Localizing Events in Videos with Multimodal Queries

Supplementary Material

In this Appendix, we present the following:

- Additional information about the dataset ICQ-Highlight and licenses for the datasets and models we have used;
- Additional technical implementations including prompts of the benchmark ICQ;
- Extended experimental results due to page limits in the main part.

A. Notations

We define the key concepts along with the notations and abbreviations used in this paper below.

Notation	Full Form	Definition
MQ	Multimodal Query	a semantic query that integrates image and text
NLQ	Natural Language Query	a semantic query expressed in text only
MQA	Multimodal Query Adaptation	an adaptation method for handling multimodal queries
SUIT	Surrogate Fine-Tuning	a fine-tuning strategy that uses a surrogate training task
v_{ref}	reference image	an image that conveys the main semantics of the query
t_{ref}	refinement text	a text used to adjust query details

Table 3. Notation Table

B. Dataset: ICQ-Highlight

B.1. License

The dataset and code are publicly accessible. We use standard licenses from the community and provide the following links to the non-commercial licenses for the datasets we used in this paper.

QVHighlights: https://github.com/jayleicn/moment_detr/blob/main/data/LICENSE

Stability Diffusion:https://github.com/Stability-AI/stablediffusion/blob/main/LICENSE

B.2. Construction Pipeline

We base our model on the original annotation from QVHighlights [42]. The whole pipeline, as shown in Fig. 7 consists of (1) *annotation*: We further conduct a quality check on the annotations in the original dataset and filter out a few samples (details can be found in Sec. B.4). In order to generate more relevant reference images, we manually augment the original captions by adding new visual details based on three frames extracted from the raw videos. To introduce refinement texts, we purposely alter certain details of the captions to generate a new one. All annotations are carried out by two individuals and evaluated by a third party for accuracy. (2) We use the augmented and altered captions to generate reference images with a suite of Text-2-Image models, including DALL-E 2 and Stability Diffusion XL for 4 variants of styles. (3) We implement an additional quality check process for all generated images to eliminate and regenerate images that might contain unsafe or counterintuitive content. We employ BLIP2 [43] to filter out generated images with lower semantic similarity with augmented captions than 0.2 and conduct a manual sanity check to control the image quality.

Data Curation and Quality check Image generation can suffer from significant imperfections in terms of semantic consistency and content safety. To address these issues, we implement a quality check in 2 stages: (1) We calculate the semantic similarity between the generated images and the text queries using BLIP2 [43] encoders, eliminating samples that score lower than 0.2; (2) We perform a human sanity check to replace images that are: i) semantically misaligned with the text, ii) mismatched with the required reference image style, iii) containing sensitive or unpleasant content (*e.g.*, violent, racial, sexual content), counterintuitive elements, or noticeable generation artifacts.

B.3. Statistics

The dataset comprises 1515 videos and 1546 test samples on average for each style. The exact numbers may vary slightly across styles and are provided in the Appendix.

Tab. 4 presents the statistics for various reference image styles in terms of the number of queries, videos, and the presence of refinement texts. Tab. 5 breaks down the statistics of refinement texts for different reference image styles across various query types: object, action, relation, attribute, environment, and others. The numbers of each type can vary slightly depending on the different styles.

Reference Image Style	#Queries	#Videos	#With Refinement Texts	#Without Refinement Texts
`scribble`	1546	1515	/	5
`cinematic`	1532	1502	1445	5
`cartoon`	1532	1501	1444	5
`realistic`	1532	1501	1446	4

Table 4. Statistics of Different Reference Image Styles

B.4. Details of Deleted Data

We removed four entries from the QVHighlight dataset that could cause violent, sexual, sensitive, or graphic content in generation in the original natural language query as listed:

