OVG-HQ: Online Video Grounding with Hybrid-modal Queries
Paper ID 7351

Supplementary Material

In the supplementary material, we provide more details and more experimental results of our work. We organize the supplementary into the following sections.

- In Section A, we present a detailed description of our proposed OVG-HQ-Unify framework, which consists of the feature extraction, memory-guided multi-modal fusion module, memory-guided moment prediction module.
- In Section B, we provide results on ICQ-Highlight dataset and results of the ablation experiments.
- In Section C, we provide a comprehensive description of the process used to construct our hybrid-modal queries dataset, QVHighlights-Unify. Additionally, we showcase some examples of the samples we generate.
- In Section D, we analyze the diversity, realism, and construction reliability of our dataset, and provide a statistical comparison with existing benchmarks.
- In Section E, we present additional experimental results, including the utility of multimodal queries for existing offline grounding methods.

A. More Details of Model and Implementation

We provide the implementation details in this section. We will release the collected dataset, source code and trained models upon acceptance.

A.1. Details of Feature Extraction

This section introduces the feature extraction process, as shown in Figure A. The feature extraction module accepts arbitrary combinations of multi-modal queries and current video window as input, and outputs query features and video features via the query feature extractor and the video feature extractor. When a query includes multiple modalities, learnable modality tokens are inserted at the beginning of each query feature sequence to enable the model to distinguish different modalities. Below, we separately introduce the video feature extractor and the query feature extractor we employ.

A.1.1. Video Feature Extractor

We process streaming video through a sliding window of size L seconds advancing at $M = 2$-second intervals. At each time step t, the CLIP image encoder [8] extracts features from current frame, with overlapping window features computed once and cached. For initial windows containing fewer than K frames, we left-pad zero vectors to maintain the feature matrix dimensionality $\mathbf{F}_v \in \mathbb{R}^{K \times D_v}$.

A.1.2. Query Feature Extractor

1) Text Query: The CLIP text encoder [8] generates textual features $\mathbf{F}_t$.
2) Segment Query: Employing the video feature extractor with 2-second sampling intervals to obtain $\mathbf{F}_s$.
3) Image Query: The CLIP image encoder [8] extracts features $\mathbf{F}_i$, which are duplicated temporally to match the segment query length.

A.2. Memory-guided Multi-modal Fusion Module

We employ a two-layer Transformer Decoder [7] to fuse the video features and query features, resulting in query-aware video representations $\mathbf{F}_{qv} \in \mathbb{R}^{K_{\text{VID}} \times D}$. Subsequently, we input $\mathbf{F}_{qv}$ into two layers of residual block composed of parametric memory blocks, integrating historical information to obtain memory-guided features $\hat{\mathbf{F}}_{qv} \in \mathbb{R}^{K_{\text{VID}} \times D}$. We set K_{VID}=16. The detailed process is illustrated in Figure A.

A.3. Memory-guided Moment Prediction Module

We employ a two-layer Transformer Decoder [7] to decode the memory-guided features. **First**, the Anchor Embeddings and the Anchors' center and width coordinates are used as queries for the Decoder. Through this decoding process, we obtain the anchor features and the refined anchor center and width coordinates. **Second**, the anchor features produced by the Transformer Decoder are input into a classifier and a regressor to obtain classification and regression results. **Third**, these results are passed to subsequent parametric modules, where they are combined with historical information to generate optimized predictions. We employ 4 anchor queries with lengths of 1, 2, 4, 8. The detailed procedure is illustrated in Figure B.

A.4. Implementation Details

We use PyTorch 2.0.0 and 1 4090 GPUs for our experiments. The model weights are initialized using Xavier initialization [2]. We optimize the model parameters using the AdamW [6] optimizer, with an initial learning rate of 1e-4 and a weight decay of 1e-4. The model is trained for 30 epochs; The batch size is set to 256. In the focal loss [10], we set $\alpha = 0.9$ and $\gamma = 2$, and apply a dropout rate of 0.5. In the distillation process, the temperature for the KL divergence is set to 2. During training, only anchors with a temporal Intersection over Union (tIoU) greater than 0.5 with the ground truth are considered training samples.

