QEVA: A Reference-Free Evaluation Metric for Narrative Video Summarization with Multimodal Question Answering

Woojun Jung and **Junyeong Kim**[*]
Department of Artificial Intelligence, Chung-Ang University
{svvma91, junyeongkim}@cau.ac.kr

Abstract

Video-to-text summarization remains underexplored in terms of comprehensive evaluation methods. Traditional n-gram overlap-based metrics and recent large language model (LLM)-based approaches depend heavily on human-written reference summaries, limiting their practicality and sensitivity to nuanced semantic aspects. In this paper, we propose QEVA, a reference-free metric evaluating candidate summaries directly against source videos through multimodal question answering. QEVA assesses summaries along three clear dimensions: Coverage, Factuality, and Chronology. We also introduce MLVU(VS)-Eval, a new annotated benchmark derived from the MLVU dataset, comprising 800 summaries generated from 200 videos using state-of-the-art video-language multimodal models. This dataset establishes a transparent and consistent framework for evaluation. Experimental results demonstrate that QEVA shows higher correlation with human judgments compared to existing approaches, as measured by Kendall's τ_b, τ_c, and Spearman's ρ. We hope that our benchmark and metric will facilitate meaningful progress in video-to-text summarization research and provide valuable insights for the development of future evaluation methods.[1]

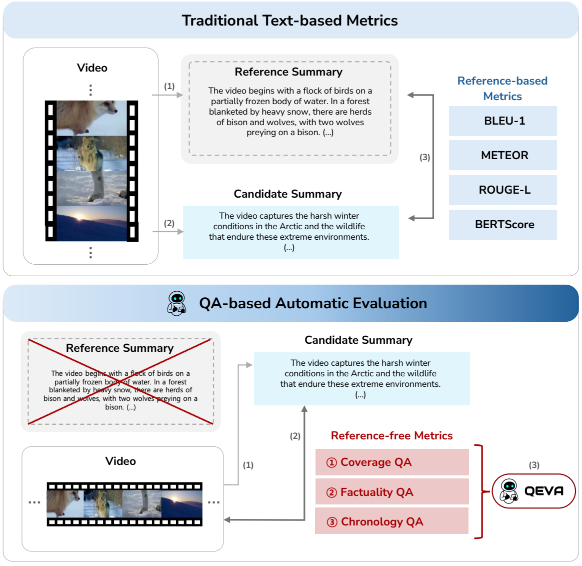

Traditional Text-based Metrics
Video
(1)
(2)
Reference Summary
The video begins with a flock of birds on a partially frozen body of water. In a forest blanketed by heavy snow, there are herds of bison and wolves, with two wolves preying on a bison. (...)
Candidate Summary
The video captures the harsh winter conditions in the Arctic and the wildlife that endure these extreme environments. (...)
(3)
Reference-based Metrics
BLEU-1
METEOR
ROUGE-L
BERTScore
QA-based Automatic Evaluation
Reference Summary
The video begins with a flock of birds on a partially frozen body of water. In a forest blanketed by heavy snow, there are herds of bison and wolves, with two wolves preying on a bison. (...)
Video
(1)
(2)
Candidate Summary
The video captures the harsh winter conditions in the Arctic and the wildlife that endure these extreme environments. (...)
Reference-free Metrics
① Coverage QA
② Factuality QA
③ Chronology QA
(3)
QEVA

Figure 1: Overview of existing video summarization evaluation approaches and our proposed method, QEVA. (Top) Traditional reference-based metrics rely solely on text-to-text comparisons between candidate and human-written reference summaries, often failing to capture nuanced semantic and multimodal content. (Bottom) QEVA leverages a fully reference-free multimodal question-answering pipeline (Coverage QA, Factuality QA, Chronology QA) to directly evaluate candidate summaries against source videos, enabling a more comprehensive and semantically grounded assessment. **Takeaway:** QEVA provides a more accurate and scalable alternative by eliminating the reliance on human-written reference summaries and directly assessing summaries against source video content.

1 Introduction

Text-based video summarization has become increasingly crucial due to the explosive growth of video content and significant advances in Video-Large Multimodal Models (Video-LMMs). Despite substantial progress in generating comprehensive textual summaries from videos, insufficient attention has been paid to reliably evaluating the quality of these summaries. Currently, evaluation primarily relies on reference-based metrics such as ROUGE or METEOR, which compare generated summaries to reference texts produced by human annotators. However, acquiring accurate and detailed reference summaries for videos requires considerable human effort and resources, making the process both costly and inefficient at scale. Consequently, the absence of efficient evaluation approaches has become a major obstacle to the advancement of video summarization research.

While reference-free metrics have gained con-

[*] Corresponding author.

[1] Our evaluation metric (QEVA) is available at https://github.com/jungnerd/QEVA

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24632–24642
November 4-9, 2025 ©2025 Association for Computational Linguistics