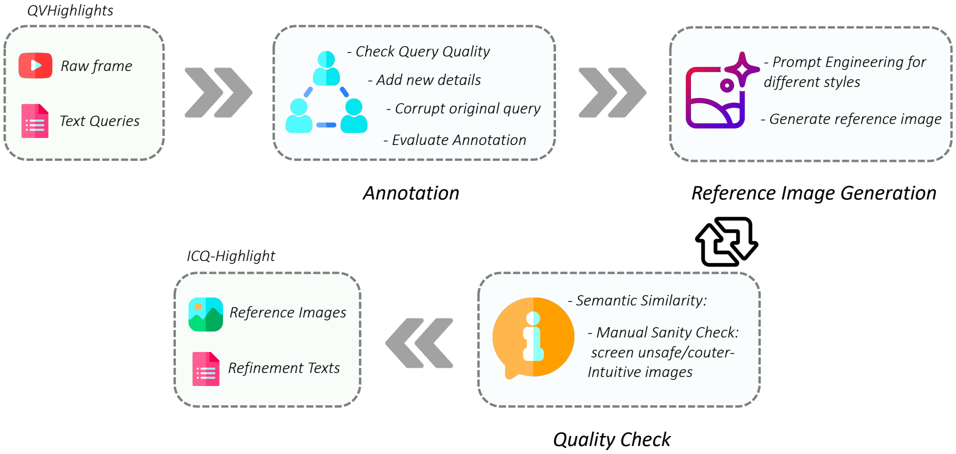


Figure 7. Dataset Construction Pipeline: We base our model with original annotations from QVHighlights and introduce a pipeline consisting of annotation, reference image generation, and quality check.

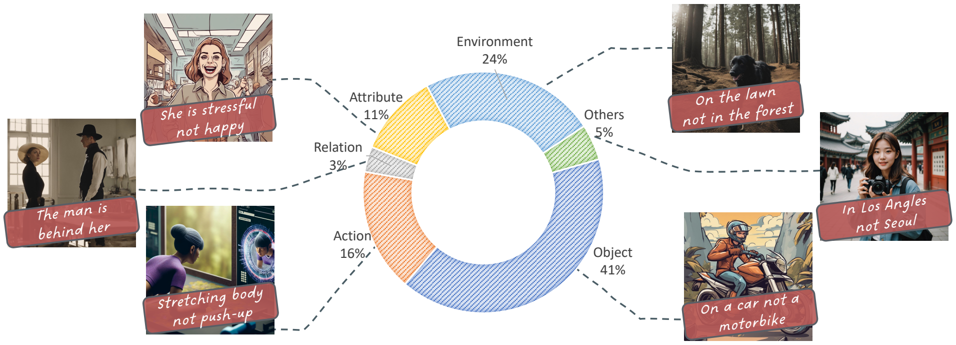


Figure 8. Distribution of Refinement Text Types. Refinement texts are designed to either *complement* or *correct* the original semantics of reference images. We identify 5 major types of refinement texts, each targeting different semantic aspects: object, action, relationship, attribute, environment, and others.

Reference Image Style	**#Queries**					
	Object	Action	Relation	Attribute	Environment	Others
scribble	594	242	50	162	343	70
cinematic	588	239	50	162	343	66
cartoon	590	239	48	161	341	68
realistic	586	241	50	161	341	70

Table 5. Statistics of Refinement Texts

- “A graph depicts penis size.” (qid: 9737)
- “People mess with the bull statues testicles.” (qid: 7787)
- “People butcher meat from a carcass.” (qid: 4023)
- “Woman films herself wearing black lingerie in the bathroom.” (qid: 7685)

C. Benchmark Details

In this section, we list the details of our selected backbone models, the implementation of our training-free MQA methods, and SUIT strategy.

C.1. Implementation Details

Automatic Pseudo-MQs Construction We build the pseudo-MQ dataset from image-text datasets Flickr30K and COCO. We generate captions for the COCO dataset with BLIP-2 [43]. To forge the original captions, we employ GPT3.5 to process the pure-text captions of each image with the prompts shown in Tab. 11. For each sample, we randomly select one template and refinement text type to generate a forged caption and the corresponding forged part as a refinement text. In total, we construct a pseudo-MQ dataset with 89 420 samples for training and 4785 samples for validation.

Model	Visual Encoder	Query Encoder	Localization Decoder*	Source
Moment-DETR (2021)	ViT-B/32 + SlowFast	CLIP Text	DETR	V
QD-DETR (2023)	ViT-B/32 + SlowFast	CLIP Text	DETR	V, V+A
EaTR (2023)	ViT-B/32 + SlowFast	CLIP Text	DETR	V
CG-DETR (2023)	ViT-B/32 + SlowFast	CLIP Text	DETR	V
TR-DETR (2024)	ViT-B/32 + SlowFast	CLIP Text	DETR	V, V+A
UMT (2022)	ViT-B/32 + SlowFast	CLIP Text	Transformer	V+A
UniVTG (2023)	ViT-B/32 + SlowFast	CLIP Text	Conv. Heads	V
UVCOM (2023)	ViT-B/32 + SlowFast	CLIP Text	Transformer Heads	V, V+A
SeViLA (2023)	ViT-G/14	CLIP Text	Glan-T5 XL (3B) [17]	V
TimeChat (2024)	ViT-G/14 + Video Q-Former	LLaMA tokenizer	LLaMA-2 7B [26]	V
VTimeLLM (2024)	ViT-L/14	LLaMA tokenizer	Vicuna v1.5 (7B) [16]	V

Table 6. Comparison of selected backbone models. *We only list the model head for the localization task if the model has multiple heads for different tasks.