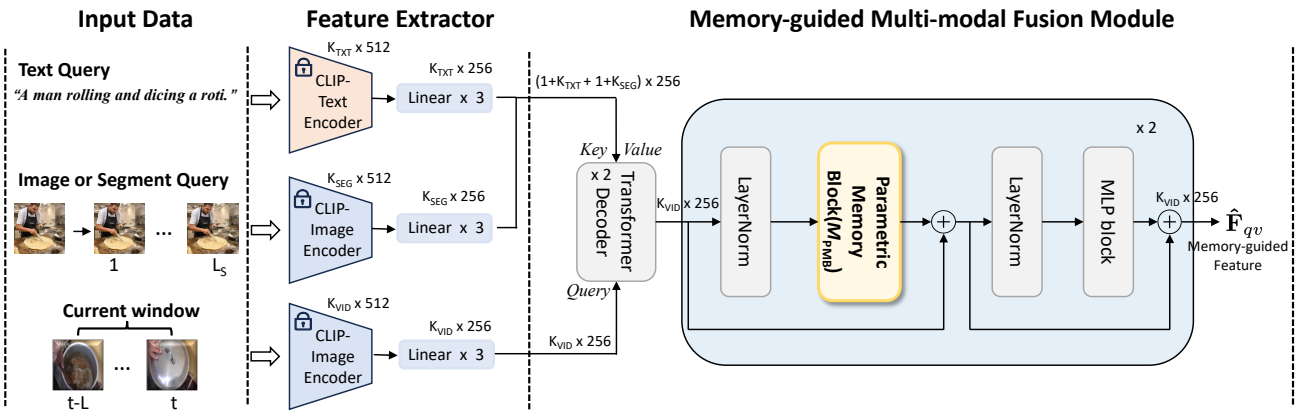


Figure A. Detailed illustration of the Input Data, Feature Extractor, and Memory-guided Multi-modal Fusion Module. "Linear × N" indicates N consecutive linear layers coupled with Layer Normalization.

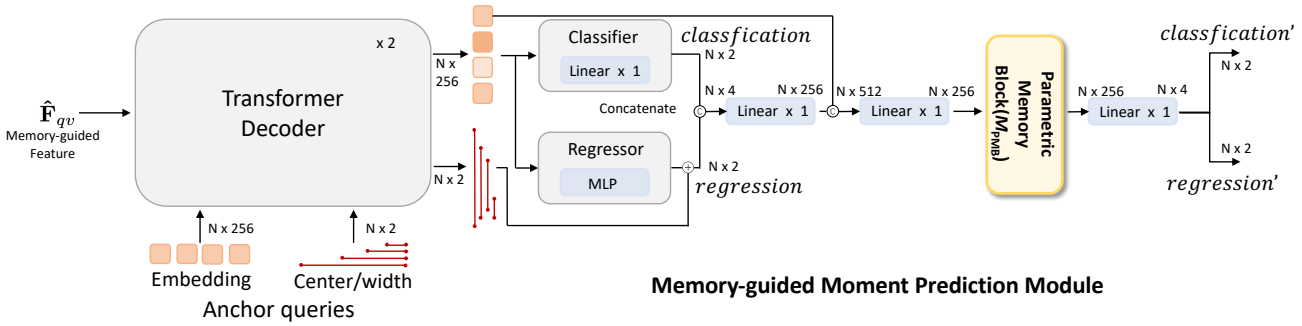


Figure B. Detailed illustration of Memory-guided Moment Prediction Module.

B. More Ablation Study Results

B.1. More results on ICQ-Highlight dataset and Comparisons with hybrid-modal retrieval methods.