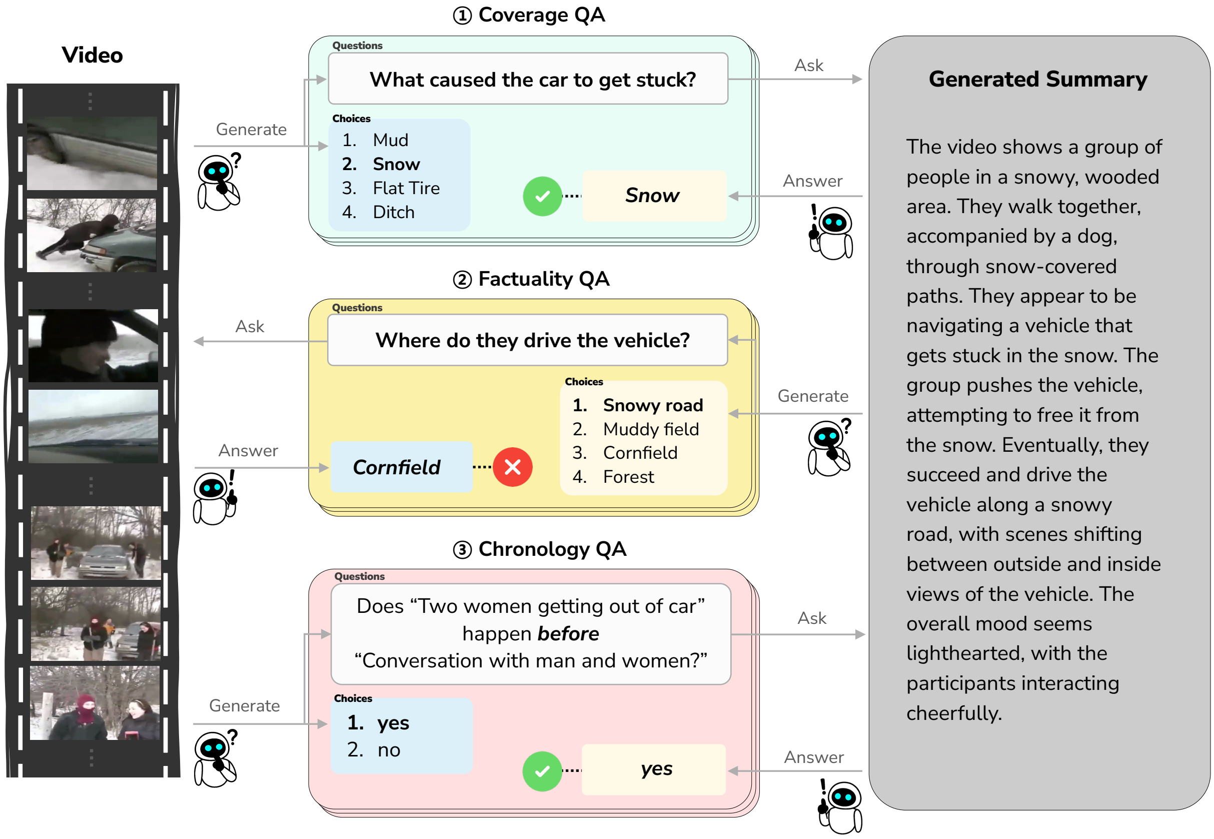

① Coverage QA
Video
Questions
What caused the car to get stuck?
Choices
1. Mud
2. Snow
3. Flat Tire
4. Ditch
Snow
Generate
Ask
Answer
Generated Summary
The video shows a group of people in a snowy, wooded area. They walk together, accompanied by a dog, through snow-covered paths. They appear to be navigating a vehicle that gets stuck in the snow. The group pushes the vehicle, attempting to free it from the snow. Eventually, they succeed and drive the vehicle along a snowy road, with scenes shifting between outside and inside views of the vehicle. The overall mood seems lighthearted, with the participants interacting cheerfully.
② Factuality QA
Questions
Where do they drive the vehicle?
Choices
1. Snowy road
2. Muddy field
3. Cornfield
4. Forest
Cornfield
Ask
Answer
Generate
③ Chronology QA
Questions
Does "Two women getting out of car" happen before "Conversation with man and women?"
Choices
1. yes
2. no
yes
Generate
Ask
Answer

Figure 2: Detailed illustration of QEVA's multimodal question-answering methodology. Given a video and candidate summary, QEVA evaluates the summary across three distinct dimensions: (1) Coverage (whether the summary comprehensively covers key video content), (2) Factuality (accuracy of the information presented), and (3) Chronology (i.e.,chronological fidelity; correctness of event ordering). QEVA generates tailored QA pairs through a structured pipeline involving question generation, question answering, and answer correctness checking. **Takeaway:** QEVA explicitly decomposes video summary quality into three complementary dimensions, enabling nuanced and interpretable evaluation.

siderable attention in text summarization tasks, these methods typically compare the generated summaries directly with the source text, leveraging linguistic similarities or embeddings. Unfortunately, extending this concept to videos is inherently challenging due to the fundamental modality gap - videos are spatio-temporal and multimodal, and thus cannot be directly compared to textual summaries in a straightforward manner. Moreover, recent attempts at multimodal evaluation, such as CLIPScore or LLM-based comparisons used in benchmarks like MLVU, still rely heavily on reference summaries or fail to reliably capture important aspects such as chronological fidelity or factual accuracy. Thus, there is a pressing need for a novel, effective, and fully reference-free evaluation paradigm tailored specifically for video summarization.

To address these challenges, we propose a **Q**uestion-answering based **E**valuation metric for **V**ideo summ**A**rization, QEVA. QEVA is built upon the intuitive principle that a high-quality video summary should be able to substitute the original video content effectively. Based on this principle, we identify three critical dimensions for evaluating summaries: Coverage (capturing all essential content), Chronology (chronological fidelity; preserving the order of events), and Factuality (ensuring factual correctness). QEVA leverages Video-LMMs to automatically generate relevant questions from the original video content across these dimensions and employs LLMs to answer these questions using only the generated summaries. By measuring the accuracy of these answers, QEVA quantitatively assesses summary quality without the need for any human reference summaries.

Furthermore, we present MLVU(VS)-Eval, a new annotated evaluation dataset derived from the MLVU benchmark, containing 800 summaries generated by Video-LMMs such as GPT-4o, QwenVL, InternVL, and Video-LLaVA. Each summary has been rigorously annotated by multiple human evaluators according to Coverage, Chronology, and Factuality, demonstrating strong inter-annotator agreement (Krippendorff's α = 0.68). Our comprehensive experiments show that QEVA consistently exhibits the highest correlation with human judgments compared to existing reference-based and multimodal evaluation methods (e.g., ROUGE, METEOR, BERTscore).