Implementation of SUIT We apply LoRA to all linear layers in the language model of LLaVA-mistral-1.6 with rank $=$ 32 and alpha $=$ 64 with one epoch on the full dataset. The training takes up to 16 hours on a single NVIDIA A40 GPU.

C.2. Model Comparison

Tab. 6 compares our selected backbone models. The query encoder denotes the text encoder of each model used to encode natural language queries. Source represents the modalities of the source data, while V and A refer to "Video" and "Audio" respectively. All models have been fine-tuned on QVHilights.

C.3. Prompt Engineering

Since the performance may highly depend on the wording in a prompt, we use 3 different prompts for MQ-Cap and MQ-Sum adaptation methods. In Tab. 7, the prompts are divided into "Prompts For Style `cartoon/cinematic/realistic`" and "Prompts for `scribble`". This distinction arises because refining `scribble` images with complementary texts involves adding new details, slightly differing from other scenarios. Despite this minor variation, the prompt style remains consistent, simulating 3 different user query styles.

For MQ-Sum(+SUIT), we use the **same** prompts as MQ-Sum in the parameter-efficient fine-tuning with LoRA.

D. Extended Results

Due to the page limits, we appended additional experiments and analyses in this section.

D.1. Main Results for Other Metrics

We present the model performance in mAP in Tab. 8 as an extension to Table 1. We find that the table aligns with the results stated in Sec. 5. Our SUIT strategy demonstrates good transferability to ICQ-Highlight. We highlight this in Fig. 10 on `scribble` images and show the performance gain with MQ-Sum(+SUIT) method.

D.2. Model Performance on Different Refinement Text Types

We calculate the model performance on different subsets of refinement texts shown in Fig. 9. We conclude even though models have close performance across reference image styles, they show varied performance on different refinement text types across styles. For `scribble` style, models generally perform for "relation" better than other styles. For `cartoon` style, models demonstrate a more balanced performance across all types. The performance is notably higher for "environment" and "attribute" in `cinematic` style. Finally, for `realistic` style, the models yield better performance in "object" and "environment".

D.3. MQ-based *vs*. NLQ-based Performance

We compare model performance on the MQ-based ICQ-Highlight and the original NLQ-based QVHighlight (results taken from the original papers) using Spearman's rank correlation coefficient [69] on R1@0.5. For `scribble`, Spearman's rank correlation coefficients are 0.89(MQ-Cap) and 0.93(MQ-Sum). The `cartoon` style yields coefficients of 0.98(MQ-Cap) and 0.94(MQ-Sum). The `cinematic` style shows coefficients of 0.93 for both MQ-Cap and MQ-Sum. Lastly, `realistic` has coefficients of 0.96(MQ-Cap) and 0.95(MQ-Sum). The high correlation scores indicate a strong positive correlation across benchmarks, suggesting queries of both benchmarks share the common semantics and yield the reliability of our benchmark.

D.4. MQ-Cap Without Refinement Text *vs*. VQ-Enc

We compare the model performance between MQ-Cap without the revision step with refinement texts and VQ-Enc, as shown in Tab. 10. Both methods only use reference images as queries without refinement texts. Overall, MQ-Cap without refinement texts still significantly outperforms pure VQ-Enc, highlighting the effectiveness of image captioning. Additionally, TR-DETR and UVCOM perform best across all styles.

	Prompts For Style `cartoon/cinematic/realistic`	**Prompts For Style `scribble`**
1	I have a caption {INPUT DATA}, adjust the {MODIFICATION TYPE} from {MODIFIED DETAIL} to {ORIGINAL DETAIL}. The revised caption should remain coherent and logical without introducing any additional details.	I have a caption {INPUT DATA}. Modify it by adding {NEW TYPE} {NEW DETAIL}. The revised caption should remain coherent and logical without introducing additional details.
2	Read this {INPUT DATA}! Change the {MODIFICATION TYPE} from {MODIFIED DETAIL} to {ORIGINAL DETAIL}. Then, write a new caption that fits and doesn't add new stuff. Only give the caption, no extra words.	Read this {INPUT DATA}! Add the {NEW TYPE} {NEW DETAIL} to it. Then, write a new caption that fits and doesn't add new stuff. Only give the caption, no extra words.
3	Here's a caption {INPUT DATA}. Can you change {MODIFICATION TYPE} from {MODIFIED DETAIL} to {ORIGINAL DETAIL}? After that, make a new caption that makes sense and doesn't add anything extra. Just write the caption; no explanations are needed.	Here's a caption {INPUT DATA}. Can you add {NEW TYPE} {NEW DETAIL}? After that, make a new caption that makes sense and doesn't add anything extra. Just write the caption, no explanations needed.

