GROUNDED PRODUCT UNDERSTANDING IN LIVESTREAM VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

E-commerce livestreams have emerged as an important channel for presenting products to online consumers, often featuring multiple products with relevant information distributed across different moments. This poses significant challenges for downstream product understanding applications, such as product-centric livestream clipping, where models need to identify the product and its relevant segments for information gathering. However, existing benchmarks for general product understanding typically evaluate product retrieval and temporal localization in isolation, leaving the critical correspondence between product identity and temporal evidence largely unassessed. To address this limitation, we introduce **GPUB**, a large-scale benchmark comprising 3,000 real-world e-commerce livestream instances with quality-controlled multi-moment temporal annotations and a catalog of over 31K fashion products. GPUB supports three evaluation tasks: given a livestream video and a candidate product set, the main task **Grounded Product Understanding (GPrU)** requires jointly identifying the product being presented and localizing its supporting moments; Product Retrieval and Product Moment Localization serve as two complementary subtasks. Evaluation of existing multimodal models shows that GPrU remains highly challenging, with the best-performing baseline achieving only 10.13 Pair mAP@.3. To narrow the performance gap, we further develop UniPro, a unified product understanding model that derives product-aligned and temporally structured representations from shared multimodal encoding, improving Pair mAP@.3 to 24.58 while achieving 38.81 Joint R@1@.3 on GPrU.

1 INTRODUCTION

Livestream e-commerce enables hosts to showcase and sell products through continuous video streams with real-time viewer interaction. Unlike information-dense product pages or edited short videos, product evidence is distributed across multiple non-contiguous moments, with visual and spoken cues providing complementary information (Fig. 1). This fragmented and multimodal evidence makes establishing reliable correspondence between catalog products and their supporting livestream moments inherently challenging, creating a key bottleneck for downstream applications such as product-centric clipping and summarization.

Existing tasks address product identification or temporal localization from different perspectives, but largely treat them as separate problems. E-commerce video benchmarks evaluate general product understanding through question answering (Liu et al., 2026), while product representation and retrieval methods focus on matching livestreams or short videos to specific products (Bai et al., 2023; Yang et al., 2023; Hu et al., 2024). Product retrieval focuses on resolving which catalog product is presented; however, it operates on temporally trimmed or product-centric video inputs, where the relevant content has already been narrowed to some extent. Temporal grounding, in contrast, explicitly evaluates when relevant evidence occurs (Hendricks et al., 2017; Lei et al., 2021; Liu et al., 2025), yet typically assumes that the target semantics are already specified by a text query, question, or task description. In general, neither setting requires the target product and its supporting moments to be inferred jointly from the same livestream video. As a result, a model may perform well on both aspects independently without establishing the correct correspondence between the identified product and its supporting moments.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

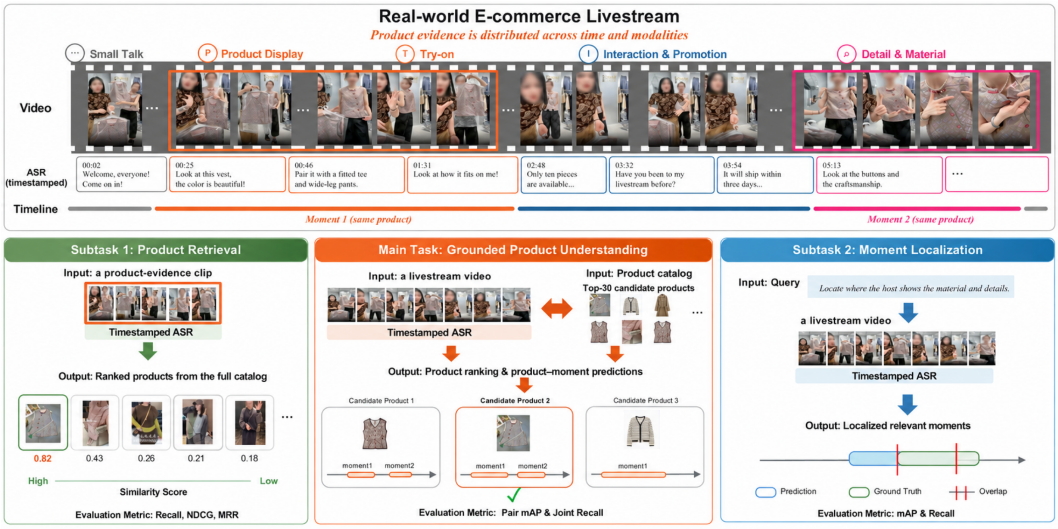


Figure 1: Grounded Product Understanding in e-commerce livestreams. Top: An example e-commerce livestream with product evidence distributed across disjoint moments, conveyed through both visual content and timestamped ASR. Bottom: The three evaluation tasks in GPUB.

To address this gap, we formulate a joint problem as **Grounded Product Understanding (GPrU)**, requiring jointly identifying the target catalog product and localizing its supporting moments from the same livestream video. We further introduce **Grounded Product Understanding Benchmark (GPUB)** for this task, containing 3,000 livestream evaluation windows, 8,975 quality-controlled product evidence moments, and a catalog of 31,831 fine-grained fashion products. GPUB defines Grounded Product Understanding as the main task, alongside Product Retrieval and Product Moment Localization as two complementary subtasks (Fig. 1). These evaluations assess product identification and temporal localization separately, as well as the correspondence between identified products and their supporting moments in GPrU.

GPUB combines three key design features to support these evaluations: (1) a fine-grained fashion product catalog that consists entirely of apparel categories and captures the challenge of distinguishing highly similar fashion products; (2) temporally distributed multimodal evidence annotations that require models to integrate visual and spoken clues across multiple moments; and (3) an identity-preserving joint annotation and evaluation protocol that ensures product identification and temporal localization refer to the same product.

To provide a domain-trained baseline for GPUB, we develop UniPro, a unified model designed to capture both product-level semantics and temporal evidence. Fine-grained product identification emphasizes semantic discrimination between similar items, while evidence localization additionally requires temporal structure. UniPro addresses this through shared multimodal encoding, then derives a product-aligned representation for product identification and temporally structured representations for evidence localization. Our main contributions are as follows:

- We introduce Grounded Product Understanding (GPrU), a new task that requires jointly identifying the target catalog product and localizing its supporting semantic evidence.
- We construct GPUB, a large-scale benchmark to support product understanding in e-commerce livestreams, with multi-moment evidence annotations and a fine-grained fashion product catalog. GPUB evaluates GPrU as its main task, with Product Retrieval and Product Moment Localization serving as two complementary subtasks.
- Extensive evaluation of existing off-the-shelf MLLMs reveals substantial limitations in modeling product–moment correspondence. We further develop UniPro, a unified product understanding model, as a task-trained solid baseline to facilitate future research on GPrU.