By introducing QEVA and the MLVU(VS)-Eval dataset, this work pioneers the first systematic exploration of reference-free evaluation for video-to-text summarization. Our approach not only addresses the pressing scalability and practicality issues but also sets a clear foundation for future advancements in multimodal summarization research, facilitating rapid and reliable assessment of Video-LMMs in real-world applications, including content platforms, news summarization, and automated content generation services.

2 Related Work

2.1 Automated Evaluation of Summarization

Automated evaluation has traditionally relied on lexical overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). While computationally efficient and widely used, these reference-based metrics primarily measure surface-level similarity between candidate and human-written reference texts, often failing to capture deeper semantic consistency, factual correctness, or coherence (Kryściński et al., 2019; Kasai et al., 2022).

To address these limitations, embedding-based metrics such as BERTScore (Zhang et al., 2020) and CLIPScore (Hessel et al., 2022) have been proposed to measure semantic similarity using learned embeddings. However, these methods still require reference summaries, limiting their practical applicability and scalability. More recently, Large Language Models (LLMs) have been leveraged as evaluators ("LLM-as-a-Judge"), demonstrating promising correlation with human judgments (Zheng et al., 2023), but they introduce challenges such as biases, reproducibility issues, and high API costs (Fu et al., 2023).

Our proposed QEVA metric addresses these challenges by employing a fully reference-free evaluation approach, directly comparing generated summaries against source videos through multimodal question-answering, thus providing a more semantically grounded and scalable alternative.

2.2 QA-based Evaluation Metrics

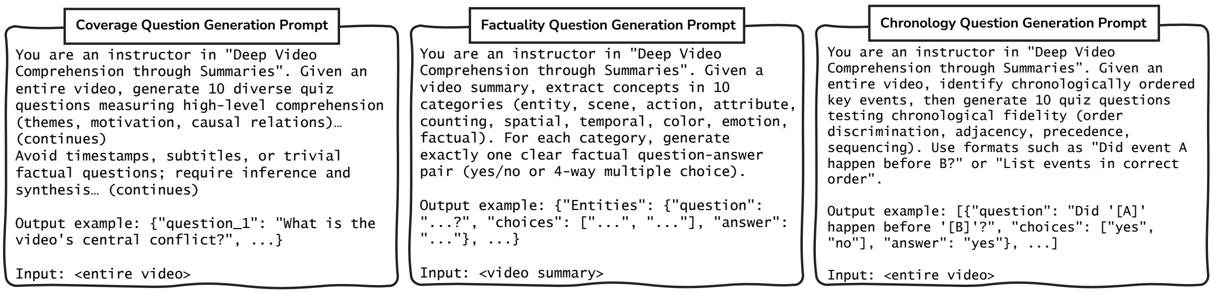
Question-answering (QA) based evaluation methods assess the semantic quality of generated texts by testing their ability to answer questions derived from relevant contexts (Wang et al., 2020). Unlike traditional lexical metrics, QA-based evaluation directly probes textual outputs for factuality, coverage, and relevance. Examples include FEQA (Durmus et al., 2020) and QAFactEval (Fabbri et al., 2022), which evaluate factual consistency by comparing answers derived from summaries and source documents.

Reference-free QA metrics such as QuestEval (Scialom et al., 2021) and Q^2 (Honovich et al., 2021) offer increased scalability by eliminating the dependency on human-written references. QuestEval integrates precision-oriented (summary-based) and recall-oriented (source-based) QA components, achieving strong correlation with human judgments. Recently, the TIFA metric (Hu et al., 2023) extended QA-based evaluation to text-to-image synthesis by generating questions from textual prompts and assessing image fidelity through visual QA.

Our work builds upon these advances by introducing QEVA, which uniquely applies multimodal QA to narrative video summarization, explicitly measuring Coverage, Factuality, and Chronology, thereby providing a comprehensive and reference-free evaluation framework tailored specifically for multimodal summarization.

3 The QEVA Method

We introduce **QEVA**, a novel evaluation framework designed to assess how comprehensively and faithfully a textual video summary captures the original video's content. QEVA is grounded in the principle that *"a good summary should serve as an effective substitute for the source video"*. To operationalize this intuition, QEVA employs multimodal question answering (QA) as its core mechanism: a high-quality summary should allow accurate answering of critical questions derived from the source video content. Unlike prior work relying solely on direct LLM-based judgments, QEVA integrates both Video-LMMs and LLMs within a collaborative pipeline, thus mitigating the inher-



Coverage Question Generation Prompt

```
You are an instructor in "Deep Video
Comprehension through Summaries". Given an
entire video, generate 10 diverse quiz
questions measuring high-level comprehension
(themes, motivation, causal relations)…
(continues)
Avoid timestamps, subtitles, or trivial
factual questions; require inference and
synthesis… (continues)

Output example: {"question_1": "What is the
video's central conflict?", ...}

Input: <entire video>
```

Factuality Question Generation Prompt

```
You are an instructor in "Deep Video
Comprehension through Summaries". Given a
video summary, extract concepts in 10
categories (entity, scene, action, attribute,
counting, spatial, temporal, color, emotion,
factual). For each category, generate
exactly one clear factual question-answer
pair (yes/no or 4-way multiple choice).

Output example: {"Entities": {"question":
"...?", "choices": ["...", "..."], "answer":
"..."}, ...}

Input: <video summary>
```

