QB-Highlights: Quality-Guided Budgeted Highlight Detection in Videos with Dense Query-Relevant Moments

Chao Gao[1†], **Hui Li**[1†], **Ying Chen**[1], **Zongyao He**[1], **Sibin Deng**[1*]
[1]Alibaba Group
{shuilang.gc, lh401704, YingChen, hezongyao.hzy, sibin.dsb}@taobao.com

Abstract

Highlight detection in content-dense untrimmed videos, such as livestreams, is particularly challenging due to the prevalence of semantically similar moments and the application-driven demand of fitting a target duration budget. Existing methods typically rank segments independently based on semantic relevance, struggling to discriminate fine-grained quality among similar moments and allocating the budget to redundant selections. We therefore reformulate highlight detection as jointly selecting high-quality and complementary segments under a duration budget. To study this formulation, we introduce QB-Highlights, a large-scale benchmark of 50,000 e-commerce livestream videos with densely occurring query-relevant events and subtle quality variations. QB-Highlights is organized around two complementary dimensions: Segment Qualification, which assesses fine-grained per-segment quality; and Budgeted Composition, which evaluates set-level complementarity under a duration budget. It provides MLLM-assisted pseudo labels for training, dense human annotations for testing, and an evaluation protocol jointly measuring quality and diversity. We further propose a Qualification-then-Composition Reasoning (QCR) framework that performs structured deliberative reasoning to produce a highlight set matching the target duration. Extensive experiments show that QCR outperforms strong MLLM, moment retrieval, and highlight detection baselines on both segment quality and compositional diversity. The code and dataset are publicly available at `https://github.com/gclonghorn/QB-Highlights`.

1 Introduction

The rapid growth of long-form videos like livestreams and vlogs has made short highlights essential for quickly grasping video content. Video Highlight Detection (HD), which identifies salient moments from a video, has thus become a core capability of modern video platforms, powering applications such as previews and recommendation. Recent years have witnessed significant progress in HD methods [1, 2, 3, 4, 5, 6, 7], which typically predict query-conditioned relevance scores for individual segments and select the top-ranked ones as highlights. However, this relevance-based segment-wise ranking paradigm becomes insufficient in real-world scenarios characterized by two factors. First, videos such as e-commerce livestreams often contain many query-relevant moments whose quality varies substantially. Second, practical highlights are typically constrained by a duration budget. Together, these factors extend the key challenge beyond finding relevant moments to selecting high-quality ones and composing them effectively within the budget.

[†]Equal contribution.
[*]Corresponding author.

Concretely, as illustrated in Fig. 1, HD methods such as Moment-DETR rely on relevance-based saliency scores, which are easily misled by dense query-relevant events and produce mismatched selections; their per-segment ranking further leaves inter-segment redundancy unaddressed, allocating the limited budget to overlapping content. General-purpose MLLMs such as Qwen3-VL-8B, despite strong semantic understanding, remain insensitive to fine-grained visual quality (e.g., motion blur, weak subject prominence, occlusion), yielding unreliable quality judgments. Existing benchmarks also offer limited support for studying these challenges. As summarized in Tab. 1, prior datasets [8, 9, 10, 11, 12] fall short in three aspects. **(i) Limited scale**, due to the high cost of manual annotation [8, 9, 10]; **(ii) Sparse occurrence of query-relevant events**, making it difficult to distinguish fine-grained quality differences among semantically similar moments due to the lack of discriminative supervision; **(iii) Absence of budget-aware diversity**, as existing benchmarks neither impose a duration budget nor evaluate inter-segment redundancy, leaving models with no incentive to avoid overlapping selections.

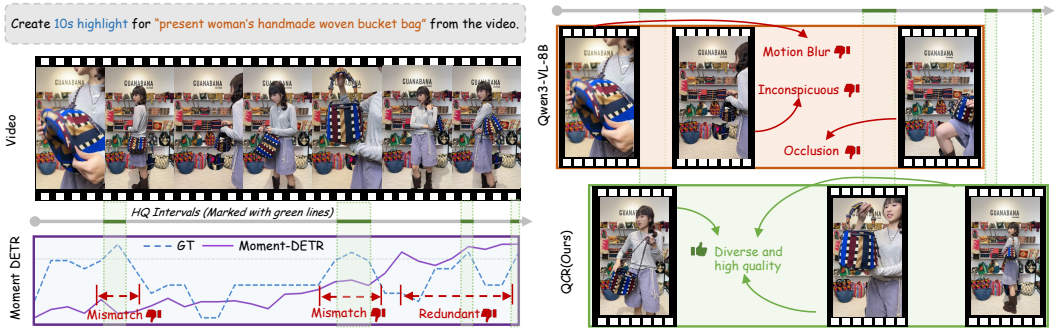

Figure 1: Qualitative comparison of highlight selection on a livestream video with densely occurring query-relevant moments. While Moment-DETR produces redundant selections that mismatch the high-quality ("HQ" in the figure) intervals and Qwen3-VL-8B overlooks fine-grained visual quality, our QCR selects segments that are both high-quality and diverse under the duration budget.

To address these issues, we reformulate highlight detection as *Quality-Guided Budgeted Video Highlight Detection*, where the goal is to select segments that are not only individually high-quality but also collectively complementary under a duration budget. We introduce **QB-Highlights**, a large-scale benchmark of 50,000 e-commerce livestream videos featuring dense occurrence of query-relevant events with subtle quality variations. QB-Highlights is organized around two complementary dimensions: Segment Qualification, which assesses individual segment quality along clarity, completeness, subject prominence, and temporal stability; and Budgeted Composition, which evaluates the selected highlights as a set, rewarding complementary coverage and penalizing redundancy under a duration budget. To enable scalable learning and reliable evaluation, QB-Highlights provides MLLM-assisted pseudo labels for training and dense human annotations for testing, together with an evaluation protocol that jointly measures segment-level quality and set-level diversity.

We further develop a **Qualification-then-Composition Reasoning (QCR)** framework for quality-guided budgeted video highlight detection. Rather than ranking segments independently by relevance, QCR casts highlight detection as a *deliberative reasoning process* and directly predicts a highlight set matching the target duration. As shown in Fig. 4, it first performs *segment qualification* to identify candidate moments and assess their quality, and then performs *budgeted composition* to assemble a complementary subset under the duration budget. To endow this reasoning capability, we adopt a two-stage training strategy. QCR first learns to follow structured reasoning traces via supervised fine-tuning, and is then further optimized with GRPO using a trace-aware reward that jointly supervises final highlight quality and intermediate reasoning behavior. Extensive experiments show that QCR outperforms a broad set of strong baselines, improving both segment quality and compositional diversity.

In summary, our contributions are four-fold: **First,** we reformulate highlight detection in videos with abundant semantically similar moments as *Quality-Guided Budgeted Video Highlight Detection*, shifting from segment-wise semantic-relevance ranking to set-level selection of high-quality, complementary segments under a duration budget. **Second,** we introduce **QB-Highlights**, a large-scale benchmark of 50,000 e-commerce livestream videos featuring densely occurring query-relevant

events and subtle quality variations, together with a budget-aware protocol that jointly evaluates segment quality and compositional diversity. **Third,** we propose a **QCR** framework that performs structured deliberative reasoning to produce a highlight set matching the target duration. **Fourth,** extensive experiments show that QCR outperforms strong MLLM, moment retrieval, and highlight detection baselines on QB-Highlights in both segment quality and compositional diversity.

2 Related Works

2.1 Moment Retrieval & Video Highlight Detection

Moment Retrieval (MR) [13, 14, 15, 16] and Video Highlight Detection (HD) are two closely related tasks in query-conditioned video understanding. MR aims to localize temporal spans that match a natural language query, typically through proposal scoring [13, 17, 18, 19, 20] or direct boundary regression [14, 21, 22, 23, 24]. HD instead focuses on identifying salient or query-relevant moments, usually by predicting clip-level saliency scores. Recent benchmarks and methods, such as QVHighlights [12], further bridge MR and HD by jointly detecting query-relevant moments and estimating their saliency under a DETR-based framework [25]. Subsequent works improve this paradigm with stronger query-video interaction modules [1, 2, 3, 26, 27], pre-trained vision-language representations [4, 28], and more efficient training strategies [6, 7]. Together, these developments make MR and HD increasingly intertwined: MR emphasizes precise temporal interval prediction, while HD emphasizes saliency-based moment selection. However, both remain relevance-centric and fall short in real-world application scenarios such as e-commerce livestreams, where query-relevant moments are abundant yet differ subtly in quality, and highlights are typically demanded with a duration budget. In contrast, we cast highlight detection as a quality-guided, budget-aware composition problem.