To further validate effectiveness of our approach, we conducted comprehensive experiments on the ICQ-Highlight [12] dataset containing abundant complementary image-text query pairs. Following the summarization paradigm proposed in [12], we employ a MLLM (LLaVA-mistral-1.6) [4, 5] to convert paired image-text inputs into unified textual queries. As demonstrated in Table A, when feeding these synthesized textual queries into models pre-trained on QVHighlights, our method significantly outperforms the TwinNet baseline across all four image style categories in evaluation metrics, thereby conclusively validating the efficacy of our model design. Notablly, our proposed QVH-HQ-Unify framework achieves state-of-the-art performance across all four image style when directly processing raw multimodal queries from ICQ-Highlight. This demonstrates the remarkable advantages of our unified framework in cross-modal semantic fusion, which effectively captures complementary semantic information from heterogeneous image-text queries.

Table A. Comparison between our model and TwinNet across four distinct styles.

Method	scribble	cartoon	cinematic	realistic
TwinNet-MLLM	18.97	18.62	17.45	18.37
Ours-MLLM	19.95	20.34	19.53	19.08
Ours-QVH-HQ-Unify	21.98	22.35	21.25	22.07

C. Dataset Construction Details

Our dataset is constructed based on QVHighlights [3], which originally contains only text queries and raw videos. To expand the modalities of queries, we enrich the dataset by incorporating additional types of queries.

C.1. Detailed construction pipeline

In this section, we present the construction processes of the four types of queries: Retrieved image (Image-R), generated image (Image-G), generated video segment (Segment-

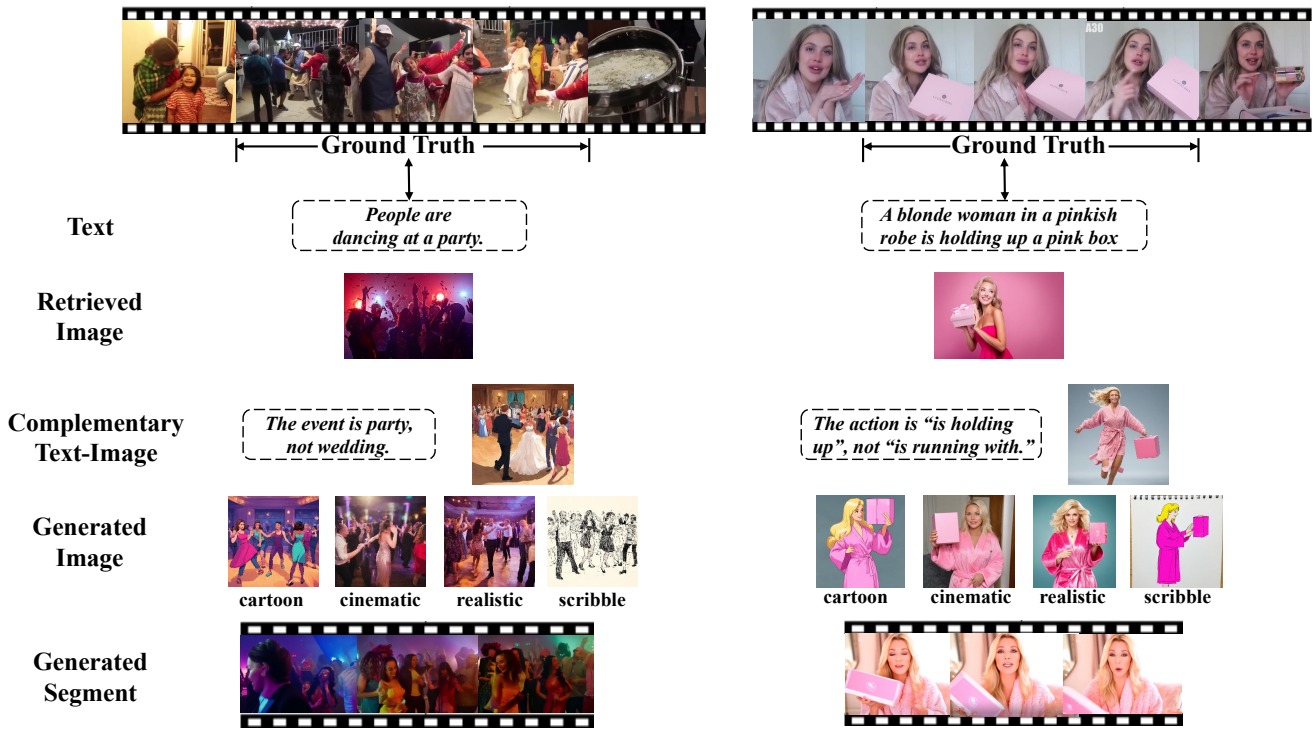
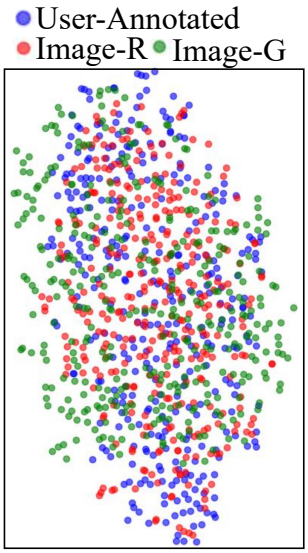