Chronology Question Generation Prompt

```
You are an instructor in "Deep Video
Comprehension through Summaries”. Given an
entire video, identify chronologically ordered
key events, then generate 10 quiz questions
testing chronological fidelity (order
discrimination, adjacency, precedence,
sequencing). Use formats such as "Did event A
happen before B?" or "List events in correct
order".

Output example: [{"question": "Did '[A]'
happen before '[B]'?", "choices": ["yes",
"no"], "answer": "yes"}, ...]

Input: <entire video>
```

Figure 3: Detailed prompts used by QEVA for multimodal question-answer generation across three distinct evaluation dimensions: (a) Coverage, (b) Factuality, and (c) Chronology. (a) Coverage QA prompts instruct a Video-LMM to generate high-level, inference-driven questions that assess whether the summary comprehensively captures essential content, themes, and events from the source video. (b) Factuality QA prompts guide an LLM to first extract critical factual elements explicitly mentioned in the summary (such as entities, actions, attributes), and subsequently generate targeted questions designed to verify factual accuracy. (c) Chronology QA prompts outline a structured workflow for event segmentation from the source video, sampling adjacent and distant pairs of events, and generating questions to evaluate whether the summary preserves the correct chronological order of these events. **Takeaway:** QEVA employs carefully structured prompts tailored to each evaluation dimension, enabling a systematic and comprehensive assessment of video summary quality along the axes of Coverage, Factuality, and Chronology.

ent limitations of Video-LMMs as direct evaluators (see Section 4.4 for detailed analyses). Given a source video and candidate summary, QEVA outputs a multidimensional quality score reflecting the summary's *coverage*, *factuality*, and *chronology (chronological fidelity)*. Figure 6 provides an overview of our QEVA workflow.

3.1 Evaluation Criteria

Inspired by prior work in text summarization evaluation (e.g., SummEval) and recent multimodal captioning evaluation metrics (e.g., ACCR in G-VEval), we define three complementary criteria to comprehensively evaluate video summaries: **Coverage**, **Factuality**, and **Chronology**.

Coverage. Coverage measures whether the summary includes essential content, key events, and the main messages of the source video. A high-coverage summary should comprehensively represent all core information and omit no major events or salient points.

Factuality. Factuality assesses whether the summary accurately reflects details present in the source video without introducing hallucinated or unsupported information. Good factuality implies all claims in the summary can be directly verified from the video.

Chronology. Chronology (chronological fidelity) evaluates whether a summary preserves the source video's order of events and correctly interprets nonlinear structures (e.g., flashbacks/flash-forwards). This criterion concerns ordering only; fine-grained timing such as duration or pace is not evaluated.

These criteria are mutually complementary. A summary may comprehensively cover all major events (*Coverage*) but misrepresent critical details (*Factuality*), or accurately summarize events but distort their order (*Chronology*). By explicitly defining these distinct evaluation dimensions, QEVA ensures consistent and interpretable evaluation, minimizing subjective interpretation and inter-annotator variance.

3.2 Question-Answer Generation

For each evaluation criterion, QEVA generates tailored question-answer (QA) pairs to probe summary quality. Figure 3 illustrates example prompts for QA generation, with complete prompts provided in Appendix A.

Coverage QA Generation. We prompt a Video-LMM with the full source video and a specialized instruction to generate N diverse, high-level questions. These prompts encourage synthesis and inference-based questions that target main events, causal relations, and overarching themes, explicitly avoiding superficial factual or timestamp-specific queries.

Factuality QA Generation. We prompt an LLM to extract salient elements from the candidate summary and categorize them (entities, actions, attributes, counting, etc.). Then, we generate targeted factual queries that require answers strictly supported by the summary itself, enabling verification of summary correctness against the source video.

Chronology QA Generation. We first employ a Video-LMM to extract a chronological sequence of key events from the video. Subsequently, we sample pairs of events (adjacent and non-adjacent) and generate three types of chronological questions: (1) order verification (yes/no), (2) temporal precedence (multiple-choice), and (3) event sequence sorting tasks. These questions explicitly evaluate if the summary accurately preserves event ordering.

3.3 Question and Answer Filtering

To ensure the quality and discriminative power of generated QA pairs, we introduce a two-stage automatic filtering step. This process is designed to eliminate questions that either do not require the source context (i.e., are trivial) or are inherently flawed (i.e., ambiguous or low-quality). Specifically, we employ an alternative Video-LMM or LLM (distinct from the model used for generating the QA pairs) to answer each question under two conditions:

- **Trivial Filtering:** If the alternative model correctly answers a question without any context (using only the question and answer choices), we consider it trivial and remove it. This step is vital to ensure that QEVA genuinely tests for faithfulness to the provided context rather than the model's own parametric knowledge.
- **Low-quality Filtering:** If the model fails to answer correctly even when provided with the appropriate context (the source video or summary), the question is considered ambiguous or unanswerable and is thus discarded. This removes noise from the evaluation.

This two-stage filtering ensures retained QA pairs are neither trivial nor excessively ambiguous.

3.4 QEVA Score Computation

For each criterion, QEVA computes an evaluation score as the proportion of correctly answered questions from the filtered set. Let Q_{Cov}, Q_{Fact}, and Q_{Chrono} denote the filtered QA sets for Coverage, Factuality, and Chronology, respectively.