2 Related Work

E-commerce and Livestream Product Retrieval. Product retrieval aims to identify corresponding items from a product catalog, often requiring fine-grained discrimination among similar products. Early work primarily studies image–text matching over large product collections (Zhan et al., 2021). With the growing prevalence of video-based shopping, subsequent studies extend product retrieval to short videos and livestreams, addressing challenges such as temporal variation, cross-domain appearance shifts, and background distractors (Zhang et al., 2021; Yang et al., 2023; Hu et al., 2024). However, existing work largely focuses on product matching and exhibition intervals, rather than associating products with moments that convey related product information.

Video Temporal Grounding and Moment Retrieval. Video temporal grounding aims to localize video segments corresponding to a given language query. Existing methods span proposal-based matching, direct boundary prediction, and unified temporal modeling (Hendricks et al., 2017; Zhang et al., 2020; Lei et al., 2021; Moon et al., 2023; Lin et al., 2023). Recent work has further broadened temporal grounding to diverse settings, including long-form videos (Soldan et al., 2022; Tan et al., 2024), grounded question answering (Xiao et al., 2024), task-oriented evidence localization (Liu et al., 2025), and MLLM-based temporal grounding (Zhang et al., 2026; Zheng et al., 2026). Despite these extensions, existing temporal grounding methods typically assume that the target is directly specified by a query, whereas GPrU requires the target product to be inferred from a candidate set.

Long-form Multimodal Video Modeling. Long-form video understanding is constrained by the amount of visual and linguistic context that can be processed effectively. Existing approaches address the visual side through adaptive frame selection, token compression, and memory-based summarization, enabling models to cover longer temporal contexts (Shen et al., 2024; Buch et al., 2025; Hu et al., 2025; Song et al., 2024; He et al., 2024). Complementary work incorporates subtitles or ASR to provide semantic information beyond visual frames, ranging from video–language pretraining (Li et al., 2020; Xu et al., 2021) to streaming multimodal modeling (Chen et al., 2025). These approaches enable more effective long-form video understanding, while product understanding in e-commerce livestreams requires identifying information relevant to specific products from complex visual scenes and semantically noisy spoken content.

3 The GPUB Benchmark

This section presents the task formulations, data construction pipeline, dataset characteristics, and evaluation protocols of GPUB.

3.1 Task Definition

Let $\mathcal{G} = \{p_1, \ldots, p_{N_g}\}$ denote the shared product catalog. A benchmark instance consists of a livestream video V centered on a single target product $p^* \in \mathcal{G}$, its timestamped ASR A, and a set of ground-truth product evidence moments $\mathcal{E}^*$. Each $e = [t_s, t_e] \in \mathcal{E}^*$ is a temporal interval that provides evidence for p^*. GPUB focuses on three tasks: **(1) Product Retrieval (subtask).** Given a product evidence clip (V_m, A_m) cropped from (V, A) and the full catalog $\mathcal{G}$, the model ranks catalog products by their correspondence to the clip. **(2) Product Moment Localization (subtask).** Given the full video (V, A) and a product-related text query q, the model returns confidence-ranked temporal intervals that support the semantics of q. **(3) Grounded Product Understanding (main task).** Given the full instance video (V, A) and a fixed candidate set $\mathcal{C} \subset \mathcal{G}$ of 30 products containing p^*, with no text query provided, the model ranks the candidate products and produces product–moment predictions $\{(p_i, \hat{m}_{ij})\}$, where $p_i \in \mathcal{C}$ and $\hat{m}_{ij} = [\hat{t}_s^{ij}, \hat{t}_e^{ij}]$ denotes the predicted moment corresponding to p_i.

3.2 Dataset Construction

We construct GPUB by progressively establishing the target product and its supporting temporal evidence for each livestream. Specifically, we first curate the product catalog by normalizing product identifiers, merging duplicate SKUs, standardizing product metadata, and retaining one quality-controlled main image for each product, resulting in a catalog of 31,831 products. Building on this

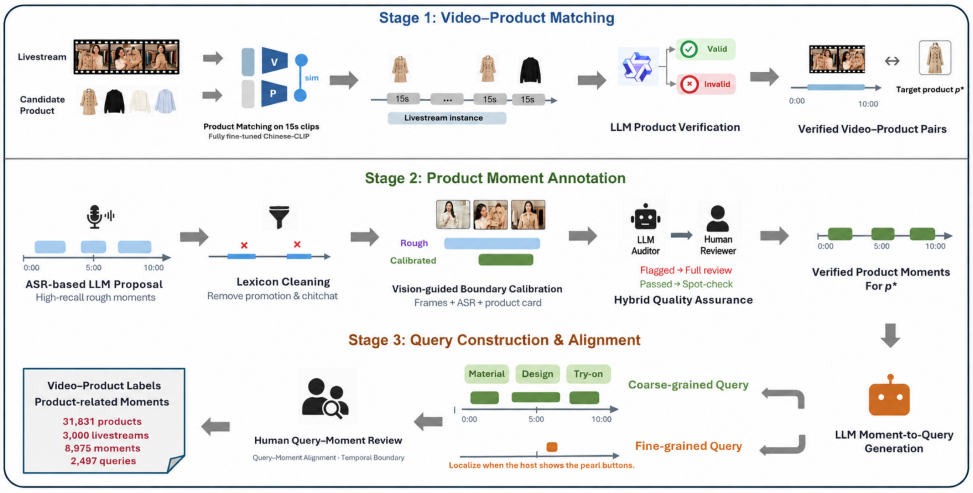


Figure 2: Video-side construction pipeline. **Stage 1** verifies video–product pairs. **Stage 2** annotates the complete set of evidence corresponding to each target product. **Stage 3** derives coarse-grained and fine-grained queries and aligns them with their supporting moments.

catalog, the video-side pipeline comprises three stages (Figure 2): video–product matching, product evidence annotation, and query construction and evidence alignment. The first two stages provide the joint annotations for Grounded Product Understanding, while the third derives textual queries and their query-specific temporal annotations for Product Moment Localization.

Stage 1: Video–product matching. We first determine the specific catalog product being discussed in the livestream instance. A fully SFT-trained Chinese-CLIP model retrieves candidate products from each 15-second clip, after which Qwen3-VL-32B-Instruct verifies the top-ranked candidate using the entire instance video, timestamped ASR, and product information. We retain only instances for which the product match can be confidently verified and discard uncertain cases. The verified product is taken as the target product p^*, which provides the basis for subsequent temporal evidence annotation.

Stage 2: Product evidence annotation. We annotate $\mathcal{E}$, the set of *product evidence moments*: temporally coherent intervals that refer to p and convey visual or spoken information about its attributes or characteristics. Non-contiguous evidence is annotated separately, while intervals without substantive product information, such as those involving promotion, pricing, inventory updates, or casual conversation, are excluded. Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025a), with thinking enabled, first proposes high-recall candidate intervals from timestamped ASR, followed by lexicon-based screening of promotional and conversational noise. The model then validates the remaining candidates using video frames, ASR, and target product information, and calibrates their temporal boundaries against the visual content. GPT-4o (OpenAI, 2024) flags uncertain annotations for full human review, while a random subset of passed instances is fully audited to verify annotation correctness and identify missing evidence.