Figure C. Examples of our constructed QVHighlights-Unify dataset.

G), and complementary text-image (Text-C and Image-C) as shown in Figure C:

- **Image-R Query**: Using the original text queries from the QVHighlights dataset, we retrieved ten semantically matching images from the internet. We then employed the advanced vision-language model, InternVL [1], to compute the similarity scores between these ten images and the text. The image with the highest similarity score was selected as our final choice.
- **Image-G Query**: Consistent with the method in ICQ-Highlight [12], we generate images in four styles: scribble, cartoon, cinematic, and realistic. We carefully design specific prompts for each style. By combining the text queries from QVHighlights with these prompts, we use the text-to-image generation model Stable Diffusion [9] to produce images in these four different styles.
- **Segment-G Query**: Each text query in the QVHighlights dataset is input into GPT-4o to generate a longer, richer, more detailed text description. This generated text is then fed into the text-to-video model CogVideoX-5B [11] to produce a 6-second video, serving as the generated segment query.
  During our review of the generated data, we identify a small number of low-quality generated segments. Specifically, we find that the model occasionally generates meaningless pure white videos, whose file sizes are generally less than 150KB. To address this issue, we first filter out low-quality segments based on their file size, considering any segment smaller than 150KB as low quality. For all such segments, we re-extract the original video frames and generate more detailed, and semantically similar text queries. We continue generating new segments until their file sizes exceed 150KB. Subsequently, we manually review all the segments and, when necessary, adjust the input text for CogVideoX-5B to ensure that the generated segments meet all required standards.
- **Image-C and Text-C**: Drawing inspiration from the data construction methodology of ICQ-Highlight [12], we created Image-C and Text-C pairs through a text modification process. Specifically, the revised text queries (e.g., replacing "Swimming" with "Dancing") were fed into the [9] Stable Diffusion model to generate corresponding Image-C, while the modification deltas between original and revised texts were formulated into Text-C (e.g., "The action is swimming, not dancing"). This procedural alignment ensures that: (1) Image-C visually embodies the semantics of the revised text through diffusion-based synthesis; (2) Text-C textually articulates the modification intent.

(a) An example of our generated dataset queries.



(b) t-SNE of image features.

Figure D. Qualitative and quantitative analysis of our constructed dataset. (a) An example of different query modalities generated for a single event. (b) A t-SNE visualization showing the feature space overlap between ground-truth frames (GT), retrieved images (Image-R), and generated images (Image-G), confirming their semantic alignment.

D. Analysis of Dataset and Query Reliability

D.1. Diversity and Realism of Multi-modal Queries

To ensure the realism and diversity of our dataset, the construction process is grounded in genuine human behaviors and interests. The text queries in the base QVHighlights dataset were collected via Amazon Mechanical Turk, where annotators described events they found interesting after watching entire videos. This methodology ensures the queries reflect natural human curiosity and language. Consequently, the multi-modal queries derived from these text descriptions inherit a high degree of realism and diversity, representative of real-world user scenarios.