- For **Coverage** and **Chronology**, the summary S is used as the context. The scores are the proportions of correctly answered questions:

$$\text{Score}_{Cov}(S, Q_{Cov}) = \frac{|Q_{Cov,\text{correct}}|}{|Q_{Cov}|} \tag{1}$$

$$\text{Score}_{Chrono}(S, Q_{Chrono}) = \frac{|Q_{Chrono,\text{correct}}|}{|Q_{Chrono}|} \tag{2}$$

- For **Factuality**, the source video V is used as the context. The score is the proportion of questions confirmed as accurate:

$$\text{Score}_{Fact}(V, Q_{Fact}) = \frac{|Q_{Fact,\text{correct}}|}{|Q_{Fact}|} \tag{3}$$

The final QEVA score is the arithmetic mean of these three component scores. This provides a final score on a normalized 0-to-1 scale. Formally, for a summary S and video V, the score is defined as:

$$\begin{aligned}\text{QEVA}(S, V) = \big(&\text{Score}_{Cov}(S, Q_{Cov}) \\ &+ \text{Score}_{Fact}(V, Q_{Fact}) \\ &+ \text{Score}_{Chrono}(S, Q_{Chrono})\big)/3\end{aligned} \tag{4}$$

3.5 Implementation Details

For our default experimental setup, we use Gemini-1.5 Pro as the primary Video-LMM and GPT-4o as the primary LLM for QA generation and answering. Alternative settings with open-source models (e.g., Qwen2.5-VL, InternVL3, LLaMA-3.1, Gemma-3) are also explored in Section 4.4. All prompts, hyperparameters, and scripts necessary for full reproducibility are included in the supplementary material and will be publicly released upon acceptance.

4 Experiments

We conduct comprehensive experiments to validate the effectiveness of our proposed QEVA metric. We first introduce **MLVU(VS)-Eval**, a novel benchmark dataset for evaluating video-to-text summarization metrics (§4.1). We then compare QEVA against existing metrics in terms of correlation with human judgments (§4.2). Additionally, we analyze model-wise performance of QEVA (§4.3), and demonstrate the robustness of QEVA via ablation studies (§4.4).

Direction	Metric	Reference Summary	Video Used	MLVU(VS)-Eval		
				Kendall's τ_b	**Kendall's τ_c**	**Spearman's ρ**
Rule-based	BLEU-1	✓		0.0217 (0.7803)	0.0219 (0.7803)	0.0431 (0.7045)
	BLEU-2	✓		0.0673 (0.3865)	0.0680 (0.3865)	0.1045 (0.3561)
	BLEU-3	✓		0.1994 (0.0152)	0.1836 (0.0152)	0.2836 (0.0108)
	BLEU-4	✓		0.2113 (0.0155)	0.1578 (0.0155)	0.2798 (0.0119)
	ROUGE-L	✓		0.1094 (0.1593)	0.1104 (0.1593)	0.1571 (0.1639)
Similarity-based	METEOR	✓		0.2663 (0.0006)	0.2689 (0.0006)	0.3822 (0.0005)
	CIDEr	✓		0.2401 (0.0015)	0.2435 (0.0014)	0.3612 (0.0012)
	SPICE	✓		0.2287 (0.0021)	0.2312 (0.0020)	0.3489 (0.0019)
	BERTscore	✓		0.1415 (0.0688)	0.1428 (0.0688)	0.1987 (0.0773)
	Video-Summary Similarity		✓	0.0278 (0.0517)	0.0266 (0.0522)	0.0401 (0.0549)
LLM-based	MLVU	✓		0.5284 (0.0000)	0.5309 (0.0000)	0.6738 (0.0000)
	Video-LMM Judge		✓	0.5376 (0.0000)	0.5441 (0.0000)	0.6810 (0.0000)
QA-based	QEVA(Ours)		✓	**0.6465 (0.0000)**	**0.6407 (0.0000)**	**0.7326 (0.0000)**

Table 1: Comparative evaluation results of QEVA and existing summarization evaluation metrics on the MLVU(VS)-Eval benchmark. We report correlations with human judgments using Kendall's τ_b, τ_c, and Spearman's ρ. The metrics are categorized into several groups: rule-based n-gram metrics (BLEU variants, ROUGE-L), similarity-based metrics (METEOR, CIDEr, SPICE, BERTScore, Video-Summary Similarity), LLM-based metrics (MLVU, Video-LMM Judge), and our proposed QA-based metric (QEVA). Reference usage and video modality usage for each metric are also indicated. **Takeaway:** QEVA consistently achieves significantly higher correlation with human judgments compared to existing metrics, demonstrating its effectiveness as a reference-free and multimodal evaluation metric.

Metric	Kendalls' τ_b				Kendalls' τ_c			
	Qwen2.5-VL-7B	**InternVL3-8B**	**Video-LLaVA-7B**	**GPT-4o**	**Qwen2.5-VL-7B**	**InternVL3-8B**	**Video-LLaVA-7B**	**GPT-4o**
BLEU-1	0.1662	0.0870	0.0758	0.0053	0.1669	0.0867	0.0758	0.0054
BLEU-2	0.2735	0.1247	0.1457	0.1333	0.2746	0.1246	0.1458	0.1339
BLEU-3	0.2914	0.1380	0.1370	0.3293	0.2800	0.1300	0.1108	0.3088
BLEU-4	0.1654	0.1130	**0.3565**	0.1717	0.1444	0.0914	0.2025	0.1463
ROUGE-L	0.4022	0.2332	0.1573	0.3423	0.4038	0.2329	0.1575	0.3429
METEOR	0.2628	0.1573	0.3555	0.4107	0.2638	0.1571	**0.3558**	0.4125
BERTscore	0.5202	0.0163	0.0701	0.3147	0.5223	0.0163	0.0700	0.3161
Video-Summary Similarity	0.0012	-0.0301	0.2277	0.2417	-0.0218	-0.0119	0.2117	0.1472
MLVU	0.3950	0.3260	-0.1254	0.2361	0.3900	0.3208	-0.1067	0.2392
Video-LMM Judge	0.4138	0.3358	0.2056	0.2178	0.2005	0.3123	0.1233	0.2198
QEVA(Ours)	**0.4509**	**0.4268**	0.1450	**0.4262**	**0.4500**	**0.4114**	0.1000	**0.4222**

Table 2: Model-wise correlation analysis of summarization evaluation metrics on the MLVU(VS)-Eval dataset. Correlations (measured by Kendall's τ_b and τ_c) are reported separately for summaries generated by four representative Video-LMM models: Qwen2.5-VL-7B, InternVL3-8B, Video-LLaVA-7B, and GPT-4o. **Takeaway:** QEVA exhibits stable and positive correlations with human judgments across diverse Video-LMM summarization models, highlighting its robustness and generalizability.