Stage 3: Query construction and evidence alignment. Given $\mathcal{E}^*$, GPT-4o (OpenAI, 2024) constructs queries and their temporal targets for Product Moment Localization at two semantic granularities. A *coarse-grained query* covers a broad aspect, such as material, design, or try-on experience, and is aligned with the verified moments conveying that information. A *fine-grained query* targets a specific attribute or demonstration, such as fabric texture or lace-paneling details, with boundaries further refined to precisely cover the supporting evidence. Human reviewers then validate each query–evidence pair and refine temporal boundaries when necessary, yielding the final correspondences $(q, \mathcal{E}_q)$.

30-candidate construction. To focus GPrU on joint product identification and temporal grounding rather than large-scale retrieval, we construct a challenging 30-product candidate set for each

Table 1: Comparison of representative benchmark test splits. Counts refer to the evaluation split where reported; – indicates data not reported or not released for that split. † denotes clothing-trajectory queries (Video2Shop), cropped clips (LPR4M), and query–video pairs (QVHighlights), rather than distinct source videos. ASR indicates whether speech transcripts are provided with the evaluation data; QVHighlights releases ASR captions for training videos only.

Benchmark	Domain	Dataset properties				Evaluation coverage			
		Catalog products	Videos	Avg. duration	ASR	Product ID	Temporal loc.	Multi-moment	Joint P–M
Video2Shop (Cheng et al., 2017)	Fashion	–	5,266†	–	–	✓	–	–	–
LPR4M (Yang et al., 2023)	Livestream commerce	66,358	20,079†	–	✓	✓	–	–	–
Charades-STA (Gao et al., 2017)	Indoor activities	–	1,334	30.6 s	–	–	✓	–	–
QVHighlights (Lei et al., 2021)	Vlogs/news	–	1,542†	≈150 s	–	–	✓	✓	–
MomentSeeker (Yuan et al., 2025)	Open-domain	–	268	1,201.9 s	–	–	✓	✓	–
GPUB (ours)	**Fashion livestream**	**31,831**	**3,000**	**265.1 s**	✓	✓	✓	✓	✓

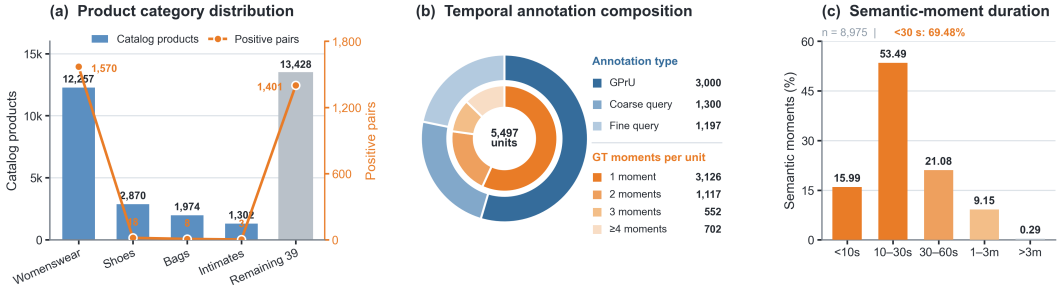


Figure 3: GPUB characteristics. **(a)** Distribution of catalog products and positive pairs across major product categories, with the remaining 39 categories aggregated. **(b)** Composition of temporal annotations by annotation type (outer ring) and number of ground-truth moments per annotation unit (inner ring). **(c)** Duration distribution of product evidence moments.

instance. We use Qwen3-VL-Embedding-2B (Li et al., 2026) to retrieve the 29 highest-ranked non-target products from the full catalog as hard negatives. Together with the target product p^*, they form the candidate set, with p^* randomly positioned to avoid positional bias. GPrU is evaluated over these candidate sets, whereas Product Retrieval is evaluated over the full catalog.

3.3 Benchmark Characteristics

As shown in Table 1, GPUB distinguishes itself from existing benchmarks through its large product space and comprehensive evaluation of both product identification and temporal evidence localization. Figure 3 further characterizes the internal data distribution of GPUB. Product evidence is both sparse and temporally distributed relative to the video context: videos are 4–5 minutes long on average, while 69.48% of product evidence moments are shorter than 30 seconds. Moreover, 43.13% of task-specific annotation units contain multiple ground-truth moments, with relevant evidence potentially occurring at multiple non-contiguous moments throughout a video. GPUB further provides text queries at different levels of specificity to capture product evidence localization at different granularities. Together, these characteristics require models not only to distinguish fine-grained products within a large product space, but also to localize and integrate temporally distributed evidence over extended video contexts.

3.4 Evaluation Metrics

Product Retrieval is evaluated over the full catalog using Recall@K, mean reciprocal rank (MRR), and normalized discounted cumulative gain (nDCG@K), measuring top-K retrieval success, the reciprocal rank of the target product, and ranking quality with position discounting (Voorhees &

Tice, 2000; Järvelin & Kekäläinen, 2002). Only the target product is treated as relevant. **Product Moment Localization** is evaluated using temporal mAP and Recall@K at temporal intersection-over-union (tIoU) thresholds, with ranked predictions matched one-to-one to query-specific ground-truth moments. For **GPrU**, we introduce two joint metrics: Pair mAP for ranked product–moment predictions and Joint Recall for instance-level joint success, both defined below.

Pair mAP. For pair-level evaluation, each predicted moment $\hat{m}_{ij}$ forms a product–moment pair $(p_i, \hat{m}_{ij})$. For benchmark instance n, pairs are processed in descending confidence order, and a pair is considered correct at threshold τ if $p_i = p_n^*$ and $\hat{m}_{ij}$ is greedily matched to the unmatched ground-truth moment $e \in \mathcal{E}_n^*$ with the highest tIoU, where $\mathrm{tIoU}(\hat{m}_{ij}, e) \geq \tau$. We compute standard AP from the ranked pairs for each instance and average it across all instances:

$$\text{Pair mAP } @\tau = \frac{1}{N} \sum_{n=1}^{N} \mathrm{AP}_n^{\tau}. \tag{1}$$

Unmatched predictions are treated as false positives, while unmatched ground-truth moments remain in the AP denominator. We report results at $\tau \in \{0.3, 0.5\}$.

Joint Recall. Joint Recall measures instance-level joint success in product identification and evidence localization. An instance is counted as successful when the target product is ranked within the top K and at least one predicted moment for that product overlaps a ground-truth evidence moment with $\mathrm{tIoU} \geq \tau$. The metric is computed as the fraction of successful instances over the benchmark. We report $K \in \{1, 5\}$.

4 METHOD

4.1 OVERVIEW

As illustrated in Fig. 4, UniPro is built on Qwen3-VL-Embedding-2B and leverages shared multimodal representations to jointly perform product identification and temporal evidence localization. Given sampled video frames and selected timestamped ASR utterances, the shared encoder first produces multimodal token representations. These representations are used to derive a product-aligned video embedding for product retrieval and are further temporally structured for moment localization. Depending on the task, moment localization is conditioned on either a text query in Product Moment Localization or candidate products in GPrU.