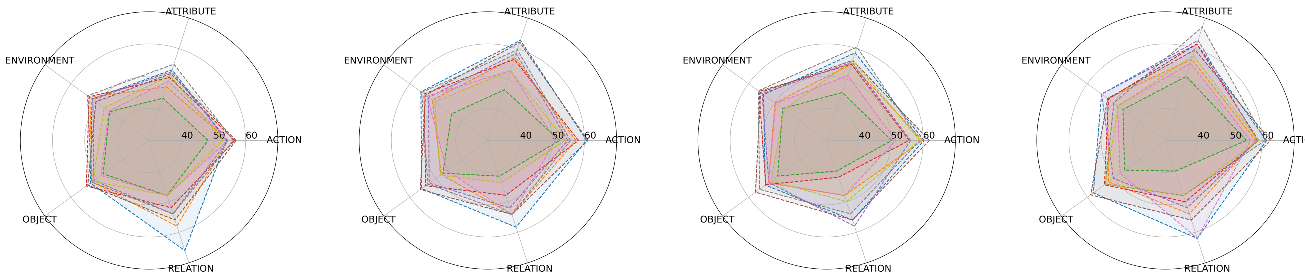
Table 7. Prompts for MQ-Cap and MQ-Sum. We use 3 different prompts and report the average performance and standard derivation in other tables.