To further simulate user behavior and mitigate biases in the visual query construction, we employed two distinct pipelines. For the **retrieval pipeline** (Image-R), we emulate a user searching for a relevant image online. We first perform a broad image search and then use the InternVL model to select the image most semantically aligned with the text query, a process shown to correlate highly with human preference (see Section D.2). For the **generative pipeline** (Image-G), we generate images in four distinct visual styles (scribble, cartoon, cinematic, and realistic) to enhance diversity and promote model generalization, as illustrated in Figure D.

To empirically validate that our generated and retrieved queries align with real user intent, we used CLIP to extract features for Image-R, Image-G, and a ground-truth frame from the user-annotated video segment for the same text query. A t-SNE visualization of these features, shown in Figure D (b), reveals a strong overlap in the feature space. This indicates that our constructed visual queries effectively capture diverse user intentions and are semantically consistent with user-annotated ground truth content.

Table B. Statistical comparison with existing datasets.

Dataset	Text Query	Image Query	Segment Query
QVHighlights	10.3K	0	0
ICQ-Highlight	1.5K	6.2K	0
Ours	19.0K	26.3K	8.8K

D.2. Reliability of the Dataset Construction Pipeline

The quality of our dataset relies on the foundational models used in its construction. To validate the reliability of using InternVL for ranking retrieved images (Image-R), we conducted a human evaluation. We randomly selected 200 text queries and had human annotators manually rank the top 10 retrieved image candidates for each query based on semantic similarity. The Pearson correlation coefficient between the manual rankings and the InternVL rankings was 0.86, indicating a very strong positive correlation. Furthermore, the top-3 consistency, where both methods yield the identical top three results in the same order, reached 96%. These results demonstrate that InternVL's automated ranking aligns closely with human judgment, confirming the

Table C. Effect of multimodal queries on representative offline video grounding methods, evaluated on $R^1_{0.5}$. Adding visual queries (+Image-R or +Segment-G) consistently improves performance over text-only queries.

Method	Publication	Text	+Image-R	+Segment-G
Moment-DETR	NeurIPS21	53.94	55.28 (**+1.34**)	55.37 (**+1.43**)
QD-DETR	CVPR23	62.68	64.12 (**+1.44**)	64.03 (**+1.35**)
TaskWeave	CVPR24	64.26	65.37 (**+1.11**)	65.84 (**+1.58**)

method's reliability and scalability for dataset construction. For all generated queries (Image-G and Segment-G), a manual filtering process was also applied to ensure high quality and semantic relevance, as detailed in Section 4.2.

D.3. Statistical Comparison with Existing Datasets

Our QVHighlights-Unify dataset significantly expands upon existing benchmarks for video grounding by incorporating a greater number of modalities and a larger volume of query samples. Table B provides a statistical comparison with the original QVHighlights and the ICQ-Highlight datasets. Our dataset not only includes the original text queries but also introduces multiple new visual query types (retrieved images, generated images) and video segment queries, resulting in a more comprehensive and practical benchmark for the OVG-HQ task.

E. Additional Experimental Results

E.1. Utility of Multimodal Queries in Offline Grounding Methods

To demonstrate that the benefit of hybrid-modal queries is not limited to our online framework, we adapted three representative offline video grounding methods (Moment-DETR, QD-DETR, TaskWeave) to accept multimodal inputs. We augmented their architectures with image and video encoders alongside their original text encoders and evaluated them on our dataset. As shown in Table C, providing visual queries (Image-R or Segment-G) in addition to text consistently improves the performance of these established offline methods. This confirms the general utility of using hybrid-modal queries for video temporal grounding tasks.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 3

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics (AISTATS)*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 1

[3] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:11846–11858, 2021. 2

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[6] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 1

[7] 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 (CVPR)*, pages 23023–23033, 2023. 1

[8] 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. 1

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[10] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2980–2988, 2017. 1

[11] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[12] Gengyuan Zhang, Mang Ling Ada Fok, Yan Xia, Yansong Tang, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. Localizing events in videos with multimodal queries. *arXiv preprint arXiv:2406.10079*, 2024. 2, 3