4.2 PRODUCT-ALIGNED MULTIMODAL REPRESENTATION

We first adapt Qwen3-VL-Embedding-2B to align livestream content with catalog products in a shared semantic space. Given a cropped product evidence moment and its corresponding catalog product, both are encoded by the shared backbone and optimized in a common embedding space using an in-batch contrastive objective. We then extend this product-level alignment from cropped evidence moments to complete livestream instances. Since livestream ASR is often lengthy and contains substantial content unrelated to the target product, encoding the full transcript consumes excessive tokens and introduces considerable redundancy. We therefore employ a lightweight Visual-Conditioned Utterance Selector that scores each timestamped utterance conditioned on a pooled visual representation and retains relevant utterances under a fixed token budget. The selected utterances preserve their original timestamps and temporal order. The selected ASR and sampled video tokens are jointly encoded by the product-aligned backbone, producing shared multimodal token states $H \in \mathbb{R}^{L \times d}$. For Product Retrieval, we normalize the last valid token state in H as the global video embedding and rank catalog products by cosine similarity.

4.3 TEMPORAL EVIDENCE GROUNDING

While the shared multimodal representation H captures visual and spoken content, moment localization additionally requires organizing this information along the video timeline. We therefore apply Temporally Anchored Cross-Attention to obtain temporally structured representations from H. Specifically, K learnable anchor queries distributed over the normalized video timeline attend to the multimodal tokens in H, producing $Z = \{z_k\}_{k=1}^{K} \in \mathbb{R}^{K \times d_t}$. During cross-attention, temporal

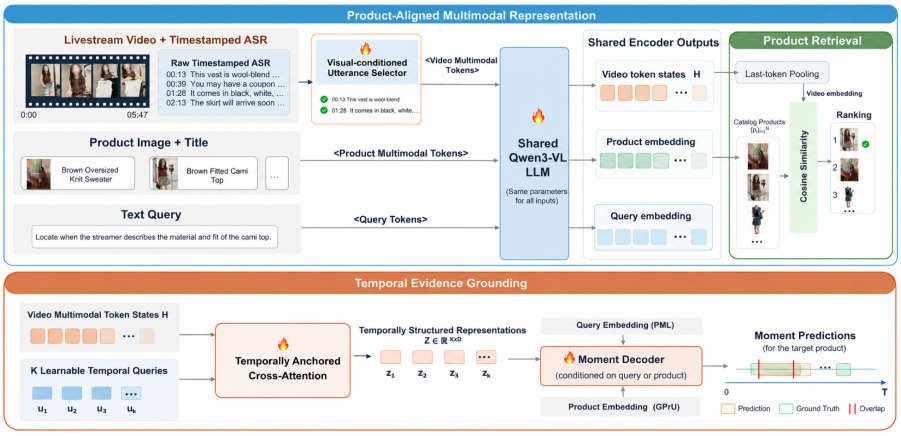


Figure 4: **Overview of UniPro.** UniPro builds a shared product-aligned multimodal representation for product retrieval and temporal evidence grounding, where temporally structured representations are further constructed for moment localization.

information is incorporated through temporal encodings of both the multimodal tokens and anchor queries, together with a temporal-distance bias. Based on the temporal representations Z, we use a Moment Decoder to localize the evidence moments corresponding to a text query or candidate product. The Moment Decoder is implemented as a three-layer Transformer decoder with a set of learnable moment queries. The semantic condition is projected and added to the moment queries, which attend to Z and jointly predict normalized temporal intervals and their confidence scores. For Product Moment Localization, the text-query embedding q serves as the condition. For Grounded Product Understanding, each candidate-product embedding p_i serves as the condition, with product–moment pairs ranked by combining the product matching score and moment confidence.

5 Experiments

5.1 Experimental Setup

We evaluate two groups of representative baselines according to the requirements of the GPUB tasks. For Product Retrieval, we consider Vision–Language embedding models for efficient fine-grained matching between product evidence moments and catalog products. For Product Moment Localization and Grounded Product Understanding, we evaluate general-purpose MLLMs for temporal evidence localization over long-form livestreams. These baseline models and their detailed evaluation settings are provided in Appendix A. UniPro is evaluated across all three tasks. Its implementation and training details, including training-data statistics and split integrity, are provided in Appendix B. All models are evaluated on the complete GPUB test set following the task-specific protocols in Section 3.1.

5.2 Main Benchmark Results

Table 2 summarizes the overall results. Results indicate that GPUB remains challenging for current models across all three tasks. On full-catalog Product Retrieval, even Qwen3-VL-Embedding-8B achieves only 15.05 R@1. The challenge extends to temporal grounding and becomes more pronounced in the joint GPrU setting, where Qwen3-VL-8B achieves only 10.13 Pair mAP@.3. Task-specific fine-tuning substantially improves its performance on both PML and GPrU, demonstrating the benefit of downstream supervision. UniPro further improves over this SFT baseline by 7.22 on PML mAP@.3 and 2.66 on GPrU Pair mAP@.3, reaching 53.50 and 24.58, respectively. On GPrU Pair mAP@.5, Qwen3-VL-8B (SFT) obtains 13.77, compared with 14.33 for UniPro. These results demonstrate the benefits of task-specific supervision and dedicated modeling across the complementary capabilities evaluated by GPUB. Additional results are provided in Appendix A.3.

Table 2: Main results on the three GPUB tasks (%). Metric definitions are provided in Section 3.4. In panel (c), Pair mAP and Joint R denote the product–moment correspondence metrics, while Ret. and Loc. denote product-retrieval and temporal-localization diagnostics, respectively. The subscript 30 indicates the fixed 30-candidate setting. Higher is better.

(a) Product Retrieval — Full Catalog

Model	Recall				nDCG			MRR
	R@1	R@10	R@30	R@50	@10	@30	@50	
Chinese-CLIP-336px	1.73	6.35	10.27	13.22	3.72	4.65	5.20	3.23
GME-Qwen2-VL-2B	3.58	11.46	19.43	24.56	6.96	8.83	9.80	6.18
SigLIP2-SO400M-Patch14-384	3.85	11.56	18.43	22.08	7.30	8.93	9.61	6.48
FG-CLIP2-Base	4.22	11.51	16.96	19.71	7.48	8.76	9.28	6.62
Qwen3-VL-Embedding-2B	13.04	34.22	45.66	51.32	22.79	25.52	26.59	20.06
Qwen3-VL-Embedding-8B	15.05	37.89	50.48	57.02	25.50	28.48	29.70	22.54
UniPro	**33.47**	**58.20**	**69.11**	**73.65**	**45.31**	**47.90**	**48.75**	**42.00**

(b) Product Moment Localization — 2,497 Queries

Model	tIoU = 0.3				tIoU = 0.5			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Qwen2.5-VL-7B	14.71	20.10	22.61	17.26	6.17	7.85	8.76	6.96
Qwen3-VL-8B	23.09	33.49	36.93	27.96	10.89	15.43	17.16	12.96
InternVL3-8B	8.98	14.70	16.87	11.64	2.99	4.71	5.29	3.74
InternVideo2.5-7B	13.80	21.65	26.22	17.65	5.37	7.78	9.09	6.55
Keye-VL-1.5-8B	11.56	19.43	22.31	15.10	4.53	7.12	7.99	5.64
MiniCPM-V-4.5	6.27	11.26	12.65	8.43	1.91	3.40	3.71	2.56
GLM-4.6V-Flash	22.63	33.03	36.65	27.42	11.08	16.12	18.11	13.31
Qwen3-VL-8B (SFT)	42.65	47.62	52.62	46.28	21.65	32.36	36.36	29.39
UniPro	**43.34**	**64.05**	**78.22**	**53.50**	**25.76**	**38.20**	**49.01**	**32.80**