	Model	`scribble`		`cartoon`		`cinematic`		`realistic`	
		mAP@0.5	Avg.	mAP@0.5	Avg.	mAP@0.5	Avg.	mAP@0.5	Avg.
VQ-Enc	Moment-DETR (2021)	14.95	6.67	16.51	7.21	17.00	7.39	17.41	7.66
	QD-DETR (2023)	19.48	10.11	19.57	10.18	18.07	9.54	18.88	9.94
	QD-DETR† (2023)	18.22	9.74	14.31	7.30	15.18	7.45	14.71	7.66
	EaTR (2023)	25.27	13.98	25.95	14.21	26.83	14.70	26.65	14.49
	CG-DETR (2023)	*30.24*	*15.57*	*30.78*	*15.70*	*30.07*	*15.48*	*30.98*	*15.83*
	TR-DETR (2024)	21.09	11.67	20.87	11.71	19.62	11.02	19.72	10.76
	UMT† (2022)	5.57	2.81	4.66	1.96	5.60	2.46	4.59	2.23
	UniVTG (2023)	24.30	13.02	20.80	11.56	19.85	10.99	19.42	10.95
	UVCOM (2023)	20.13	11.15	20.19	11.96	20.67	12.37	20.73	12.03
MQ-Cap	Moment-DETR (2021)	46.98 (± 2.3)	26.15 (± 1.5)	48.14 (± 1.2)	27.22 (± 0.7)	48.98 (± 0.4)	27.96 (± 0.4)	49.00 (± 0.82)	27.72 (± 0.5)
	QD-DETR (2023)	50.69 (± 3.1)	31.01 (± 2.4)	54.15 (± 0.9)	33.04 (± 0.9)	55.32 (± 0.9)	34.06 (± 0.7)	54.75 (± 0.7)	34.31 (± 0.7)
	QD-DETR† (2023)	50.78 (± 3.9)	31.44 (± 3.0)	53.91 (± 1.2)	33.94 (± 1.0)	54.06 (± 0.5)	34.67 (± 0.3)	53.82 (± 0.8)	34.18 (± 0.7)
	EaTR (2023)	*52.11 (± 2.8)*	32.88 (± 2.6)	53.23 (± 0.7)	33.60 (± 0.7)	54.00 (± 0.7)	34.54 (± 0.3)	54.36 (± 0.8)	34.73 (± 0.3)
	CG-DETR (2023)	51.13 (± 3.0)	32.13 (± 2.1)	*56.15 (± 0.8)*	36.08 (± 0.6)	55.15 (± 1.0)	35.22 (± 0.7)	*56.63 (± 0.8)*	36.57 (± 0.9)
	TR-DETR (2024)	51.07 (± 2.5)	32.15 (± 2.1)	55.72 (± 1.1)	35.98 (± 1.2)	55.87 (± 0.8)	36.29 (± 0.5)	56.32 (± 0.4)	36.76 (± 0.5)
	UMT† (2022)	42.35 (± 2.7)	26.47 (± 2.0)	45.03 (± 1.3)	28.64 (± 1.0)	46.43 (± 0.8)	30.01 (± 0.7)	45.93 (± 0.8)	29.67 (± 0.8)
	UniVTG (2023)	40.68 (± 2.5)	24.71 (± 1.9)	42.68 (± 0.7)	26.03 (± 0.6)	43.53 (± 0.4)	26.43 (± 0.5)	43.64 (± 0.8)	26.76 (± 0.5)
	UVCOM (2023)	51.27 (± 3.2)	*33.39 (± 2.5)*	54.40 (± 0.7)	*36.50 (± 0.7)*	*55.99 (± 0.7)*	*37.11 (± 0.3)*	54.98 (± 0.8)	*36.83 (± 0.6)*
	SeViLA (2023)	14.45 (± 0.8)	9.30 (± 0.6)	19.52 (± 0.5)	13.12 (± 0.4)	22.16 (± 0.3)	14.64 (± 0.4)	22.48 (± 0.6)	14.55 (± 0.5)
	TimeChat (2024)	9.08 (± 0.6)	4.45 (± 0.4)	11.01 (± 0.9)	5.13 (± 0.5)	10.58 (± 0.7)	4.82 (± 1.0)	10.69 (± 1.0)	4.78 (± 0.2)
	VTimeLLM (2024)	18.48 (± 1.0)	8.15 (± 0.5)	21.90 (± 0.3)	9.16 (± 0.1)	24.03 (± 0.5)	10.15 (± 0.3)	23.45 (± 0.7)	10.10 (± 0.1)
MQ-Sum	Moment-DETR (2021)	44.40 (± 2.5)	23.96 (± 1.8)	47.31 (± 2.1)	26.03 (± 1.4)	46.62 (± 1.9)	25.55 (± 1.3)	47.29 (± 2.2)	26.07 (± 1.3)
	QD-DETR (2023)	47.09 (± 2.8)	28.27 (± 2.4)	51.06 (± 3.3)	30.90 (± 2.5)	50.89 (± 3.3)	30.52 (± 2.8)	50.05 (± 3.6)	30.49 (± 2.7)