4.1 MLVU(VS)-Eval: A Novel Benchmark for Evaluating Video Summarization Metrics

Existing evaluation datasets for video-to-text summarization lack human annotations, limiting accurate metric evaluation. To address this, we propose **MLVU(VS)-Eval**, a novel human-annotated dataset built upon the MLVU benchmark (Zhou et al., 2024).

We select 200 video clips (average length ∼15 minutes) from the MLVU Video Summarization task. Each video has a human-written reference summary. We generate candidate summaries using four widely-used Video-LMM models: GPT-4o, InternVL3-8B, Qwen2.5-VL-7B, and Video-LLaVA-7B, resulting in 800 candidate summaries. For our evaluation, we recruited 20 annotators (comprising graduate and undergraduate students) to assess each summary based on three criteria: *Coverage*, *Factuality*, and *Chronology*, using a 5-point Likert scale. Each summary received evaluations from two independent annotators. The inter-annotator agreement, measured using Krippendorff's α, is 0.68, indicating substantial reliability of the annotations.

4.2 Correlation with Human Judgments

We compare QEVA with various existing evaluation metrics in terms of correlation with human judgments.

Baseline Metrics. We compare QEVA with several representative categories of evaluation metrics: reference-based n-gram overlap metrics, embedding-based similarity metrics, and LLM-based evaluation approaches. For the n-gram and embedding-based metrics, we report results using widely adopted methods in each category. Additionally, we introduce two novel baselines: (1) employing a Video-LMM as a direct judge by prompting it with our human annotation guidelines, and (2) computing multimodal embedding similarity between the original video and generated summary using a state-of-the-art multimodal encoder[2].

Results and Analysis. Table 1 presents correlations measured by Kendall's τ_b, τ_c, and Spearman's ρ. QEVA consistently achieves significantly higher correlations than all existing metrics, highlighting its superior alignment with human evaluations, especially notable given QEVA is a reference-free metric.

4.3 Model-wise Correlation Analysis

To further examine QEVA's consistency, we analyze correlation results separately for each Video-LMM summarization model. Table 2 shows that QEVA consistently yields positive and stable correlations across different summarization models. In contrast, some baseline metrics even show negative or inconsistent correlations for certain models.

We observe relatively lower correlations for Video-LLaVA-generated summaries. Upon qualitative analysis, we find these summaries are generally shorter and of lower quality, making it challenging for QEVA to generate meaningful questions and answers.

4.4 Ablation Studies

We perform ablation studies to analyze QEVA's robustness and generalizability.

Evaluation Criteria-wise Ablation. We separately calculate correlations between human judgments and QEVA scores for each evaluation criterion (Coverage, Factuality, Chronology). As shown in Table 3, QEVA demonstrates strong correlations across all individual criteria, validating its capability to accurately capture different aspects of summarization quality.

Robustness to Different Video-LMM and LLM Models. We further examine QEVA's internal robustness by replacing the original Video-LMM (Gemini-1.5 Pro) and LLM (GPT-4o) components with alternative models. Specifically, we test open-source Video-LMM variants (QwenVL, InternVL) and LLM variants (LLaMA, Gemma). Table 4 indicates that QEVA maintains high correlations even with open-source alternatives. This demonstrates QEVA's practical applicability and cost-effectiveness by alleviating reliance on costly API-based models.

Criteria	τ_b	τ_c	ρ
Coverage	0.6005	0.5872	0.7349
Factuality	0.5731	0.5586	0.7123
Chronology	0.5610	0.5479	0.6984

Table 3: Ablation study of QEVA across the three individual evaluation criteria (Coverage, Factuality, Chronology). Correlations with human judgments (measured by Kendall's τ_b, τ_c, and Spearman's ρ) are reported separately for each dimension. **Takeaway:** QEVA demonstrates strong and consistent correlations across all individual evaluation criteria, validating its comprehensive and multidimensional evaluation approach.

Video-LMM + LLM	τ_b	τ_c	ρ
Qwen2.5-VL + LLaMA-3.1	0.6211	0.6152	0.7012
Qwen2.5-VL + Gemma-3	0.6075	0.6021	0.7120
InternVL3 + LLaMA-3.1	0.5988	0.5933	0.6761
InternVL3 + Gemma-3	0.5827	0.5780	0.6894
Gemini-1.5-Pro + GPT-4o	**0.6465**	**0.6407**	**0.7326**

Table 4: Ablation study examining the robustness of QEVA to different combinations of Video-LMM and LLM models compared to the default setting (Gemini-1.5 Pro + GPT-4o). We report correlations with human annotations (Kendall's τ_b, τ_c, and Spearman's ρ) when replacing original models with alternative open-source models (Qwen2.5-VL, InternVL3, LLaMA-3.1, Gemma-3). **Takeaway:** QEVA maintains high correlations with human judgments even when using alternative open-source models, indicating practical applicability and cost-effectiveness without relying solely on costly API-based models.

5 Conclusion

We propose QEVA, a novel reference-free metric for evaluating narrative video summarization leveraging multimodal question answering. QEVA demonstrates significantly higher correlation with human judgments compared to existing metrics, while eliminating reliance on costly reference summaries. Our approach facilitates scalable, accurate, and practical evaluation of video summarization systems, accelerating the development and deploy-

[2] We use LanguageBind (Zhu et al., 2023) for multimodal embedding in this paper.

ment of Video-LMMs in real-world multimodal applications.