(c) Grounded Product Understanding — 30 Candidates

Model	Product–Moment Correspondence				Diagnostics			
	Pair mAP@.3	Pair mAP@.5	Joint R@1@.3	Joint R@5@.3	Ret. $R@1_{30}$	Ret. mAP_{30}	Loc. R@1@.3	Loc. mAP@.3
Qwen2.5-VL-7B	3.53	1.57	4.13	7.07	19.20	27.82	10.47	11.22
Qwen3-VL-8B	10.13	5.03	18.03	19.93	39.10	51.63	15.92	19.55
InternVL3-8B	5.34	2.46	5.43	11.07	22.23	31.38	14.71	14.81
InternVideo2.5-7B	0.44	0.07	1.23	2.20	11.63	16.86	1.58	5.26
Keye-VL-1.5-8B	2.27	1.13	4.13	4.17	17.47	17.62	9.14	9.87
MiniCPM-V-4.5	2.42	0.92	3.30	5.90	17.83	26.17	4.80	6.31
GLM-4.6V-Flash	7.37	3.87	12.53	16.10	22.70	27.78	8.30	16.34
Qwen3-VL-8B (SFT)	21.92	13.77	35.47	35.47	38.60	47.43	39.83	52.15
UniPro	**24.58**	**14.33**	**38.81**	**65.54**	**39.40**	**52.54**	**40.82**	**56.76**

Table 3: Grounded Product Understanding ablations under the fixed 30-candidate protocol (%). **(a)** Representation architecture. **(b)** ASR content selection.

(a) Architecture design.

Variant	Ret. $R@1_{30}$	Loc. mAP @.3	Loc. mAP @.5	Joint R@1@.3	Pair mAP @.3	Pair mAP @.5
w/o Separate Retrieval Representation	13.64	44.59	22.00	13.50	10.23	5.13
w/o Temporally Structured Representations	38.80	50.66	28.75	37.94	21.57	11.60
Full UniPro	**39.40**	**56.76**	**32.80**	**38.81**	**24.58**	**14.33**

(b) ASR content selection.

ASR policy	Ret. $R@1_{30}$	Loc. mAP @.3	Loc. mAP @.5	Pair mAP @.3	Pair mAP @.5	Joint R@1@.3
No ASR	37.97	48.70	26.83	19.94	10.33	36.47
Random ASR	39.23	54.36	32.04	22.57	12.85	38.27
Unfiltered ASR	38.20	56.12	**33.52**	23.76	13.92	37.64
Selected ASR	**39.40**	**56.76**	32.80	**24.58**	**14.33**	**38.81**

5.3 Analysis

Similar product identification performance but different GPrU performance. As shown in Fig. 5(a), strong product identification does not necessarily translate into strong joint grounding performance. UniPro achieves a Ret. mAP_{30} of 52.54 and a normalized Pair mAP@.3 of 46.78. In the Top-1 comparison in Fig. 5(b), Qwen3-VL-8B and UniPro achieve nearly identical Ret. $R@1_{30}$ scores of 39.10 and 39.40, respectively, yet their Joint R@1@.3 scores differ substantially, at 18.03 and 38.81. For UniPro, the outcome decomposition comprises 38.81 joint success, 0.59 correct product identification with failed localization, and 60.60 incorrect product identity. This contrast highlights the need to jointly evaluate product identity and its supporting temporal evidence.

Fine-grained catalog ambiguity challenges product identification. We group samples by the similarity between each target product and its hardest negative, with results reported in Appendix

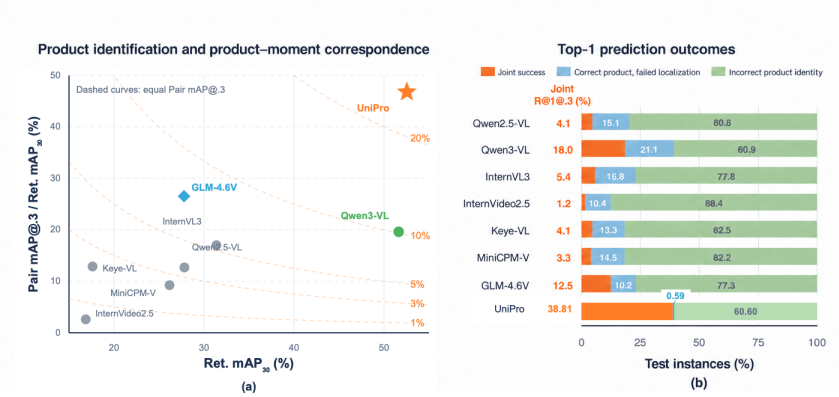


Figure 5: (a) Comparison of product identification and product–moment correspondence across models. Ret. mAP_{30} measures product identification, while normalized Pair mAP@.3 reflects product–moment correspondence; dashed curves indicate equal Pair mAP@.3. (b) Outcome decomposition of GPrU Top-1 predictions into joint success, localization failure after correct product identification, and incorrect product identity.

Table 5(a). Identification performance degrades as similarity increases: from low to high similarity, Ret. R@1_{30} drops by 25.98, 9.76, and 27.15 for Qwen3-VL-8B, GLM-4.6V-Flash, and UniPro, respectively. The prevalence of visually and semantically similar products in GPUB makes fine-grained product identification particularly challenging. These errors also limit GPrU performance, since joint success requires both correct product identification and evidence localization.

Semantic and distributed evidence pose distinct challenges to temporal localization. As shown in Appendix Fig. 6, all seven general-purpose MLLMs perform better on fine-grained queries than on coarse-grained queries, suggesting that explicit attribute cues may facilitate temporal localization. This suggests that such cues provide anchors for temporal localization, whereas semantically defined evidence remains more challenging to ground. Meanwhile, GT-Recall@10 consistently decreases as the number of ground-truth moments increases (Appendix Table 5(b)), indicating that evidence distributed across multiple moments is more difficult to recover completely.

Ablation Study. Table 3 examines representation design and ASR content selection. Using temporally structured representations for product retrieval instead of a separate retrieval representation reduces Ret. R@1_{30} by 25.76 (Table 3(a)), indicating that temporal structure is less suitable for product matching. Removing temporal structuring reduces Loc. mAP@.3 by 6.10 while leaving retrieval largely unchanged, confirming its importance for moment localization. For ASR selection (Table 3(b)), removing ASR reduces Loc. mAP@.3 and Pair mAP@.3 by 8.06 and 4.64, respectively, showing that ASR provides useful evidence. Random and unfiltered ASR recover much of this loss. Compared with unfiltered ASR, selected ASR reduces the ASR token budget while maintaining comparable performance and achieving the highest Pair mAP@.3 (24.58). Overall, these results highlight the importance of task-specific representations and efficient ASR selection in UniPro.

