1	vs. C1
C0 Baseline	33.4	−14.9
C1 +Visual Search	48.3	—
Single-pass	46.5	−1.8
Full ViS-CoT	**58.5**	+10.2

Table 7: Single-pass vs. interleaved reasoning (micro-F1%, average over seven categories). C0: images only; C1: retrieval → direct extract; Single-pass: retrieval + captions + ASR in one flat prompt (no Stage-1); Full ViS-CoT: same evidence, staged design. Adding all auxiliary evidence in one prompt actually *underperforms* visual search alone by 1.8 pp, while the staged design improves over C1 by +10.2 pp.

Findings. Two observations stand out. *(1) Interleaved reasoning outperforms single-pass by +11.9 pp.* The contrast between Single-pass (46.5%) and Full ViS-CoT (58.5%) isolates the effect of the staged design itself: both rows use the same retrieved evidence, the same captions, and the same ASR transcripts; only the prompting structure differs. The +11.9 pp swing is attributable solely to whether evidence is organized into two reasoning stages or concatenated in one prompt. *(2) Single-pass actually underperforms retrieval-only.* Despite having strictly *more* context than C1 (captions and ASR added on top of retrieval), Single-pass loses 1.8 pp relative to C1. We interpret this as a *context-overload effect*: when many heterogeneous evidence sources compete in one prompt, the model struggles to integrate them and is more easily distracted by visual-similarity-induced noise from retrieved neighbors. The staged design mitigates this by anchoring reasoning on visual evidence first (Stage 1) and then refining it with retrieved and auxiliary context (Stage 2). This experiment provides direct empirical support for framing ViS-CoT's novelty as a staged design choice, not merely a concatenation of existing components.

D Label-Leakage Robustness

Because the visual-search index is constructed from product images in the training split of VideoAVE, a legitimate concern is whether ViS-CoT's gains stem from retrieving *near-duplicate* items that share their attribute labels with the query — effectively label propagation rather than visual

reasoning. We address this concern in three steps.

Step 1: Quantifying overlap. We compute `product_id` overlap between the test and training splits of VideoAVE. Overlap ranges from **19% to 37%** across the 14 categories, reflecting the fact that VideoAVE (and Amazon product data more broadly) draws from a shared marketplace where the same product may appear in both splits.

Step 2: Implementation safeguard. We enforce `skip_self=True` in the retrieval step: any retrieved neighbor whose `product_id` matches the query is dropped before the top-k list is formed. This prevents trivial self-matches even when overlap exists, but does not by itself eliminate the possibility that retrieved *nearby* items share the same attribute labels.

Step 3: Clean-subset re-evaluation. We construct a *clean subset* of test products by removing every test item whose `product_id` appears anywhere in the training split. This yields a strict no-overlap evaluation set (the clean subset retains 63–81% of samples per category). We then re-evaluate C0 (baseline) and C1 (+Visual Search) on this clean subset.

Metric	Full set	Clean subset	Δ
C0 baseline (F1%, raw)	28.0	26.4	−1.6
C1 Visual Search gain	+14.9 pp	+13.3 pp	−1.6 pp

Table 8: Visual search performance on the full evaluation set vs. the clean subset (test products whose `product_id` does *not* appear in training). The +13.3 pp clean-subset gain is only 1.6 pp (≈11% relative) below the +14.9 pp full-set gain, confirming that the improvement is not driven by near-duplicate label leakage.

Findings. As shown in Table 8, the gain from visual search drops by only 1.6 pp (≈11% relative) when all overlapping products are removed. The bulk of the improvement therefore arises from *visually similar but distinct* training products, which the model uses for attribute-level transfer rather than direct label copying.

Two additional mechanisms further support this conclusion. First, the retrieved context is not consumed as a single nearest-neighbor label but rather as an *aggregated frequency distribution* over the top-k neighbors' attribute values, forcing the model to reason over noisy multi-candidate evidence rather than copy a single label. Second, ViS-CoT's two-stage design (Appendix C) ensures that the Stage-1 visual hypothesis is formed *before* retrieved labels are revealed, further mitigating label-anchoring effects in the reasoning trace.

E Open-Source Reproducibility

To evaluate whether ViS-CoT depends on proprietary model components, we replace the GPT-4.1 modules in the original pipeline with the open-source Qwen2.5-7B-Instruct model (Yang et al., 2024a). Specifically, Qwen2.5-7B is used for (1) transcript summarization, which condenses raw ASR transcripts into concise product descriptions, and (2) attribute definition generation (A_{def}), which produces structured attribute descriptions for each product category. We then re-run the full ViS-CoT pipeline using Qwen2.5-7B for both components and evaluate performance on seven representative categories.

Table 9 reports the resulting end-to-end micro-F1 scores. Using Qwen2.5-7B yields an average F1 of 58.3%, compared with 58.5% when using GPT-4.1, corresponding to a difference of −0.2 percentage points. Performance differences across categories are small, and Qwen2.5-7B matches or slightly exceeds GPT-4.1 in three categories (Arts, Pet, and Toys). These results indicate that the performance of ViS-CoT remains stable when replacing proprietary components with open-source models.

Category	GPT-4.1	Qwen2.5-7B
Arts	49.1	$49.3_{\uparrow 0.2}$
Baby	63.6	$63.0_{\downarrow 0.6}$
Clothing	67.9	$67.7_{\downarrow 0.2}$
Grocery	56.5	$55.8_{\downarrow 0.7}$
Patio	61.3	$61.2_{\downarrow 0.1}$
Pet	53.0	$53.1_{\uparrow 0.1}$
Toys	57.9	$57.9_{\uparrow 0.0}$
Avg	**58.5**	**58.3**$_{\downarrow 0.2}$

Table 9: End-to-end ViS-CoT micro-F1 (%) with GPT-4.1 vs. open-source Qwen2.5-7B for transcript summarization and A_{def} generation. Subscripts show change vs. GPT-4.1. The −0.2pp average gap confirms that ViS-CoT's gains do not depend on proprietary APIs.