	QD-DETR† (2023)	48.10 (± 3.2)	29.49 (± 2.9)	50.72 (± 3.3)	31.11 (± 3.0)	49.94 (± 2.8)	31.38 (± 2.4)	50.30 (± 3.8)	30.85 (± 2.6)
	EaTR (2023)	*49.07 (± 2.6)*	*30.92 (± 2.0)*	50.82 (± 2.6)	31.38 (± 1.7)	50.71 (± 3.2)	31.34 (± 2.7)	51.37 (± 3.0)	32.02 (± 2.0)
	CG-DETR (2023)	48.41 (± 3.5)	29.86 (± 2.9)	52.31 (± 2.9)	33.21 (± 2.3)	51.59 (± 2.8)	32.34 (± 2.5)	52.31 (± 3.1)	32.91 (± 2.0)
	TR-DETR (2024)	46.69 (± 3.6)	29.72 (± 2.8)	*52.41 (± 2.6)*	33.48 (± 1.9)	*52.39 (± 3.1)*	33.14 (± 2.6)	*52.87 (± 3.1)*	*33.57 (± 2.5)*
	UMT† (2022)	40.99 (± 2.7)	25.88 (± 1.8)	43.03 (± 2.0)	27.02 (± 1.5)	42.88 (± 2.0)	26.73 (± 1.6)	43.89 (± 1.3)	27.38 (± 1.0)
	UniVTG (2023)	38.86 (± 2.7)	22.76 (± 1.8)	40.13 (± 2.8)	24.43 (± 1.7)	40.73 (± 2.7)	24.02 (± 1.9)	40.20 (± 2.4)	24.11 (± 1.6)
	UVCOM (2023)	47.33 (± 3.2)	30.75 (± 2.5)	52.22 (± 3.4)	*34.00 (± 2.7)*	51.37 (± 4.2)	*33.36 (± 3.1)*	51.64 (± 3.8)	33.52 (± 2.6)
	SeViLA (2023)	14.54 (± 1.7)	9.24 (± 1.3)	22.13 (± 1.8)	14.07 (± 1.1)	22.17 (± 1.4)	14.52 (± 0.9)	22.87 (± 1.8)	14.45 (± 1.3)
	TimeChat (2024)	9.12 (± 0.4)	4.07 (± 0.2)	9.63 (± 1.7)	4.64 (± 0.7)	10.18 (± 1.2)	4.94 (± 0.9)	9.46 (± 1.8)	4.16 (± 1.3)
	VTimeLLM (2024)	19.40 (± 1.4)	8.54 (± 0.4)	21.59 (± 0.8)	8.98 (± 0.4)	22.74 (± 0.3)	9.44 (± 0.3)	23.2 (± 1.6)	9.65 (± 0.7)
MQ-Sum	***+ SUIT***								
	Moment-DETR (2021)	49.46 (± 0.6)	28.36 (± 0.47)	49.01 (± 0.3)	28.0 (± 0.2)	49.32 (± 0.5)	28.07 (± 0.3)	48.39 (± 0.4)	27.34 (± 0.2)
	QD-DETR (2023)	55.82 (± 0.2)	35.19 (± 0.1)	54.12 (± 0.5)	33.94 (± 0.2)	55.05 (± 0.2)	34.59 (± 0.2)	54.62 (± 0.2)	34.45 (± 0.2)
	QD-DETR† (2023)	54.71 (± 0.5)	35.29 (± 0.3)	54.20 (± 0.1)	35.48 (± 0.2)	54.05 (± 0.17)	35.2 (± 0.4)	53.14 (± 0.6)	34.54 (± 0.2)
	EaTR (2023)	55.2 (± 0.7)	35.86 (± 0.4)	52.88 (± 0.2)	34.18 (± 0.2)	54.07 (± 0.7)	34.66 (± 0.1)	52.68 (± 0.3)	33.92 (± 0.4)
	CG-DETR (2023)	55.6 (± 0.6)	36.16 (± 0.2)	55.5 (± 0.4)	35.47 (± 0.3)	55.93 (± 0.7)	35.85 (± 0.3)	55.34 (± 0.6)	35.43 (± 0.3)
	TR-DETR (2024)	***56.75*** (± 0.4)	***37.25*** (± 0.2)	***55.76*** (± 0.2)	***36.31*** (± 0.1)	***56.36*** (± 0.5)	***36.84*** (± 0.5)	***56.18*** (± 0.3)	***37.05*** (± 0.3)
	UMT† (2022)	46.55 (± 0.3)	30.45 (± 0.3)	46.44 (± 0.6)	30.71 (± 0.3)	46.86 (± 0.4)	30.9 (± 0.3)	46.54 (± 0.2)	29.94 (± 0.2)
	UniVTG (2023)	43.36 (± 0.4)	26.87 (± 0.2)	42.2 (± 0.4)	26.42 (± 0.2)	43.23 (± 0.5)	26.81 (± 0.3)	42.89 (± 0.58)	26.45 (± 0.4)
	UVCOM (2023)	54.18 (± 0.3)	36.92 (± 0.4)	54.56 (± 0.3)	36.91 (± 0.1)	54.43 (± 0.4)	37.29 (± 0.1)	53.31 (± 0.5)	36.53 (± 0.2)