2.2 Temporal Reasoning in Video-MLLMs

Recent general-purpose Multimodal Large Language Models (MLLMs) have made substantial progress in video understanding and temporal reasoning [29, 30, 31, 32]. However, these gains mainly come from proprietary large-scale MLLMs, whose closed-source access and high computational cost limit practical deployment. This motivates a growing line of research [33, 34, 35] that equips smaller, open-source video-MLLMs with stronger temporal reasoning capabilities through targeted fine-tuning and reinforcement learning. For instance, VideoMind [33] introduces a Chain-of-LoRA mechanism that allows a single backbone to switch among multiple reasoning roles via lightweight LoRA adapters, achieving strong performance on temporal-grounded video understanding. Video-R1 [34] and Time-R1 [16] extend GRPO to video reasoning, encouraging models to explicitly exploit temporal cues during inference. Video-Thinker [35] further integrates grounding and captioning into the chain-of-thought process, improving performance on video question answering. Building on this line of MLLM-based temporal reasoning, we formulate highlight detection as a deliberative reasoning process and train the MLLM in two stages: supervised fine-tuning to learn structured reasoning traces, followed by GRPO to jointly reward final highlight quality and intermediate reasoning behavior.

3 Problem Formulation

When videos contain many query-relevant moments under a duration budget, relevance-based selection is insufficient: it cannot discriminate fine-grained quality differences among similar moments, and independent ranking tends to produce redundant selections that waste the budget. We therefore formulate highlight detection as *Quality-Guided Budgeted Video Highlight Detection*, aiming to select high-quality segments that jointly provide complementary information within the target duration. Given a video V and a query q, the goal is to predict a set of temporal segments $S = \{s_i\}_{i=1}^{N}$, where each $s_i = [t_i^s, t_i^e]$ denotes a candidate segment with duration $|s_i| = t_i^e - t_i^s$. Under a target duration budget B, the feasible set is defined as

$$\mathcal{F}(B) = \left\{ S = \{s_i\}_{i=1}^{N} \;\middle|\; \left| \sum_{i=1}^{N} |s_i| - B \right| \leq \epsilon,\ s_i \cap s_j = \varnothing \ \ \forall i \neq j \right\}, \tag{1}$$

where the tolerance ϵ allows the total selected duration to approximately match the budget rather than enforcing a hard equality constraint. The optimal highlight set is then defined as

$$S^* = \arg \max_{S \in \mathcal{F}(B)} U(S \mid V, q), \tag{2}$$

where the utility $U(\cdot)$ is governed by two coupled criteria: *Segment Qualification* evaluates the standalone quality of each candidate segment for the query, covering clarity, completeness, subject prominence, and temporal stability; *Budgeted Composition* evaluates the selected set as a whole, encouraging complementary coverage of informative content while discouraging redundant selections under the fixed duration budget.

4 QB-Highlights Benchmark

Data Overview. We introduce **QB-Highlights** (Table 1), a large-scale benchmark for *quality-guided budgeted highlight detection*. We build QB-Highlights from e-commerce livestreams, a representative domain with abundant semantically similar moments and strong practical demand for short-form highlights. The benchmark contains 50K video–query pairs collected from a leading e-commerce platform. All videos are 1080p clips sampled from popular livestream sessions, averaging 80 seconds in length with varied durations (Fig. 2(a)). Each query is a presentation request over a fine-grained product description (e.g., "present the women's double-breasted mid-length leather jacket"), and the queries collectively cover diverse product categories and visual attributes. Due to the dense recurrence of product-presentation events in livestreams, the benchmark emphasizes quality-aware selection among dense query-relevant moments. The target highlight duration is set to $B = 10$ seconds, matching the typical length of short-form video highlights used in practice. QB-Highlights adopts a hybrid supervision scheme to balance scale and annotation quality: 49.1K training samples use automatically constructed highlight labels for scalable learning, while 0.9K test samples are annotated by humans with dense clip-level quality labels to ensure reliable evaluation.

Table 1: Comparison with existing highlight detection datasets.

Dataset	Domain	#Videos	Avg. Len	Test Label	Occurrence of Query-Relevant Events	Budgeted
TVSum [8]	Open-domain	50	203s	Frame	Sparse	✗
SumMe [9]	Open-domain	25	147s	Frame	Sparse	✗
YouTube-HL [10]	Activity	600	143s	Clip	Sparse	✗
ActivityThumb [11]	Activity	4K	60s	Clip	Sparse	✗
QVHighlights [12]	Vlog/News	10.2K	150s	2s Clip	Sparse	✗
QB-Highlights (ours)	E-commerce Livestream	**50K**	80s	2s Clip	**Dense**	✓

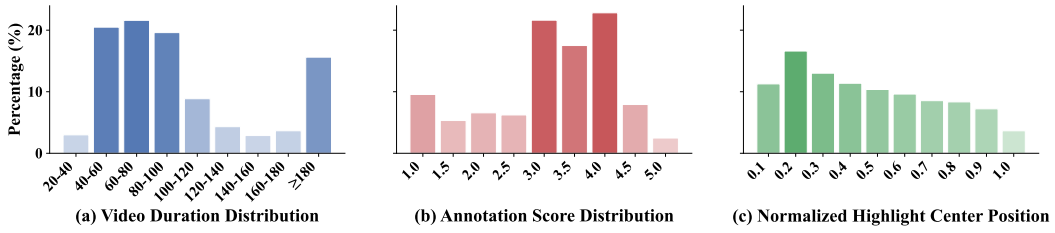

Figure 2: Data distribution of QB-Highlights.

Human-Annotated Test Set. Following QVHighlights [12], the QB-Highlights test split is densely annotated with 2-second clip-level quality scores. Each clip is rated on a 5-point Likert scale, where annotators assign a single overall quality score that holistically reflects the *Segment Qualification* criteria defined in Sec. 3, covering clarity, completeness, subject prominence, and temporal stability. Detailed annotation criteria are provided in Appendix A.2. Following the qualification protocol of TVR [36], we conduct a qualification test to screen annotators: among 100 candidate workers, the top 20% are retained for the annotation task. Each clip is then independently rated by three qualified annotators, and the final clip-level score is computed by averaging the three ratings. As shown in Fig. 2(b), the aggregated scores cover the full 1–5 range, while high-score clips form only a small

fraction. The retained annotations yield a Krippendorff's α [37] of 0.703 and an ICC [38] of 0.708, indicating substantial inter-annotator agreement.

Teacher-Assisted Training Set. As shown in Fig. 3, the training set is built with a teacher-assisted coarse-to-fine pipeline. Given a raw video and its query, Gemini-2.5 Pro [30] first scans each video at 360p to select candidate high-quality segments, and then reviews the candidates at 720p to compose a highlight set that satisfies the budgeted composition criteria. The composed highlight sets are then screened by quality and diversity filters, which remove samples with low-quality segments or redundant compositions; passing samples are assembled into $\langle$*video*, *query*, *highlight set*$\rangle$ tuples for training. The resulting highlight labels span the full video timeline (Fig. 2(c)), indicating that highlights are not biased toward any particular temporal region. For a randomly sampled 17K subset, we further use Qwen3-Max [29] to generate reasoning traces following the format defined in Sec. 5.1. Filtering details and teacher prompts are deferred to Appendices A.3 and B.2.

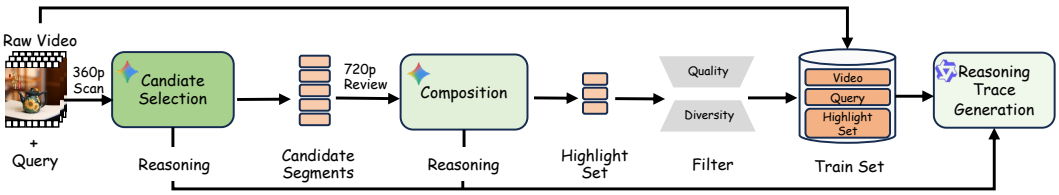

Figure 3: Data synthesis pipeline for QB-Highlights training set.