6 Conclusion

We introduced GPUB, a large-scale benchmark for product understanding in livestream videos, centered on the new GPrU task of jointly identifying the target product and localizing its supporting temporal evidence. Our evaluation reveals persistent challenges arising from fine-grained catalog ambiguity, temporally distributed evidence, and the need to establish correct product–moment correspondence. We further develop UniPro, a unified baseline that addresses all three tasks within a shared framework. We hope GPUB will provide a foundation for future research on grounded multimodal understanding in realistic long-form videos.

AI USE STATEMENT

Generative AI tools were used for literature search, language editing, and polishing. All AI-assisted changes were reviewed and verified by the authors. The authors take full responsibility for the final content of this paper.

PROJECT PAGE

The anonymous project page and accompanying code are available at `https://anonymous.4open.science/api/repo/gpub-evaluation-C2E4/file/project-page/index.html`.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Xuehan Bai et al. Cross-domain product representation learning for rich-content e-commerce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5697–5706, 2023.

Shyamal Buch et al. Flexible frame selection for efficient video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Joya Chen et al. LiveCC: Learning video LLM with streaming speech transcription at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhi-Qi Cheng et al. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4048–4056, 2017.

Jiyang Gao et al. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Bo He et al. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Lisa Anne Hendricks et al. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Kai Hu et al. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xiaowan Hu et al. Spatiotemporal graph guided multi-modal network for livestreaming product retrieval. *arXiv preprint arXiv:2407.16248*, 2024.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Kwai Keye Team. Keye-VL-1.5-8B. Hugging Face model card, 2025. URL `https://huggingface.co/Kwai-Keye/Keye-VL-1_5-8B`.

Jie Lei, Tamara L. Berg, and Mohit Bansal. QVHighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, 2021.

Linjie Li et al. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Mingxin Li et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Kevin Qinghong Lin et al. UniVTG: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Xiangrui Liu et al. TimeScope: Towards task-oriented temporal grounding in long videos. *arXiv preprint arXiv:2509.26360*, 2025.

Xianjie Liu et al. E-VAds: An e-commerce short videos understanding benchmark for MLLMs. *arXiv preprint arXiv:2602.08355*, 2026.

WonJun Moon et al. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xiaoqian Shen et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Mattia Soldan et al. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5026–5035, 2022.

Enxin Song et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Chaolei Tan et al. SynopGround: A large-scale dataset for multi-paragraph video grounding from tv dramas and synopses. *arXiv preprint arXiv:2408.01669*, 2024.

Michael Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 200–207, 2000. doi: 10.1145/345508.345577.

Yi Wang et al. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025.

Junbin Xiao et al. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

Chunyu Xie et al. FG-CLIP 2: A bilingual fine-grained vision-language alignment model. *arXiv preprint arXiv:2510.10921*, 2025.

Hu Xu et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

An Yang et al. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *arXiv preprint arXiv:2211.01335*, 2022.

Wenjie Yang et al. Cross-view semantic alignment for livestreaming product recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Tianyu Yu et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Huaying Yuan et al. Momentseeker: A task-oriented benchmark for long-video moment retrieval. *arXiv preprint arXiv:2502.12558*, 2025.

Z.ai Team. GLM-4.6V-Flash. Hugging Face model card, 2025. URL `https://huggingface.co/zai-org/GLM-4.6V-Flash`.

Xunlin Zhan et al. Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Hao Zhang et al. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Jun Zhang et al. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026.

Xin Zhang et al. GME: Improving universal multimodal retrieval by multimodal large language models. *arXiv preprint arXiv:2412.16855*, 2024.

Yanhao Zhang et al. Fashion focus: Multi-modal retrieval system for video commodity localization in e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Minghang Zheng et al. TaRO: Temporal-aware reasoning optimization for video temporal grounding. *arXiv preprint arXiv:2606.09248*, 2026.

Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A BASELINE EVALUATION DETAILS

This section describes the evaluation settings for the GPUB baselines, including input construction, prompting, and model-output processing.

A.1 VISION–LANGUAGE EMBEDDING BASELINES

For Product Retrieval, each annotated product evidence moment is represented by 10 uniformly sampled RGB frames and the corresponding ASR transcript, while each catalog product is represented by its main image and title. For each model, we use its officially released processor and follow the official inference configuration for embedding extraction. On the video side, the sampled frames and corresponding ASR are used to construct the query representation; on the product side, the product image and title are used to construct the catalog-product representation.

Qwen3-VL-Embedding-2B/8B (Li et al., 2026) and GME-Qwen2-VL-2B (Zhang et al., 2024) directly encode the corresponding multimodal inputs to produce the video and product representations. For the dual-tower models, including FG-CLIP2-Base (Xie et al., 2025), SigLIP2-SO400M-Patch14-384 (Tschannen et al., 2025), and Chinese-CLIP-336px (Yang et al., 2022), the visual and textual inputs are encoded separately. On the video side, the visual representations of the sampled frames are first mean-pooled and then fused with the corresponding ASR representation. On the product side, the product image and title representations are fused in the same manner. Both sides use a visual–text fusion weight of $0.7{:}0.3$. The resulting video and product representations are ℓ_2-normalized, and products are ranked over the full catalog using cosine similarity.

A.2 MULTIMODAL LLM BASELINES

For a consistent comparison across MLLM baselines, including Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), InternVL3-8B (Zhu et al., 2025), InternVideo2.5-7B (Wang et al., 2025), Keye-VL-1.5-8B (Kwai Keye Team, 2025), MiniCPM-V-4.5 (Yu et al., 2025), and GLM-4.6V-Flash (Z.ai Team, 2025), we use the same input preprocessing and inference settings. For each livestream video, we uniformly sample 64 frames and resize each frame so that its longer side is 448 pixels. The timestamped ASR is preserved in its original temporal order and truncated to at most 4,096 characters, with timestamps expressed relative to the beginning of the video. All models use greedy decoding and the same task-specific Chinese instructions. For Product Moment Localization, the instruction additionally provides the text query, whereas for Grounded Product Understanding, it provides the 30 candidate products through their titles and main images without an additional text query. For readability, the prompts are presented below in English translation.

Prompt Box — Product Moment Localization

Given an e-commerce livestream video represented by uniformly sampled frames and timestamped ASR, together with a text query:

Video:
{`uniformly sampled video frames`}

Timestamped ASR:
{`timestamped ASR`}

Query:
{`query`}

Localize all temporal moments in the video that satisfy the query.

Assign an independent confidence score in $[0, 1]$ to each predicted moment and rank the predictions by confidence. Return at most 10 moments.

Output only structured JSON containing `moments`, where each moment is represented by its start and end timestamps and confidence score.