Table 8. Model performance (mAP) on ICQ. We highlight the best score in *italic* for each adaptation method and the overall best scores in **bold**. For MQ-Cap and MQ-Sum, we report the standard deviation of 3 runs with different prompts and for MQ-Sum(+SUIT) we report the average performance with different seeds in training. † uses extra audio modality.

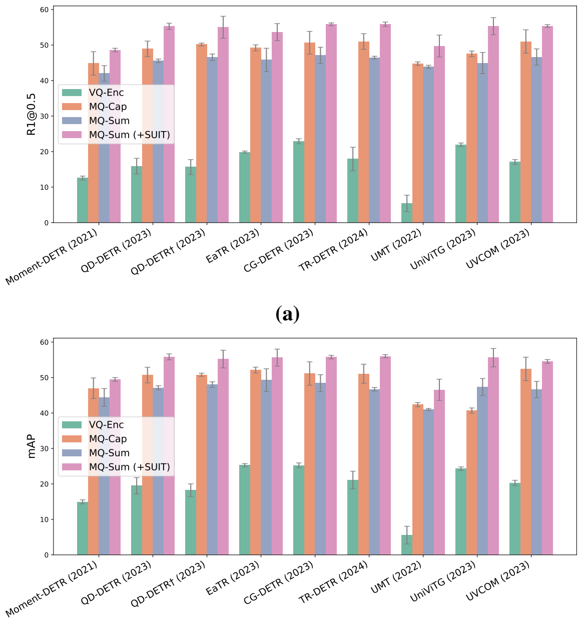
Method	original NLQ (Performance on QVHighlights)				
	R1@0.5	R1@0.7	mAP@0.5	mAP@0.7	Avg.
Moment-DETR (2021)	54.92 (-4.6%)	36.87 (-3.3%)	55.95 (-4.2%)	31.59 (-4.5%)	32.54 (-3.8%)
QD-DETR (2023)	62.87 (-8.6%)	46.70 (-12.5%)	62.66 (-7.6%)	41.59 (-12.4%)	41.23 (-10.3%)
QD-DETR† (2023)	63.71 (-6.2%)	47.67 (-8.1%)	62.9 (-5.6%)	42.07 (-6.6%)	41.73 (-6.4%)
EaTR (2023)	60.93 (-8.0%)	46.12 (-9.5%)	62.01 (-5.9%)	42.11 (-7.6%)	41.39 (-6.7%)
CG-DETR (2023)	67.27 (-8.9%)	51.94 (-13.6%)	65.48 (-7.6%)	45.64 (-12.4%)	44.88 (-11.3%)
TR-DETR (2024)	67.08 (-7.5%)	51.36 (-8.3%)	66.20 (-7.3%)	46.28 (-9.3%)	44.99 (-8.1%)
UMT† (2022)	60.22 (-10.0%)	44.24 (-14.1%)	56.62 (-9.5%)	39.85 (-15.2%)	38.54 (-12.9%)
UniVTG (2023)	59.70 (-8.7%)	40.82 (-7.2%)	51.22 (-8.0%)	32.84 (-9.9%)	32.53 (-9.0%)
UVCOM (2023)	65.01 (-5.6%)	51.75 (-8.0%)	64.88 (-5.3%)	46.96 (-9.0%)	45.83 (-8.2%)
SeViLA (2023)	56.57 (-56.2%)	40.45 (-62.1%)	47.14 (-56.8%)	32.69 (-62.3%)	33.10 (-60.6%)

Table 9. Performance comparison between the original NLQ (in QVHighlights) and forged NLQ with refinement texts introduced in ICQ-Highlight. The performance drop highlighted in the parenthesis indicates that the modifications on natural language query are non-trivial. † indicates the usage of additional audio modality.



(a) scribble **(b)** cartoon **(c)** cinematic **(d)** realistic

Figure 9. Model performance on different subsets of refinement text types. We observe that model performance with different refinement text types varies across styles.



(a)

(b)

Figure 10. Model performance between different MQA methods on `scribble`.

D.5. Original NLQs in QVHighlights vs. Forged NLQs in ICQ-Highlight

We have evaluated the model performance based on the original NLQs in QVHighlights and our refinement texts introduced in MQs to assess the significance of the refinement texts and the sensitivity of different models to NLQs. [60] points out that the impact of the NLQs may be minimal for some existing models, such as Moment-DETR. As shown in Tab. 9, Moment-DETR exhibits relatively smaller drops across all metrics, supporting this claim. On the other hand, the latest models, such as CG-DETR and TR-DETR, experience more significant performance drops, indicating a higher sensitivity to query modifications. Furthermore, SeViLA is extremely sensitive to query modifications, shown by severe performance declines across all evaluated metrics. Overall, the considerable performance decline across various models demonstrates that our modifications significantly affect the original queries. This also shows that our introduced refinement texts are not semantically trivial for localizing with MQs.