Evaluation Protocol. Unlike standard temporal grounding, which targets a unique reference span for a queried event, QB-Highlights is inherently multi-solution: in semantically redundant videos, multiple compositions of qualified segments can serve as valid highlights under the same duration budget. Fixed-reference IoU-style metrics, while informative, cannot fully capture this multi-solution nature. We therefore structure our evaluation into three complementary groups and report IoU-based results as secondary references:

- **Human-Annotated Metrics (Core):** Let $G = \{g_i\}_{i=1}^{M}$ denote the annotated clips, and let $G^\tau = \{g_i \in G \mid \text{score}(g_i) \geq \tau\}$ be the subset above threshold τ. We report *Score@Pred*, the duration-weighted average of clip-level human scores over the predicted segments, and *Prec@*τ, the fraction of predicted duration overlapping clips in G^τ (default $\tau = 4.0$; sensitivity analysis in Appendix D.2). Our primary conclusions are grounded on these human ratings.
- **Same-Origin Metrics:** We report *MD-VQA* [39] for perceptual video quality and *Compositional Diversity* (average pairwise cosine dissimilarity in CLIP feature space [40]). Because these two metrics share origins with our training-set pseudo-label filters, they are provided solely as same-origin references.
- **Third-Party Metrics:** To strictly rule out evaluation circularity, we include two external third-party metrics entirely independent of our data generation and training pipeline: **DOVER** [41] for non-reference aesthetic and technical video quality, and ***Diversity(DINO)***, computed as the average pairwise cosine dissimilarity among DINOv2 [42] embeddings of frames uniformly sampled across predicted highlight clips.

5 Method

We propose **QCR**, a framework for quality-guided budgeted video highlight detection. Unlike existing methods that score clips independently by relevance, QCR uses deliberative reasoning to first qualify candidate segments and then compose a high-quality, complementary highlight set that satisfies the target duration budget (Sec. 5.1). To align this reasoning process with the task objective, we train QCR with supervised fine-tuning on structured reasoning traces, and further refine its decision-making with GRPO under a trace-aware reward (Sec. 5.2).

5.1 Deliberative Reasoning

Given a video V, a query q, and a duration budget B, QCR generates a structured reasoning trace within a single autoregressive pass, recording both intermediate judgments and the final highlight

set, as illustrated in Fig. 4. The model first performs *segment qualification*: it scans the video to propose candidate intervals $C = \{c_j\}_{j=1}^{M}$ (with M determined by the model itself), grounds each candidate with <time> tags, and attaches an explicit quality assessment using <eval> tags. This stage produces an intermediate representation in which per-candidate evidence is made explicit, decoupling evaluation from selection. In the second stage, *budgeted composition*, QCR consolidates the qualification evidence and selects a subset $P = \{p_i\}_{i=1}^{N}$ from the qualified candidates that satisfies $\mathcal{F}(B)$, accompanied by a rationale specifying which candidates are retained or rejected. Reasoning at the set level rather than scoring segments in isolation allows QCR to balance per-segment quality with complementary coverage under the budget.

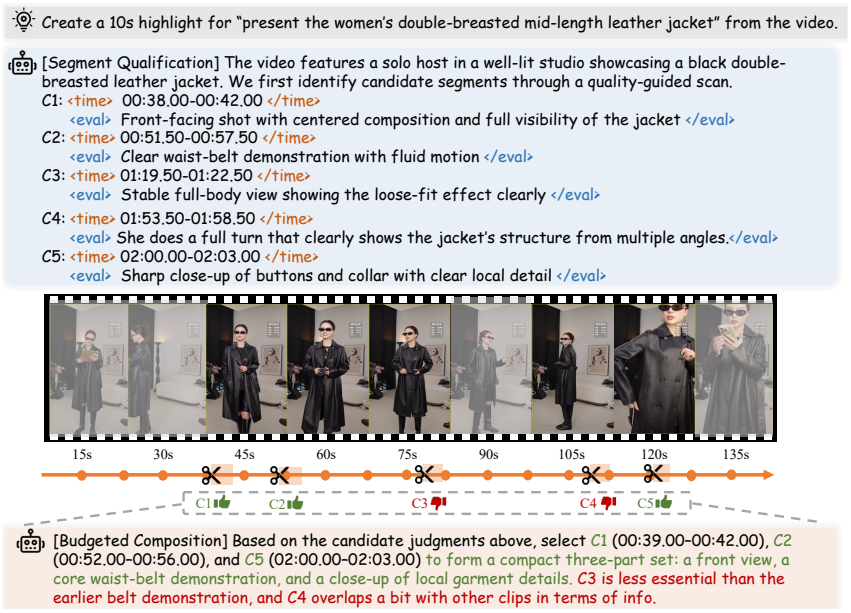

Figure 4: Overview of the QCR framework. QCR performs deliberative reasoning in two stages: segment qualification followed by budgeted composition.

5.2 GRPO with Trace-Aware Reward

To further align the reasoning policy with the objectives of quality-guided budgeted highlight detection, we refine it with GRPO using a trace-aware reward:

$$\mathcal{R} = \lambda_{\text{qual}} R_{\text{qual}} + \lambda_{\text{comp}} R_{\text{comp}} + \lambda_{\text{ground}} R_{\text{ground}},$$

Here, R_{qual} measures alignment between the prediction and high-quality reference highlights, R_{comp} encourages complementary budgeted composition, and R_{ground} rewards final predictions grounded in the intermediate candidate trace.

Multi-Segment Quality Reward. As illustrated in Fig. 5(a), R_{qual} aligns the predicted highlight set P with the reference set $G = \{g_i\}$ of high-quality segments from our training data:

$$R_{\text{qual}} = \alpha R_{\text{global}} + (1 - \alpha) R_{\text{local}}.$$

The global term measures set-level agreement via union IoU between P and G, while the local term pairs predictions with references chronologically and evaluates each pair with NGIoU [43] augmented by a boundary-offset penalty $\mathcal{P}$.

Budgeted Composition Reward. Beyond matching reference segments, R_{comp} provides a structural signal for budget-aware and complementary temporal composition. As shown in Fig. 5(b), it combines a gap term that penalizes overlapping or overly adjacent predictions with a coverage term that rewards hitting multiple temporal regions of the video. This encourages the policy to allocate the limited duration budget to complementary moments rather than overly concentrated selections.

Reasoning-Answer Grounding Reward. To ensure that the final highlight set is supported by the model's deliberative process rather than produced independently of it, R_{ground} rewards predictions

grounded in the explicitly evaluated candidates in the qualification trace (Fig. 5(c)):

$$R_{\text{ground}} = \frac{1}{N}\sum_{i=1}^{N}\left(\max_{c_j \in C}\frac{|p_i \cap c_j|}{|p_i|}\right)^{\gamma}.$$

For each predicted segment p_i, the reward measures the fraction of p_i covered by its best-matching candidate $c_j \in C$, with γ controlling the sharpness of the reward. Detailed formulations of R_{global}, R_{local}, and R_{comp} are provided in Appendix C.

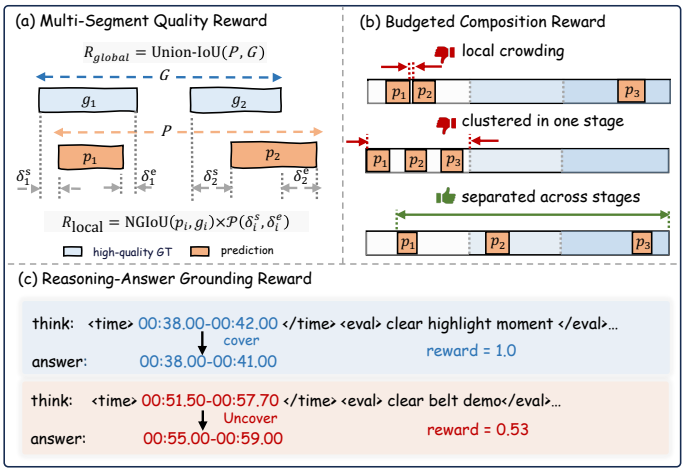

Figure 5: Trace-aware GRPO rewards. The reward jointly encourages high-quality highlight selection, complementary budgeted composition, and final decisions grounded in the deliberative candidate trace.

6 Experiments

6.1 Experimental Setup

Implementation details. We implement QCR with Qwen3-VL-8B [29] as the backbone. During both training and inference, videos are resized to 360p and sampled at 1 fps. We first perform supervised fine-tuning on 17K reasoning-trace samples for 1 epoch (lr $= 1 \times 10^{-5}$, batch size $= 128$). We then apply GRPO for 1 epoch on the full 49.1K training set (lr $= 5 \times 10^{-6}$, group size $= 8$, KL penalty $= 0.1$), with reward weights $(\lambda_{\text{qual}}, \lambda_{\text{comp}}, \lambda_{\text{ground}}) = (0.6, 0.2, 0.2)$. We set $\alpha = 0.5$ for R_{qual} and $\gamma = 1.0$ for R_{ground}.