Prompt Box — Grounded Product Understanding

Given an e-commerce livestream video represented by uniformly sampled frames and timestamped ASR, together with N candidate products, each represented by its title and main image:

Video:
{`uniformly sampled video frames`}

Timestamped ASR:
{`timestamped ASR`}

Candidate products:
$\{(i, \text{title}_i, \text{image}_i)\}$ for $i = 1, \ldots, N$

Identify the candidate product being discussed in the video and localize all temporal moments that provide visual or spoken evidence about its attributes or characteristics.

Mere product appearances and segments primarily concerning promotion, price, inventory, or casual conversation should not be considered evidence moments.

Return three ranked sets of predictions:
1. candidate products;
2. evidence moments;
3. product–moment pairs.

Assign an independent confidence score in $[0, 1]$ to each prediction and rank each set by confidence. Return at most 10 moments and 10 product–moment pairs.

Output only structured JSON containing `ranked_products`, `moments`, and `pairs`, where each moment is represented by its start and end timestamps and each pair additionally contains the corresponding candidate product index.

Output Processing. For Product Moment Localization, the returned temporal intervals and their confidence scores are used directly for temporal localization evaluation. For Grounded Product Understanding, each model returns a ranked list of candidate products, temporal intervals, and product–moment pair predictions. Product–moment predictions are ranked by the confidence scores generated by each model for AP computation. Candidate indices are mapped back to their corresponding product IDs. Temporal intervals are restricted to the current evaluation window and converted from window-relative timestamps to absolute timestamps in the original video. Invalid entries that cannot be parsed into the required format, contain invalid candidate indices, or do not define valid temporal intervals are discarded. If an output cannot be parsed, it is treated as an empty prediction. For Pair mAP, at most the top 10 ranked product–moment pairs per instance are retained for all methods, matching the maximum number of ground-truth evidence moments in an instance. The resulting predictions are evaluated using the metrics defined in Section 3.4.

A.3 Supervised Fine-Tuning Results

To provide a controlled comparison under downstream supervision, we fine-tune representative MLLMs on the GPUB training split. Each model is trained following the GPrU formulation to jointly identify the target product and localize its supporting moments. The resulting models are evaluated on both PML and GPrU using the same evaluation protocols as in the main experiments. Table 4 reports the detailed results. Beyond the main metrics discussed in Sec. 5.2, UniPro shows larger gains at higher recall cutoffs on PML, indicating better coverage of relevant evidence moments. On GPrU, improvements in both product identification and evidence localization diagnostics suggest that its gain in joint product–moment prediction is supported by both capabilities.

Table 4: **Supervised fine-tuning results on GPUB (%).** **(a)** Product Moment Localization. **(b)** Grounded Product Understanding under the fixed 30-candidate protocol. The best SFT baseline in each column is underlined; UniPro is highlighted in blue.

(a) Product Moment Localization

Model	tIoU = 0.3				tIoU = 0.5			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Qwen2.5-VL-7B (SFT)	34.23	41.84	46.38	36.17	16.53	25.22	29.16	21.16
Keye-VL-1.5-8B (SFT)	32.81	41.37	46.50	35.05	16.09	25.42	29.64	21.08
GLM-4.6V-Flash (SFT)	40.28	44.48	47.13	42.05	21.61	32.93	35.63	28.10
Qwen3-VL-8B (SFT)	42.65	47.62	52.62	46.28	21.65	32.36	36.36	29.39
UniPro	**43.34**	**64.05**	**78.22**	**53.50**	**25.76**	**38.20**	**49.01**	**32.80**

(b) Grounded Product Understanding — 30 Candidates

Model	Product–Moment Correspondence				Diagnostics			
	Pair mAP@.3	Pair mAP@.5	Joint R@1@.3	Joint R@5@.3	Ret. $R@1_{30}$	Ret. mAP_{30}	Loc. R@1@.3	Loc. mAP@.3
Qwen2.5-VL-7B (SFT)	12.68	8.54	20.34	23.16	20.71	27.51	31.42	41.37
Keye-VL-1.5-8B (SFT)	11.47	7.63	18.21	21.26	18.92	21.34	30.16	39.82
GLM-4.6V-Flash (SFT)	18.74	13.26	24.86	31.82	29.63	31.37	36.72	47.36
Qwen3-VL-8B (SFT)	21.92	13.77	35.47	35.47	38.60	47.43	39.83	52.15
UniPro	**24.58**	**14.33**	**38.81**	**65.54**	**39.40**	**52.54**	**40.82**	**56.76**

A.4 Difficulty Analysis

GPrU difficulty factors. As shown in Table 5, we examine two sources of difficulty in GPrU: product ambiguity and temporal evidence complexity. The product-ambiguity analysis uses the subset of examples with valid hardest-negative similarity scores. Product identification degrades as the target product becomes more similar to its hardest negative, while temporal localization becomes more difficult as the number of evidence moments increases.

Table 5: **Analysis of product identification and temporal localization under different difficulty factors (%).** **(a)** Product identification performance across different levels of similarity between the target product and its hardest negative. **(b)** Temporal localization performance across different numbers of ground-truth moments. GT-Recall@10 measures the recall of annotated moments among the top 10 predictions for the ground-truth product at tIoU = 0.3.

(a) Product Ambiguity

Similarity	Ret. $R@1_{30}$			Joint R@1@.3		
	Qwen3-VL	GLM-4.6V	UniPro	Qwen3-VL	GLM-4.6V	UniPro
Low	52.54	35.35	67.97	7.03	12.11	65.04
Medium	39.65	28.91	52.73	6.64	9.96	50.59
High	26.56	25.59	40.82	3.71	11.52	38.48

(b) Temporal Evidence Complexity

GT moments	N	GT-Recall@10			Pair mAP@.3		
		Qwen3-VL	GLM-4.6V	UniPro	Qwen3-VL	GLM-4.6V	UniPro
1	1,127	49.59	33.19	91.75	17.12	12.03	28.37
2	731	34.43	23.59	83.79	8.99	7.02	23.83
≥ 3	1,142	18.11	11.95	67.53	3.96	2.99	21.32
Overall	3,000	33.92	22.77	80.59	**10.13**	**7.37**	**24.58**

Query-type breakdown. The Product Moment Localization evaluation set contains 1,300 coarse-grained queries with 1,703 ground-truth moments and 1,197 fine-grained queries with 1,506 ground-truth moments. Figure 6 compares the two query types for the seven general-purpose MLLMs using mAP@0.3.

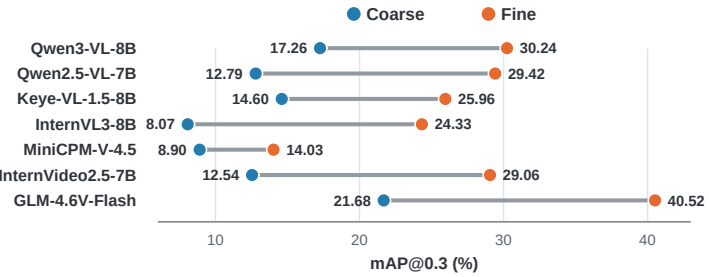


Figure 6: Product Moment Localization performance by query type for seven general-purpose MLLMs, measured by mAP@0.3.