Limitations

Despite the effectiveness and practicality demonstrated by QEVA, our proposed metric inherits several limitations inherent to its underlying models and design.

Hallucination. As QEVA employs a Large Multimodal Model (LMM) for question generation and answering, it may occasionally produce hallucinated content in the generated questions or answers not actually present in the video. Although our filtering process significantly mitigates this issue, the possibility of subtle hallucinations remains, potentially affecting evaluation reliability in edge cases.

API Cost and Processing Speed. QEVA relies heavily on inference from Large Language Models (LLMs) and Video-LMMs. Such models typically require significant computational resources and incur relatively high API costs, particularly when evaluating large-scale datasets or numerous summaries. This dependence may limit QEVA's practical applicability in resource-constrained environments or real-time scenarios.

Necessity of Post-processing. QEVA occasionally produces outputs that deviate from the predefined format or scoring criteria. Although infrequent, these cases necessitate additional post-processing to ensure compliance with the intended evaluation guidelines, slightly complicating the evaluation pipeline.

Preference for LLM-based Outputs. Recent evaluations using LLMs have identified a subtle preference bias towards outputs generated by LLMs themselves. QEVA may similarly exhibit a slight bias favoring summaries produced by certain Video-LMMs, potentially influencing the fairness and objectivity of the evaluation. This phenomenon warrants further investigation to quantify and mitigate such biases in future research.

Ethics Statement

Ethics Statement Our research involved a user study with human participants to collect judgments for evaluating our proposed metric. The study was conducted in accordance with established ethical guidelines. We recruited 20 annotators who participated on a voluntary basis. While our institution's review board determined that formal IRB approval was not required for this type of user study (involving subjective evaluation of anonymized system outputs), we took several measures to ensure the protection of participants.

Prior to the study, all participants were provided with a clear description of the research objectives and the annotation task, and we obtained informed consent from each individual. To acknowledge their valuable contribution, participants were fairly compensated for their time and effort. All data collected during the study was fully anonymized to protect the privacy and confidentiality of the participants, and no personally identifiable information (PII) was collected or stored.

Acknowledgement

This work was supported by IITP grant funded by the Korea government(MSIT) (No. RS-2020-II200004, Development of Previsional Intelligence based on Long-term Visual Memory Network) and supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grand funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program(Chung-Ang University)].

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1478–1490.

Jinlan Fu, See-Kiong Lu, and W Lam. 2023. GPTScore: Evaluate as you desire. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *Preprint*, arXiv:2104.08202.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. *Preprint*, arXiv:2111.08940.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5477–5488.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Thomas Scialom, Daniel Deutsch, Anthony R Fabbri, Vasili Zhong, Sascha Rothe, and Alexander M Rush. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10036–10050.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Preprint*, arXiv:2004.04228.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

A Full Prompts

In this section, we provide the complete prompts utilized for generating QA pairs across the three distinct QEVA evaluation dimensions—**Coverage**, **Factuality**, and **Chronology**. We present these prompts exactly as they were given to the Large Multimodal Models (Video-LMMs) and Large Language Models (LLMs) in our experiments. Researchers can use these prompts to precisely reproduce the QA generation process described in Section of the main paper.

```
You are an expert instructor in the course “Deep Video Understanding through Summarization”.

## Objective
Given an entire video as input, your task is to generate exactly 10 diverse quiz questions. These questions will assess
students' deep comprehension of the video based solely on reading a textual summary.

## Abilities to test
- Identifying key ideas, causal connections, motivations, and overarching themes in the video.
- Summarizing, comparing, and synthesizing long segments rather than recalling isolated details.
- Reasoning and inference beyond surface-level details.

## Guidelines for question generation
- **DO NOT** reference timestamps, subtitles, screen texts, or narration explicitly.
- **DO NOT** generate superficial factual questions ("When…?", "How many…?").
- **DO** formulate high-level inference questions asking to explain, infer, or synthesize key ideas (e.g., "What is the central
conflict of the story?", "Why did Character A decide to do action B?").
- Ensure answers can be clearly derived from a well-written textual summary (without requiring detailed visual-only cues).
- Questions should cover diverse aspects of the video (themes, conflicts, character motivations, cause-effect relations)
without redundancy.

## Output format
Return a JSON object with exactly 10 keys as follows:

{
  "question_1": "Question text here",
  "question_2": "Question text here",
  ...
  "question_10": "Question text here"
}
```

Figure 4: Methodology figure of QEVA.

```
You are an expert instructor in the course “Deep Video Understanding through Summarization”.

## Objective
Given a textual video summary, your task is to generate exactly 10 clear and precise quiz questions. These questions will
measure whether a generated video summary is factually accurate and consistent with the original video content.

## Abilities to test
- Accurate recognition and recall of explicit factual information from the summary.
- Validation of objective details such as entities, scenes, actions, and attributes.

## Guidelines for question generation
- **DO NOT** ask overly obvious or trivial questions (e.g., "Is there a person?").
- Generate exactly one quiz question per each of the following 10 categories:
  Entities, Scene, Actions, Attributes, Counting, Spatial, Temporal, Colors, Emotion, Factual.
- For each category, first extract one meaningful concept explicitly mentioned in the summary, then generate one question-
answer pair.
- Each question must be answerable directly from the given summary (no external knowledge or inference required).
- Question formats allowed:
  (1) yes/no (Choices: yes, no), or
  (2) multiple-choice with four options (exactly one correct option).

## Output format
Return a JSON object with exactly 10 questions as follows:

{
  "Entities": {
    "question": "Question text here",
    "choices": ["option1", "option2", "option3", "option4"], // or ["yes", "no"]
    "answer": "correct option here"
  },
  "Scene": {
    "question": "...",
    "choices": [...],
    "answer": "..."
  },
  ...
  "Factual": {
    "question": "...",
    "choices": [...],
    "answer": "..."
  }
}
```