Baselines. We consider SOTA baselines from three categories (Table 2): MLLMs, moment retrieval (MR) methods, and highlight detection (HD) methods. To ensure a fair comparison under the quality-guided budgeted highlight detection setting, all methods are evaluated with the same query and the same duration budget B. We do not impose our reasoning scaffold on baselines, preserving each method's native inference paradigm. Concretely, MLLMs directly output temporal intervals via a prompt that specifies the task requirements; MR methods are fine-tuned on the QB-Highlights training split, and their predicted temporal intervals are directly used as final highlights; HD methods produce clip-level saliency over fixed 2-second units, from which we greedily select the highest-scoring clips until budget B is reached. To ensure identical training conditions with our method, all HD baselines are fine-tuned on the QB-Highlights training set. We convert training pseudo-highlight intervals into binary clip-level saliency labels on the 2-second temporal grid (assigning 1 to pseudo-selected intervals and 0 otherwise), which directly aligns with their native saliency ranking objectives.

6.2 Results and Analysis

Quantitative Results. Table 3 shows that QCR achieves the best performance on all evaluated metrics, covering human-annotated, same-origin, and independent third-party evaluations. Relative to the best external baseline on each metric, our method improves Score@Pred by 0.268, Prec@4.0 by

Table 2: Baseline paradigm comparison. All methods are evaluated under the same query and duration budget B, while retaining their native inference paradigm.

Method	Type	Base Model	Open-source	FT on QB-Highlights	Reasoning-based	Output Form
Gemini-2.5-Pro[30]	MLLM	Gemini-2.5-Pro	×	×	✓	Temporal interval
GPT-4o[32]	MLLM	GPT-4o	×	×	✓	Temporal interval
Qwen3-VL-8B[29]	MLLM	Qwen3-VL-8B	✓	×	✓	Temporal interval
Time-R1[16]	MR	Qwen2.5-VL-7B	✓	✓	✓	Temporal interval
TRACE[44]	MR	Mistral-7B	✓	✓	✓	Temporal interval
VideoChat-T[15]	MR	Mistral-7B	✓	✓	✓	Temporal interval
QD-DETR[1]	HD	–	✓	✓	×	Saliency score
UniVTG[4]	HD	–	✓	✓	×	Saliency score
Moment-DETR[12]	HD	–	✓	✓	×	Saliency score

3.52, MD-VQA by 1.658, Compositional Diversity by 0.022, DOVER by 0.040, and Diversity(DINO) by 0.056. These improvements indicate that QCR more effectively selects high-quality and complementary moments under a fixed duration budget. Compared with its backbone Qwen3-VL-8B, QCR substantially improves Score@Pred by 1.194, Prec@4.0 by 13.85, MD-VQA by 2.358, and Compositional Diversity by 0.028, demonstrating clear gains over directly prompting the backbone. We further compare against QCR-SFT, an SFT-only variant that directly predicts highlight timestamps without deliberative reasoning or GRPO. QCR surpasses it by 0.498 on Score@Pred, 8.45 on Prec@4.0, 1.046 on MD-VQA, and 0.013 on Compositional Diversity, indicating the combined benefit of our deliberative reasoning paradigm and trace-aware GRPO over direct supervised timestamp prediction.

Table 3: Main results on QB-Highlights. Best results are shown in **bold** and second-best results are underlined. * denotes methods fine-tuned on QB-Highlights.

Method	**Human-Annotated**		**Same-Origin**		**Third-Party**	
	Score@Pred↑	Prec@4.0↑	MD-VQA↑	Comp. Div.↑	DOVER↑	Diversity(DINO)↑
MLLMs						
Qwen3-VL-8B[29]	2.313	23.82	85.812	0.091	0.601	0.198
Gemini-2.5-Pro[30]	2.954	34.15	86.371	0.092	0.638	0.214
GPT-4o[32]	2.173	24.23	86.124	0.090	0.598	0.195
MR Methods						
TRACE*[44]	2.986	29.87	86.402	0.079	0.631	0.201
VideoChat-T*[15]	3.105	33.11	86.512	0.082	0.643	0.208
Time-R1*[16]	3.239	30.22	86.348	0.070	0.647	0.210
HD Methods						
Moment-DETR*[12]	2.658	25.43	86.487	0.086	0.621	0.196
QD-DETR*[1]	2.703	26.78	86.503	0.078	0.627	0.200
UniVTG*[4]	3.198	29.04	86.412	0.097	0.638	0.216
Ours						
QCR-SFT	3.009	29.22	87.124	0.106	0.661	0.249
QCR	**3.507**	**37.67**	**88.170**	**0.119**	**0.687**	**0.272**

Qualitative Results. Figure 6 compares the highlight detection behaviors of different methods. Moment-DETR predicts intervals without explicit reasoning and only partially aligns with human-annotated high-quality intervals. Gemini-2.5-Pro [30] and Time-R1 [16] provide plausible but flawed rationales, yet their timestamp selections are not consistently aligned with the high-quality intervals and include less informative content. In contrast, QCR leverages deliberative reasoning to qualify candidates and compose a higher-quality, more informative highlight within the budget.

Extended Evaluation. For compatibility with standard temporal grounding protocols, we derive deterministic reference intervals on QB-Highlights from human annotations (Appendix A.4) and report IoU-based metrics. As shown in Table 4, QCR achieves the best results on all **QB-Highlights** metrics, confirming strong alignment with conventional localization criteria. Fine-tuning on QVHighlights [12] further yields 64.38 R1@0.5 and 50.10 R1@0.7, outperforming Moment-DETR and QD-DETR and remaining competitive with UniVTG (within a 1.05-point gap). These results indicate that our deliberative reasoning paradigm and trace-aware GRPO **transfer** beyond QB-Highlights to standard highlight detection benchmarks.

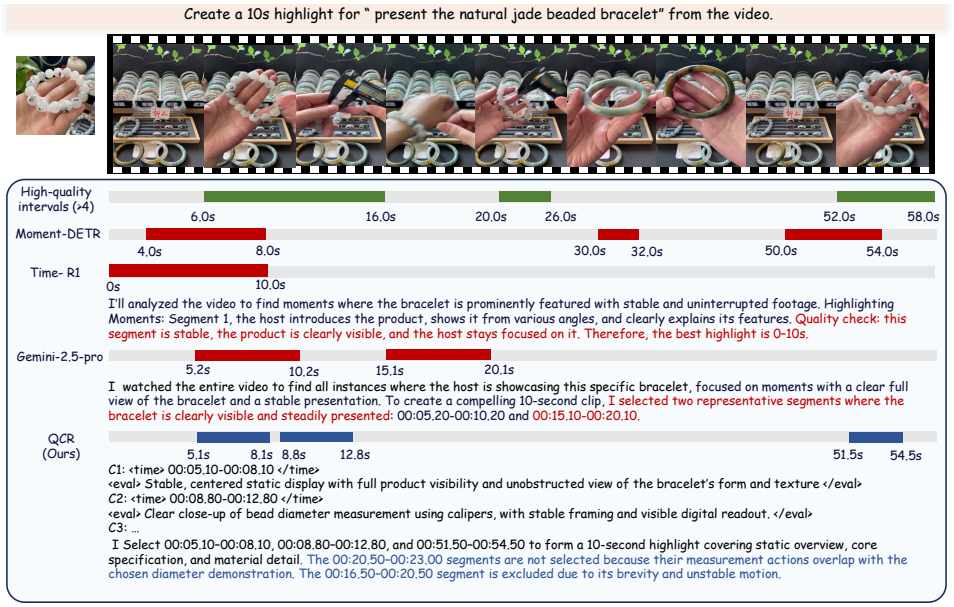

Figure 6: Qualitative comparison of highlight detection results on QB-Highlights.

Table 4: Evaluation using standard temporal grounding metrics on QB-Highlights (left) and re-trained results on the highlight detection benchmark QVHighlights (right).