B UniPro Implementation Details

B.1 Architecture Details

Utterance Selector. Each timestamped utterance is encoded into a representation u_i, augmented with its start time, end time, and duration, while the video is summarized into a global visual representation v. The selector computes the selection score as follows:

$$s_i^{\text{sel}} = \text{MLP}([u_i;\, v;\, u_i \odot v;\, |u_i - v|])\,. \tag{2}$$

Here, $\odot$ denotes element-wise multiplication and $|u_i - v|$ the element-wise absolute difference; the four components are concatenated before scoring. Higher-scoring utterances are retained as complete units, preserving their timestamps and temporal order.

Temporal Evidence Grounding. Video tokens are associated with frame timestamps, while ASR tokens inherit their utterance intervals; both are represented by normalized temporal center and duration. The number of uniformly spaced temporal anchors is determined by video duration T as $K(T) = \text{clip}(\lceil T/(5\,\text{s})\rceil, 16, 128)$. Temporal encodings are added to both the anchor queries and timestamped multimodal tokens, while a temporal-distance bias between each anchor center and token interval is applied to the cross-attention logits. Cross-attention then yields K temporally structured representations, one per anchor. The moment decoder uses $N_m = 10$ queries per semantic condition, distinct from the K temporal anchors. For GPrU, all candidate products share the same temporal representation Z, avoiding repeated video encoding. Product–moment pairs (p, j) are ranked by $s_{p,j} = \log \text{softmax}_p(s_p^{\text{ret}}/0.07) + \log \sigma(c_{p,j})$, where s_p^{ret} is the product retrieval similarity, $c_{p,j}$ is the moment confidence logit, and the softmax is over candidate products.

B.2 Training Strategy

UniPro is trained in three stages using the datasets described below.

Stage I: Video–Product Alignment. We adapt Qwen3-VL-Embedding-2B with LoRA using an in-batch InfoNCE objective:

$$\mathcal{L}_{\text{S1}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(v_i, p_i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(v_i, p_j)/\tau)}. \tag{3}$$

Here, v_i and p_i are the embeddings of a video and its matched product, B is the contrastive batch size, and sim denotes cosine similarity. The matched product is the positive, and the remaining products in the batch serve as negatives.

Stage II: ASR-Aware Multimodal Adaptation. Stage II initializes from the product-aligned model learned in Stage I, while the Qwen3-VL-Embedding-2B backbone remains frozen. The utterance selector is optimized on timestamped livestream inputs using the video–product matching objective, without utterance-level supervision. Hard utterance selection is applied in the forward pass, with a straight-through gate providing surrogate gradients to the selector and a soft budget loss regularizing the retention ratio.

Stage III: Temporal Evidence Grounding. During Stage III, the multimodal encoder remains frozen while the temporal grounding modules are optimized; the selector is updated after an initial warm-up. Predicted moments are assigned to ground-truth moments by Hungarian matching with

$$\mathcal{C}_{\text{match}} = 0.5\mathcal{C}_{\text{conf}} + 5\mathcal{C}_1 + 2\mathcal{C}_{\text{tGIoU}}, \tag{4}$$

where the terms measure confidence, normalized temporal-boundary distance, and temporal generalized IoU, respectively. The training objective is

$$\begin{aligned}\mathcal{L}_{\text{S3}} = 3\mathcal{L}_{\text{conf}} + 5\mathcal{L}_1 + 2\mathcal{L}_{\text{tGIoU}} + 0.3\mathcal{L}_{\text{rank}} \\ + \lambda_d(t)\left[\mathcal{L}_{\text{fg}} + 0.1\left(\mathcal{L}_{\text{start}} + \mathcal{L}_{\text{end}}\right)\right].\end{aligned} \tag{5}$$

$\mathcal{L}_1$ measures boundary error on the normalized timeline, and $\mathcal{L}_{\text{tGIoU}} = 1 - \text{mean}(\text{tGIoU})$. Confidence targets gradually transition from matching labels to localization-quality scores, defined as the maximum IoU between each proposal and any ground-truth moment. The pairwise ranking loss $\mathcal{L}_{\text{rank}}$ encourages higher-quality proposals to receive higher confidence, weighting each pair by its quality gap. For the auxiliary dense supervision, $\mathcal{L}_{\text{fg}}$ uses weighted binary cross-entropy with soft targets given by the overlap ratio between each temporal slot and the ground-truth intervals, while $\mathcal{L}_{\text{start}}$ and $\mathcal{L}_{\text{end}}$ use Gaussian heatmaps centered on the nearest ground-truth boundaries.

B.3 GPUB Training Split and Split Integrity

The GPUB training split is constructed using the same verification, evidence-annotation, and query-construction pipeline as the evaluation set. It contains 6,145 livestream windows spanning 465.29 hours, and the training candidate catalog contains 136,180 products. Detailed statistics are reported in Table 6. To prevent data leakage, windows sharing the same source livestream or target product are grouped and assigned to the same split, ensuring that neither source livestreams nor target products overlap across the training, validation, and test sets.

Table 6: GPUB split statistics. PML includes coarse- and fine-grained localization queries.

Statistic	Train	Validation	Test
Livestream windows	6,145	500	3,000
Source livestreams	2,311	239	1,798
Unique target products	2,693	214	1,641
Product evidence moments	15,708	1,373	8,975
PML queries	5,393	433	2,497

B.4 Implementation and Optimization Settings

The implementation and optimization settings are specified separately for the three training stages.

Stage I is trained for 8,000 steps on 4 GPUs with a per-GPU batch size of 8 and gradient accumulation of 4, using AdamW with a learning rate of 10^{-4}, LoRA rank 8, scaling factor $\alpha = 32$, and InfoNCE temperature $\tau = 0.1$. **Stage II** is trained for 2,000 steps using the same learning rate as Stage I. **Stage III** is trained for 6,000 steps on 2 GPUs with a per-GPU batch size of 8 and a learning rate of 5×10^{-5}. We uniformly sample 64 frames from each video. ASR selection is performed in 30-s blocks with 3–8 utterances per block and an overall 2,048 token ASR budget. We use AdamW with cosine learning-rate decay to 0.1 times the initial rate, 200 warm-up steps, weight decay of 0.01, and a maximum gradient norm of 1.0. The selector begins updating after 500 steps. The ranking loss forms ordered pairs with a quality gap above 0.1 and uses a confidence-logit margin of 0.2, while the boundary heatmaps use $\sigma = 1$ temporal slot.

C Ethics and Data Governance

GPUB is constructed from real-world e-commerce livestream data. To protect privacy, the original audio tracks are removed from all benchmark videos, and spoken content is provided only as de-identified timestamped ASR. Faces of livestream hosts and other identifiable individuals are blurred. Visible host nicknames and account IDs, QR codes, contact details, and store names are masked or

replaced. Viewer account identifiers appearing in video frames are detected using OCR and pixelated. Personal names and nicknames in ASR are replaced, while personally identifiable or sensitive information in product images, titles, and attributes is removed or obscured. Internal storage paths and original livestream and product identifiers are replaced with benchmark-specific identifiers, and timestamps are expressed relative to each video window. The released data have undergone privacy review and will be distributed under specified access and licensing conditions.