Figure 5: Methodology figure of QEVA.

```
# Objective:
Analyze the provided video to generate Question-Answering (QA) pairs designed to evaluate the temporal order of events within
it. These QA pairs will be used to assess how well a video summary preserves the chronological flow of the original video.

# Input:
[Provide the video input here, e.g., file path, link, or direct upload if supported]

# Execution Steps:

## Step 1: Event Segmentation and Listing
1.  Analyze the entire video to identify key events (significant actions, scene changes, state transitions, etc.).
2.  Sort the identified events chronologically based on their occurrence in the video.
3.  Describe each event concisely and clearly (e.g., "A man opens the door," "A car passes through the intersection," "A cat
jumps off the sofa").
4.  Output the result as a numbered list. (Assume the total number of events in this list is L).

    **Example Output (Event List):**
    1.  [Event 1 Description]
    2.  [Event 2 Description]
    3.  ...
    L.  [Event L Description]

## Step 2: Event Pair Sampling
Using the chronologically ordered event list (L events total) generated in Step 1, sample pairs of events (Event i, Event j)
according to the following criteria:

1.  **Adjacency Pairs:**
    * Definition: Pairs of events that occur immediately one after the other in the list (Event i, Event i+1).
    * Purpose: To check the order of logically connected, consecutive events.
    * Sampling: Select **all** Adjacency Pairs (Event i, Event i+1) for i from 1 to L-1.

2.  **Distant Pairs:**
    * Definition: Pairs of events that are temporally separated in the list (|i - j| ≥ k, where k=2).
    * Purpose: To detect global temporal order distortions (e.g., flashbacks, scene inserts).
    * Sampling Quantity: Sample approximately [Number, e.g., 5-10 or desired number] Distant Pairs.
    * Sampling Strategy: Use a combination of the following strategies to ensure diversity:
        * **Edge Emphasis:** Try to include pairs involving events from the beginning-middle, middle-end, or beginning-end of
the video.
        * **Diversity:** Prefer pairs involving different subjects or actions to avoid redundancy. (Pairs with similar subject-
verb structures to already selected pairs should be given lower priority).
        * Sample pairs satisfying |i - j| ≥ 2, considering the strategies above.

3.  **Output:** Clearly list the sampled pairs, distinguishing between 'Adjacency' and 'Distant' types.

    **Example Output (Sampled Pair List):**
    * Adjacency Pairs: (Event 1, Event 2), (Event 2, Event 3), ... (Event L-1, Event L)
    * Distant Pairs: (Event 1, Event 4), (Event 3, Event 7), (Event 2, Event L), ... (Total of [Sampled Quantity] pairs)

## Step 3: QA Pair Generation
For each sampled event pair (A, B) from Step 2, or for selected triplets of events (X, Y, Z) as needed, generate question-
answer pairs based on the provided QA templates. Replace <A>, <B>, <X>, <Y>, <Z> with the actual event descriptions from Step 1.

1.  **Apply QA Templates:**
    * **Order Discrimination (Boolean):** (Applicable to all pairs)
        * Question: "Did <A> happen before <B>?"
        * Answer: (Based on original video) Yes / No
    * **Adjacency Check (Boolean):** (**Only** applicable to Adjacency Pairs)
        * Question: "Does <B> happen immediately after <A>?"
        * Answer: (Based on original video) Yes / No
    * **Precedence Selection (Multiple Choice):** (Applicable to all pairs)
        * Question: "Which of the following two events happened first? 1. <A> 2. <B>"
        * Answer: (The event that occurred earlier in the original video) 1 or 2
    * **Sequence Ordering (Ordering):** (Generate about [Number, e.g., 2-3] questions using Distant Pairs or arbitrary triplets
(X, Y, Z))
        * Question: "List the following events in the order they occurred: A. <X>, B. <Y>, C. <Z>"
        * Answer: (The correctly ordered sequence based on the original video) e.g., ACB

2.  **Output:** Present the generated QA pairs clearly. Include both the question and the corresponding ground truth answer
(based on the original video). **You MUST generate 10 QA pairs.**

    **Example Output (QA Pairs):**
    * Question: "Did '[Event 1 Description]' happen before '[Event 2 Description]'?" / Answer: Yes
    * Question: "Does '[Event 1 Description]' happen immediately after '[Event 2 Description]'?" / Answer: No
    * Question: "Which of the following two events happened first? 1. [Event 5 Description] 2. [Event 3 Description]" / Answer:
2
    * Question: "List the following events in the order they occurred: A. [Event 1 Description], B. [Event 7 Description], C.
[Event 4 Description]" / Answer: ACB
    * ... (3 Boolean QA pairs, 4 Multiple choice QA pairs, 3 Ordering QA pairs)

# Final Output Format:
Present the results for Step 3 in json format.

**Example Output**
[{"question": "Did '[Event 1 Description]' happen before '[Event 2 Description]'?",
"choices": ['yes', 'no'],
"answer": 'yes'},
{"question": "Does '[Event 1 Description]' happen immediately after '[Event 2 Description]'?",
"choices": ['yes', 'no'],
"answer": 'no'},
{"question": "Which of the following two events happened first? 1. [Event 5 Description] 2. [Event 3 Description]",
"choices": ['1', '2'],
"answer": '2'},
{"question": "List the following events in the order they occurred: A. [Event 1 Description], B. [Event 7 Description], C.
[Event 4 Description]",
"choices": ['ABC', 'ACB', 'BAC', 'BCA', 'CAB', 'CBA'],
"answer": 'ACB'},
...
]
```

Figure 6: Methodology figure of QEVA.