Method	QB-Highlights R1@0.3↑	R1@0.5↑	mIoU↑	QVHighlights R1@0.5↑	R1@0.7↑
Qwen3-VL-8B[29]	13.25	7.09	8.65	–	–
Gemini-2.5-Pro[30]	<u>19.79</u>	<u>11.74</u>	<u>12.80</u>	–	–
GPT-4o[32]	18.62	10.31	11.15	–	–
TRACE[44]*	12.84	6.42	7.91	–	–
VideoChat-T*[15]	7.92	3.10	5.55	–	–
Time-R1*[16]	16.06	8.01	9.60	–	–
Moment-DETR*[12]	14.38	8.19	8.24	52.89	33.02
QD-DETR*[1]	11.12	7.21	6.94	63.06	45.10
UniVTG*[4]	19.31	11.64	11.29	**65.43**	<u>50.06</u>
QCR	**21.51**	**14.17**	**13.00**	<u>64.38</u>	**50.10**

6.3 Ablation Study

Framework Design Ablations. As shown in Table 5, removing GRPO and using only supervised fine-tuning (*w/o* GRPO) leads to a 6.73 drop in Prec@4.0, indicating that GRPO with the trace-aware reward improves the reliability of the deliberative reasoning process. Replacing deliberate reasoning with direct budgeted composition (*w/o* Segment Qualification), i.e., removing segment qualification, reduces Prec@4.0 by 3.85 and Composition Diversity by 0.007, confirming the importance of explicit candidate qualification before selecting the final budgeted highlight set.

Reward term ablations. Among the reward components, the Multi-Segment Quality Reward has the largest impact on highlight quality, with its removal causing a 0.379 drop in Score@Pred and a 6.10 drop in Prec@4.0. The Budgeted Composition Reward mainly influences set-level structure, and removing it decreases Compositional Diversity by 0.010. Removing the Reasoning-Answer Grounding Reward degrades all metrics, underscoring the importance of grounding the final prediction in the intermediate reasoning trace. To verify the rewards provide effective optimization signals, we plot the three reward components in Fig. 7, all rewards increase steadily during training.

Table 5: Ablation results on QB-Highlights. The upper part compares different framework designs, and the lower part compares the contribution of individual reward components.

Method	Score@Pred↑	Prec@4.0↑	MD-VQA↑	Compositional Diversity↑
w/o GRPO	3.094	30.94	87.923	0.107
w/o Segment Qualification	3.256	33.82	88.051	0.112
w/o $\mathcal{R}_{\text{qual}}$	3.128	31.57	87.964	0.110
w/o $\mathcal{R}_{\text{comp}}$	3.382	36.43	88.101	0.109
w/o $\mathcal{R}_{\text{ground}}$	3.337	33.38	88.072	0.113
QCR (Full)	**3.507**	**37.67**	**88.170**	**0.119**

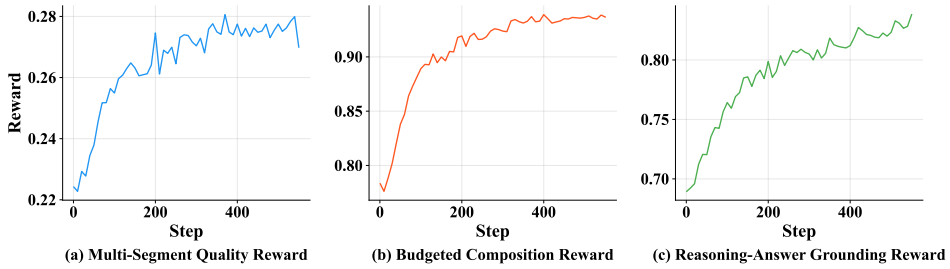

Figure 7: Training curves of different reward components.

7 Conclusion

We reformulate video highlight detection under densely occurring query-relevant events as quality-guided, budget-constrained selection, and introduce QB-Highlights, a benchmark dedicated to this setting. We further propose QCR, a framework that first qualifies candidate segments and then composes a budget-matched highlight set through deliberative reasoning, optimized with GRPO using a trace-aware reward. Experiments show consistent gains over strong baselines, highlighting the value of modeling quality and compositional diversity beyond relevance. **Broader Impact:** Our method has been deployed in the recommendation feed of a large e-commerce platform, where online A/B tests confirm consistent gains in user engagement and business metrics, demonstrating its practical value in real-world applications.

References

[1] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023.

[2] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3042–3051, 2022.

[3] Yicheng Xiao, Zhuoyan Luo, Yong Liu, Yue Ma, Hengwei Bian, Yatai Ji, Yujiu Yang, and Xiu Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18709–18719, 2024.

[4] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2794–2804, 2023.

[5] Jing Liu, Zhuo He, Weizhi Nie, Zongbing Zhang, and Yuting Su. What and where: Semantic grasping and contextual scanning for moment retrieval and highlight detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[6] David Pujol-Perich, Sergio Escalera, and Albert Clapés. Sparse-dense side-tuner for efficient video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21515–21524, 2025.

[7] Xun Jiang, Liqing Zhu, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Query as supervision: Toward low-cost and robust video moment and highlight retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):3955–3968, 2024.

[8] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.

[9] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014.

[10] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer, 2014.

[11] Yitian Yuan, Lin Ma, and Wenwu Zhu. Sentence specified dynamic video thumbnail generation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2332–2340, 2019.

[12] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[14] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1984–1990, 2019.

[15] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024.

[16] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025.

[17] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. 2019.

[18] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 200–216, 2018.

[19] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1247–1257, 2019.

[20] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12870–12877, 2020.

[21] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8211–8225, 2020.

[22] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020.

[23] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 2046–2065, 2020.

[24] Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. A hierarchical multi-modal encoder for moment localization in video corpus. *arXiv preprint arXiv:2011.09046*, 2020.

[25] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[26] Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Zien Xie, Youyao Jia, and Sidan Du. Multi-modal fusion and query refinement network for video moment retrieval and highlight detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[27] Runhao Zeng, Jiaqi Mao, Minghao Lai, Minh Hieu Phan, Yanjie Dong, Wei Wang, Qi Chen, and Xiping Hu. Ovg-hq: Online video grounding with hybrid-modal queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21085–21096, 2025.

[28] Donghoon Han, Seunghyeon Seo, Eunhwan Park, Seong-Uk Nam, and Nojun Kwak. Unleash the potential of clip for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8275–8279, 2024.

[29] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[31] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

[33] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv e-prints*, pages arXiv–2503, 2025.

[34] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[35] Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. Video-thinker: Sparking" thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025.

[36] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020.

[37] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4th edition, 2018.

[38] Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[39] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1746–1755, 2023.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[41] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20087–20097. IEEE, 2023.

[42] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[43] Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, et al. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *arXiv preprint arXiv:2505.20715*, 2025.

[44] Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. In *The Thirteenth International Conference on Learning Representations*.

[45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[46] Donggyu Lee, Youngbin Ki, Jeonghun Kang, and Taehwan Kim. Svhighlights: Towards extremely long sport video highlight detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 9254–9265, 2026.

A Benchmark Analysis

A.1 Category Distribution

Figure 8 shows the distribution of target product categories in the queries of QB-Highlights. QB-Highlights covers major e-commerce domains, including Clothing (29.8%), Living (22.1%), Accessories (21.2%), and Shoes&Bags (9.8%), reflecting the diversity of practical e-commerce content.

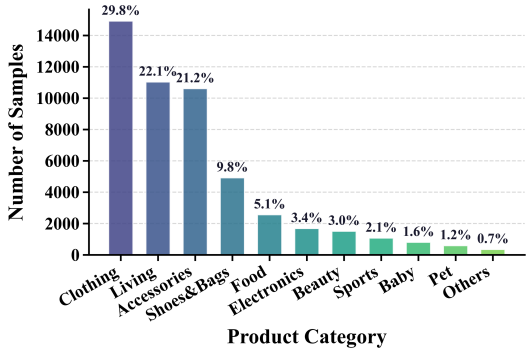

Figure 8: Distribution of QB-Highlights by product category.

A.2 Test Set Annotation Protocol

For the human-annotated test set, each video is divided into non-overlapping 2-second clips, and every clip is independently rated on a 1–5 scale. Following the Segment Qualification criteria defined in Sec. 3, annotators assess each clip along four dimensions:

- **Clarity:** whether the content relevant to the target query is visually clear and free from severe blur, low resolution, or other visual distortions.
- **Completeness:** whether the content relevant to the target query is fully captured within the frame, without significant truncation or occlusion.
- **Subject prominence:** whether the content relevant to the target query is the visual focus of the clip, sufficiently salient rather than marginal or overwhelmed by distractors.
- **Temporal stability:** whether the clip is temporally stable and continuous, free from severe shake, abrupt cuts, or rapid scene changes.

Annotators assign a single overall quality score that holistically reflects the above criteria, indicating the suitability of the 2-second clip as a highlight candidate:

- **5 points:** Excellent highlight candidate. The content relevant to the target query is clearly visible, complete, and prominent, and the clip is temporally stable and visually clean.
- **4 points:** Good highlight candidate. The content relevant to the target query is well presented across all four dimensions, with only minor visual imperfections or slight instability.
- **3 points:** Usable but imperfect candidate. The content relevant to the target query is identifiable, but the clip exhibits moderate issues in one or more dimensions, such as limited prominence, partial occlusion or truncation, or mild instability.
- **2 points:** Weak candidate. The content relevant to the target query appears only briefly, unclearly, or non-prominently, or the clip suffers from substantial issues such as severe occlusion, truncation, or instability.
- **1 point:** Unsuitable as a highlight candidate. The content relevant to the target query is absent, barely visible, severely obstructed or truncated, or the clip is visually unusable.

A.3 Training Set Filtering

We apply automatic post-filtering to the teacher-generated pseudo highlights along the quality and diversity criteria defined in Sec. 3. For *clarity* and *temporal stability*, we retain only samples whose selected segments have an average MD-VQA [39] score above 80, as MD-VQA jointly captures low-level perceptual quality and temporal smoothness. For *completeness*, we use Grounding DINO [45] to localize the target product and discard samples whose bounding box touches the image boundary for more than 50% of the highlight duration, as such cases typically indicate truncated or partially-visible products. For *subject prominence*, we further discard samples whose target bounding box is smaller than a category-dependent area threshold throughout the highlight, where the threshold is set to 10% of the frame area for regular products and 5% for small-object categories such as accessories and jewelry.

To additionally suppress redundant compositions at the set level, we uniformly sample three frames across the concatenated timeline of each composed highlight, extract CLIP [40] features, and discard samples whose average pairwise CLIP similarity exceeds 0.91, indicating insufficient visual-semantic diversity across the selected segments. As shown in Table 6, the filtered labels achieve a high human acceptance rate together with favorable automatic scores, indicating that the resulting training set provides supervision of sufficient quality for large-scale training.

Table 6: Human acceptance check of filtered training pseudo labels on 500 randomly sampled instances from the QB-Highlights training split.

Source	Acceptance Rate↑	MDVQA↑	Composition Diversity↑
Filtered training pseudo labels	84.3%	87.24	0.107

A.4 Construction of Reference Highlights for IoU-based Evaluation

QB-Highlights is designed for quality-aware highlight composition under a fixed duration budget, where multiple valid solutions may exist for the same video-query pair. That is, the target highlight is inherently *multi-solution*: different combinations of high-quality segments can satisfy the budget while preserving strong semantic value. However, conventional temporal grounding metrics such as R1@0.3, R1@0.5, and mIoU require a single fixed reference target.

To enable supplementary IoU-based evaluation, we construct a single deterministic reference highlight for each video–query pair as follows. We first retain 2-second clips with human scores no lower than 3.0 as a high-quality candidate pool, and form candidate segments by taking all contiguous spans of adjacent qualified clips with duration between 2 and 6 seconds. We then enumerate all feasible combinations whose union duration matches the evaluation budget, and select the optimal one by maximizing

$$S^* = \arg\max_S \mathcal{U}(S) = \hat{Q}(S) + \hat{D}(S),$$

where $\hat{Q}(S)$ is the duration-weighted average of clip-level human scores over the 2-second clips covered by S, and $\hat{D}(S)$ measures content diversity via CLIP-feature dissimilarity by uniformly sampling three frames across the concatenated timeline of S and computing the average pairwise cosine dissimilarity of their CLIP embeddings. Both terms are min–max normalized to $[0, 1]$ across all feasible combinations so that quality and diversity contribute on a comparable scale, jointly favoring references composed of high-scoring yet visually distinct moments. We emphasize that this construction is introduced solely for compatibility with conventional grounding metrics, and does not imply that QB-Highlights has a unique ground-truth solution.

Prompt Template for Training and Evaluation

User Prompt: Your task is to create a 10-second highlight video clip for {query} from the {video}. High-quality segments must simultaneously satisfy:
① The target query content appears clearly, fully, and without obstruction;
② The shot is stable and continuous with no cuts or transitions;
③ The segment carries substantive query-relevant visual content with high information density.
When combining multiple segments, ensure they cover different angles or states to avoid redundancy;
Reason step by step inside the <think> tag using the following structure:
[Quality-guided Global Scan]
Briefly describe the overall video pattern in 1–2 sentences. Then scan for candidate segments and evaluate each one:
C1: <time> MM:SS.xx-MM:SS.xx </time>
<eval> One concise sentence on clip quality. </eval>
C2: <time> MM:SS.xx-MM:SS.xx </time>
<eval> One concise sentence on clip quality. </eval>
(add more Ci as needed)
[Budgeted Set Composition]
In 1–3 sentences, explain which clips are selected to form the final 10-second highlight and why unselected clips are excluded.
Then output a JSON inside the <answer> tag. Each segment in the clips array must include: clip_id, timestamp (format MM:SS.xx-MM:SS.xx), description, and target query.

Prompt for Candidate Selection in Training Data Construction

Identify 1--10 high-quality candidate clips from the input video for downstream highlight composition. Select clips whose visual content directly fulfills the {query}, considering the entire {video} when making your choice. The goal is to extract candidate segments that are visually strong, semantically aligned with the query, and suitable for later composition into a concise highlight. Use visual information only and ignore all audio, including speech, music, and sound effects. Determine the query-relevant content from the provided query together with the visual content of the video. Each selected clip should satisfy the following quality requirements:

1. Query relevance as the visual focus. The clip should clearly fulfill the query, with the query being actively shown as the main visual content.
2. Clear and complete visibility. The query-relevant content should remain clearly and completely visible throughout the clip, without major occlusion or poor framing.
3. Continuous shot. The clip should come from a single continuous shot, without cuts, abrupt transitions, or severe content changes.
4. Strong visual quality. The clip should be stable, sharp, and well lit, with minimal blur, shake, or distracting background content.
5. Reasonable clip length. Each clip should be 3--10 seconds long.
6. Content diversity across clips. Different clips should capture distinct visual segments rather than adjacent, overlapping, or repetitive moments. Favor clips that provide complementary coverage of the query across the video

The output JSON should include three parts: (1) an analysis of the screening and selection process, explaining how the query is interpreted and how clips are selected; (2) the timestamps of the selected candidate clips; and (3) a description of the visual content of each selected clip.

B Prompt Templates

B.1 Train and Evaluation

B.2 Data Construction

C Trace-Aware GRPO Reward Details

This appendix provides the concrete definitions of R_{global}, R_{local}, and R_{comp} used in Sec. 5.2. The overall reward weights and the hyperparameters α and γ are given in Sec. 6.

Global Quality Term. Given the predicted highlight set $P = \{p_i\}_{i=1}^{N}$ and the reference set $G = \{g_j\}_{j=1}^{M}$, the global quality term is defined as the union IoU:

$$R_{\text{global}} = \frac{\left|\left(\bigcup_{i=1}^{N} p_i\right) \cap \left(\bigcup_{j=1}^{M} g_j\right)\right|}{\left|\left(\bigcup_{i=1}^{N} p_i\right) \cup \left(\bigcup_{j=1}^{M} g_j\right)\right|}.$$

This term measures set-level agreement between predicted and reference highlight regions.

Local Quality Term. For local alignment, predicted and reference segments are sorted chronologically and paired by index. Let $L = \max(N, M)$. Unmatched predictions or references contribute zero. For each valid pair (p_i, g_i), where $p_i = [t^s_{p_i}, t^e_{p_i}]$ and $g_i = [t^s_{g_i}, t^e_{g_i}]$, the local score is

$$r_i^{\text{local}} = \text{NGIoU}(p_i, g_i) \cdot \exp\left[-\kappa(\delta_s^i + \delta_e^i)\right],$$

where $\text{NGIoU}(\cdot,\cdot)$ follows [43] and $\kappa = 1.0$. The local reward is

$$R_{\text{local}} = \frac{1}{L}\sum_{i=1}^{L} r_i^{\text{local}}.$$

Prompt for Composition in Training Data Construction

Select and organize the best clips from the {candidate segments} to form a compact 10-second highlight. The final compositio n should be visually strong, tightly paced, and consistently fulfill the {query} throughout. Use visual information only and ign ore all audio. All selected clips must fulfill the same query. Each selected clip should satisfy the following quality requirement s:

1. Query consistency. All clips must fulfill the same query, and the query-relevant content should be clearly visible from the beginning of each clip.
2. Sustained quality throughout. Every selected clip and the entire composition should consistently satisfy strong visual quality, including stability, sharpness, lighting, clear visibility of the subject, and absence of blur, shake, or distracting background content.
3. High information density. Each selected clip should carry substantive query-relevant visual content, avoiding static, idle, or low-information moments where little informative change occurs.
4. Content diversity. The selected clips should differ in viewpoint, action, framing, or information content while maintaining high overall quality.
5. Logical composition. The clip order should balance quality, diversity, and viewing coherence. Clips may be reordered from their original temporal sequence, but they must not overlap or repeat.
6. Duration control. The final composition should be approximately 10 seconds long. Internal trimming within a candidate clip is allowed, as long as the extracted segment remains continuous and natural.

Output three parts: 1. An analysis of the screening and selection process, explaining how the final clips are chosen and balanced for quality, diversity, and temporal structure. 2. The timestamps of the selected clips, including the final trimmed time ranges used from the original video. 3. A description of each selected clip, covering the visual content, action, and framing relevant to the query.

Prompt for Reasoning Trace Generation in Training Data Construction

Under the results of {candidate selection} and {composition}, rewrite the original intermediate reasoning into a concise and s tandardized reasoning trace. The rewritten trace should be compact, structured, and method-like, with minimal narrative style. Output a JSON object with three fields:

1. scan_overview: summarize the overall visual pattern of the video in 1--2 concise sentences, focusing on the dominant vi sual style, camera setup, and the typical way content relevant to the target query appears across the video. Do not make f inal selection judgments here.
2. segment_qualification: for each candidate clip from the candidate selection, provide its timestamp and a brief reasoning sentence grounded only in observable visual properties such as clarity, stability, framing, visibility of the target query co ntent, and information density. Do not compare clips against each other or make any keep/remove decision in this field.
3. budgeted_composition: summarize the final set-level reasoning under the 10-second budget using the fine-stage composi tion result. Explain why the selected clips form an effective final highlight and why each unselected candidate is exclude d, using concrete reasons such as overlap, weaker framing, lower visual quality, lower information density, or limited ad ditional content relevant to the target query.

Keep the language concise and structured. Separate clip-level quality assessment from final budget-aware composition reasoni ng. Avoid long narrative monologues, subjective storytelling, and unsupported candidate invention. Preserve the core reasonin g logic of the coarse and fine stages while rewriting it.

The boundary offsets are normalized by the duration of the reference segment:

$$\delta_s^i = \frac{|t_{p_i}^s - t_{g_i}^s|}{|g_i|}, \qquad \delta_e^i = \frac{|t_{p_i}^e - t_{g_i}^e|}{|g_i|}.$$

Budgeted Composition Term. The composition reward combines a local anti-crowding term and a coarse stage-coverage term:

$$R_{\text{comp}} = \beta R_{\text{gap}} + (1-\beta) R_{\text{cov}},$$

with $\beta = 0.5$.

Given chronologically sorted predictions $p_i = [t_i^s, t_i^e]$, the gap term is

$$R_{\text{gap}} = \begin{cases} 0, & N \le 1, \\ \frac{1}{N-1} \sum_{i=1}^{N-1} \min\left(\frac{\Delta_i}{\Delta_{\text{target}}}, 1\right), & N > 1, \end{cases}$$

where

$$\Delta_i = \max(0,\, t_{i+1}^s - t_i^e), \qquad \Delta_{\text{target}} = 0.2 \cdot T_{\text{video}}.$$

This term penalizes overlapping or overly adjacent predictions and saturates once the adjacent gap reaches the target threshold.

For stage coverage, we divide the video into K uniform temporal bins and compute the fraction of bins intersected by at least one predicted segment:

$$R_{\text{cov}} = \frac{|\mathcal{B}_P|}{K}, \qquad K = 3,$$

where

$$\mathcal{B}_P = \{b \mid p_i \cap I_b \neq \emptyset \text{ for some } p_i \in P\},$$

and I_b denotes the b-th temporal bin of the video. This term encourages the selected highlight set to cover different temporal stages when multiple qualified moments are available.

D Additional Evaluation

D.1 Bucket Analysis

By category. As shown in Table 7, QCR shows generally stable performance across different product categories. It obtains relatively high Prec@4.0 on *Consumer Electronics* (44.02), *Jewelry & Accessories* (42.71), and *Home & Living* (41.44), while most categories achieve comparable Score@Pred values around the overall average. Performance is lower on *Clothing* and *Pet Supplies*, suggesting that some categories remain more challenging.

By duration. Table 8 presents the duration-based analysis. QCR performs worse on very short videos (20–40s), where the limited candidate pool restricts budget-aware composition. Its performance improves on medium and moderately long videos, achieving the best Prec@4.0 in the 120–140s bucket (47.84). The model also remains stable on long videos ($\geq$180s), reaching Prec@4.0 = 41.18. These results suggest that QCR benefits from selecting among multiple qualified candidate moments and maintains robust performance as video duration increases.

Table 7: Bucket analysis of QCR by product category.

Category	Count	Score@Pred↑	Prec@4.0↑	MD-VQA↑	Compositional Diversity↑
Home & Living	177	3.500	41.44	87.990	0.103
Clothing	231	3.461	33.81	87.010	0.132
Shoes & Bags	69	3.570	30.79	89.750	0.132
Beauty & Personal Care	50	3.525	39.77	88.600	0.098
Consumer Electronics	59	3.548	44.02	89.190	0.099
Food & Beverage	36	3.555	33.53	89.360	0.086
Sports & Outdoors	38	3.656	36.32	91.900	0.108
Mother & Baby	66	3.453	35.61	86.810	0.129
Pet Supplies	10	3.449	25.74	86.700	0.124
Jewelry & Accessories	153	3.520	42.71	88.490	0.135
Others	14	3.456	30.64	86.870	0.116
Total	903	3.507	37.67	88.170	0.119

Table 8: Bucket analysis of QCR by video duration.

Duration	Count	Score@Pred↑	Prec@4.0↑	MD-VQA↑	Compositional Diversity↑
20–40	27	3.094	25.82	77.780	0.108
40–60	185	3.299	33.87	82.950	0.111
60–80	195	3.462	35.44	87.040	0.109
80–100	177	3.592	40.95	90.300	0.122
100–120	80	3.589	37.57	90.240	0.135
120–140	39	3.794	47.84	95.390	0.111
140–160	26	3.720	38.90	93.530	0.114
160–180	33	3.438	36.42	86.420	0.134
$\geq$180	141	3.667	41.18	92.180	0.134
Total	903	3.507	37.67	88.170	0.119

D.2 Sensitivity to Metric Threshold

As shown in Table 9, QCR consistently ranks first across all threshold choices under Prec@τ. This indicates that our main conclusion is robust to the threshold choice, and that the results are not driven by a narrowly tuned cutoff.

D.3 Cross-Domain Evaluation

To validate the generalizability of our formulation and method beyond the e-commerce domain, we construct a zero-shot sports evaluation set using videos from SVHighlights [46]. The evaluation set covers 8 distinct sports categories: American football, baseball, basketball, ice hockey, racing, rugby,

Table 9: Threshold sensitivity of Prec@τ. Best results are shown in **bold** and second-best results are underlined. * denotes methods fine-tuned on QB-Highlights.

Method	Prec@3.0	Prec@3.5	Prec@4.0	Prec@4.5
Qwen3-VL-8B[29]	51.64	35.66	23.83	7.71
Gemini-2.5-Pro[30]	67.34	48.65	34.15	12.55
GPT-4o[32]	50.68	35.53	24.24	7.93
TRACE*[44]	67.68	47.84	29.87	10.12
VideoChat-T*[15]	69.02	49.10	33.11	10.65
Time-R1*[16]	70.16	46.32	30.22	9.35
Moment-DETR*[12]	56.57	38.76	25.43	7.93
QD-DETR*[1]	58.65	40.29	26.78	9.30
UniVTG*[4]	63.31	41.83	29.04	10.45
QCR-SFT	61.36	43.06	29.22	11.02
QCR	**79.11**	**55.51**	**37.67**	**14.21**

soccer, and volleyball. We sample 6 videos per category (48 videos in total) and trim each video to 80 seconds to match the average duration of QB-Highlights. For each category, we define a query targeting densely recurring actions (e.g., *"a player attempting a shot at the basket"* for basketball) and set the target highlight budget to 10 seconds.

We annotate this set following the exact test protocol of QB-Highlights (3 annotators per clip, 2-second clip units, 5-point Likert scale across clarity, completeness, prominence, and stability). All models are evaluated zero-shot without target-domain fine-tuning. As shown in Table 10, while the fast-paced and dynamic nature of sports videos leads to an overall drop across all methods in quality metrics alongside higher compositional diversity, QCR consistently achieves the top performance on all four metrics. This confirms that QCR's deliberative reasoning paradigm captures general quality-guided selection principles rather than domain-specific visual shortcuts.

Table 10: Zero-shot cross-domain evaluation on the constructed sports test set from SVHighlights (48 videos across 8 categories). * denotes methods fine-tuned on QB-Highlights.

Method	Score@Pred↑	Prec@4.0↑	MD-VQA↑	Compositional Diversity↑
Gemini-2.5-Pro [30]	2.57	26.8	82.90	0.129
UniVTG [4]	2.41	21.8	80.87	0.121
QD-DETR [1]	2.38	20.1	80.40	0.105
Time-R1* [16]	2.44	23.2	81.05	0.098
VideoChat-T* [15]	2.46	24.7	81.20	0.102
QCR (Ours)	**3.04**	**30.9**	**83.10**	**0.141**

D.4 Cross-Benchmark Zero-Shot Transfer

To investigate whether the representations learned on QB-Highlights transfer effectively to standard highlight detection benchmarks, we evaluate QCR zero-shot on the QVHighlights [12] test split without any target-domain fine-tuning. We compare against (i) the base Qwen3-VL-8B model without QB-Highlights training and (ii) UniVTG transferred zero-shot from its original training corpus (Ego4D + VideoCC + pseudo labels).

As shown in Table 11, training on QB-Highlights yields substantial zero-shot improvements over the untrained base model (+22.68 on R1@0.5 and +19.02 on R1@0.7). Furthermore, QCR outperforms UniVTG by 12.82 on R1@0.5 and 15.87 on R1@0.7 under zero-shot transfer, indicating that the quality and composition supervision from QB-Highlights provides strong transferability across highlight detection benchmarks.

D.5 Statistical Robustness and Confidence Intervals

To verify the statistical robustness of our results, we compute paired bootstrap 95% confidence intervals (CIs) over 1,000 resamples of the 0.9K-video test set for the primary evaluation metrics. As

Table 11: Zero-shot transfer evaluation on QVHighlights test split.

Method	Training Data	R1@0.5↑	R1@0.7↑
QCR w/o QB-Highlights training	–	15.30	5.80
UniVTG [4]	Ego4D + VideoCC + pseudo labels	25.16	8.95
QCR w/ QB-Highlights training	**QB-Highlights**	**37.98**	**24.82**

shown in Table 12, the 95% CIs of QCR do not overlap with those of the best-performing baselines across all metrics. Furthermore, paired statistical tests confirm that the performance gains of QCR over the strongest baseline for each metric are statistically significant at $p < 0.001$. This demonstrates that the observed improvements are stable and not driven by test set sampling variance.

Table 12: Bootstrap 95% confidence intervals (1,000 resamples on the test set). For each metric, the strongest baseline is reported for direct comparison. * denotes methods fine-tuned on QB-Highlights.

Method	Score@Pred	Prec@4.0	MD-VQA	Compositional Diversity
Strongest baseline per metric:				
Time-R1* [16]	3.239 [3.184, 3.294]	–	–	–
Gemini-2.5-Pro [30]	–	34.15 [32.18, 35.24]	–	–
VideoChat-T* [15]	–	–	86.512 [86.203, 86.821]	–
UniVTG* [4]	–	–	–	0.097 [0.093, 0.101]
Ours:				
QCR-SFT	3.009 [2.957, 3.061]	29.22 [27.27, 31.17]	87.124 [86.815, 87.433]	0.106 [0.102, 0.110]
QCR	**3.507 [3.460, 3.558]**	**37.67 [35.39, 39.86]**	**88.170 [87.863, 88.483]**	**0.119 [0.115, 0.124]**
Δ (QCR vs. best)	**+0.268 [0.201, 0.335]**	**+3.52 [1.38, 5.61]**	**+1.658 [1.298, 2.026]**	**+0.022 [0.017, 0.027]**
p-value	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$

E Limitations

QB-Highlights currently focuses on e-commerce livestreams, a representative content-dense domain where budgeted highlight generation has substantial practical value. While this setting reflects an important real-world application, extending the benchmark and evaluation to other content-dense video domains, such as sports broadcasts and instructional videos, remains a valuable direction for future work.

F Compute Resources

All experiments are conducted on NVIDIA H20 GPUs with bf16 mixed-precision training, DeepSpeed ZeRO-3, and gradient checkpointing. Our base model is Qwen3-VL-8B, trained on the proposed QB-Highlights dataset.

G Code and Data Availability

Ethics, Privacy & Takedown. Videos are sourced from platform-hosted livestreams under compliance review (adhering to GDPR/CCPA), where creators are notified of research usage via platform terms. Sensitive metadata is anonymized. We maintain an active takedown policy: creators may request the removal of their footage at any time, upon which corresponding entries will be promptly purged.

DUA-Gated Raw Video Access. To prevent misuse, raw videos will not be publicly hosted, but made accessible solely for non-commercial academic research under a signed Data Usage Agreement (DUA). The DUA strictly prohibits redistribution, commercial use, and any form of biometric/facial identification or tracking of appearing individuals.

Release Scope. All core metadata, annotations, pseudo-labels, prompts, reasoning traces, and code are released under the ODC-BY license at `https://github.com/gclonghorn/QB-Highlights.git`.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's four main contributions: the task formulation, QB-Highlights dataset, QCR framework, and extensive experiments. These claims are supported by the task formulation in Sec. 3, benchmark description in Sec. 4, method section in Sec. 5, and experimental results in Sec. 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in Appendix E.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: Our paper focuses on empirical contributions (a new problem formulation, a dataset, and a learning framework) and does not include formal theoretical results or proofs.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide reproduction details for the main experiments, including the dataset construction pipeline and evaluation metrics in Sec. 4, the QCR architecture and reasoning scaffold in Sec. 5, the SFT/GRPO training procedure, hyperparameters, and training schedules in Sec. 6.1, reward formulations in Appendix C, and prompts in Appendix B.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: See Appendix G for the release scope, anonymous repository link, and data access protocol.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We report training and evaluation details including data splits (Sec. 4), learning rate, batch size, number of training steps, GRPO group size, and reward weights $\lambda_{\text{qual}}, \lambda_{\text{comp}}, \lambda_{\text{ground}}$ in Sec. 6.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report paired bootstrap 95% confidence intervals over 1,000 resamples and statistical significance tests ($p < 0.001$) in Appendix D.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the compute resources used for both SFT and GRPO stages, including GPU type, memory configuration, and distributed setup in Appendix F.

   Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics in all respects.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: We discuss broader impacts in Sec. 7. QCR improves user engagement and reduces editing cost in real-world deployment, while potential risks common to content selection systems are mitigated through platform-level moderation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [Yes]

Justification: To protect privacy and prevent misuse, raw videos are not hosted via public download and are strictly gated under a Data Usage Agreement (DUA) that explicitly prohibits biometric/facial identification and commercial use (see Appendix G).

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite all baseline methods, pre-trained backbones, and external datasets used in our experiments. License terms of all open-source assets are respected.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: QB-Highlights and QCR are fully documented in the paper, and will be released with a datasheet and model card covering collection/training details, licensing, and limitations (see Appendix G).

    Guidelines:

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

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

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

Answer: [Yes]

Justification: Human annotators participated in the construction of QB-Highlights. We provide annotation guidelines and the quality criteria shown to annotators in Appendix A.2. Annotators were paid at or above the local minimum wage.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The annotation task involves only public livestream content without sensitive data or human-subject interventions, and does not require IRB review under our institution's regulations.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: QCR is built on an MLLM backbone, with an LLM-based judge used for dataset filtering and scoring (see Sec. 4, Sec. 6.1, and Appendix B).

Guidelines:

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