D.6. Case Study: the Impact of Potential Generation Artifact

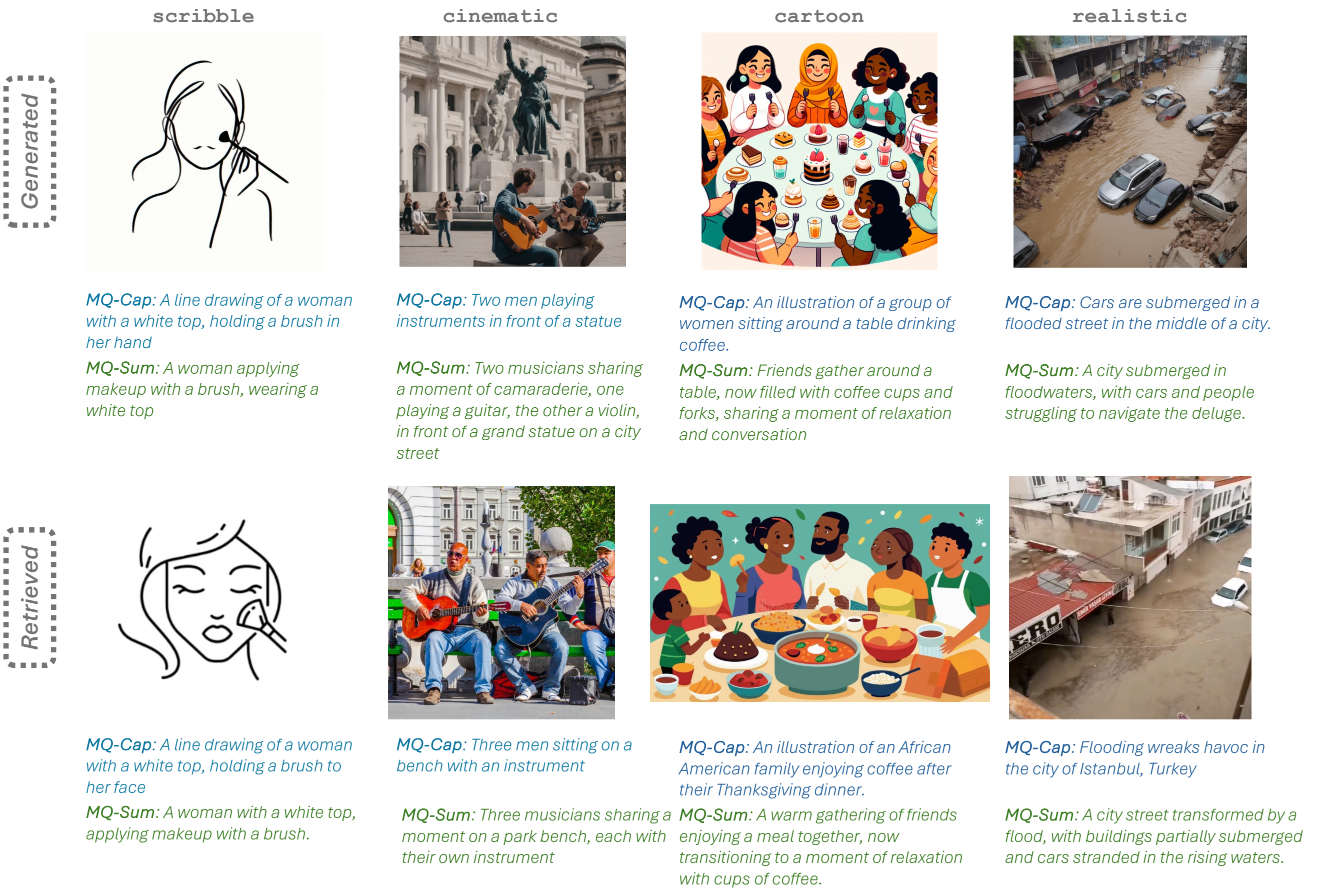
Along with the controlled experiment shown in Sec. 5.3, we conduct a qualitative case study with samples in the subsets

	Model	`scribble`		`cartoon`		`cinematic`		`realistic`	
		R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7	R1@0.5	R1@0.7
MQ-Cap wo/ revision	Moment-DETR (2021)	45.15	28.72	43.60	27.94	44.06	29.70	44.06	28.98
	QD-DETR (2023)	49.81	33.70	49.87	34.33	49.67	34.73	50.52	35.25
	QD-DETR† (2023)	51.29	36.03	48.69	33.88	49.48	34.99	49.93	35.05
	EaTR (2023)	52.01	37.77	47.45	33.09	48.56	34.33	49.61	35.64
	CG-DETR (2023)	51.42	37.84	49.35	35.90	48.89	34.79	51.04	36.55
	TR-DETR (2024)	52.01	37.19	51.04	36.62	50.00	36.03	**52.28**	**37.53**
	UMT† (2022)	46.25	31.57	45.82	30.61	46.34	29.96	46.08	31.85
	UniVTG (2023)	47.87	33.76	45.56	29.24	45.43	29.05	46.80	30.42
	UVCOM (2023)	**52.26**	**39.39**	**51.50**	**37.99**	**50.98**	**36.75**	51.70	**37.53**
VQ-Enc	Moment-DETR (2021)	12.55	5.69	13.38	6.59	14.36	6.01	14.88	6.53
	QD-DETR (2023)	15.91	9.12	14.88	8.62	13.90	8.49	14.62	8.36
	QD-DETR† (2023)	15.65	10.03	12.60	6.79	12.34	6.72	12.34	7.44
	EaTR (2023)	19.86	**13.00**	19.91	12.99	21.15	**13.45**	21.48	13.38
	CG-DETR (2023)	**22.90**	**13.00**	**24.93**	13.58	**23.24**	13.12	**24.74**	**14.23**
	TR-DETR (2024)	17.92	11.19	17.36	11.10	15.14	9.86	15.60	9.53
	UMT† (2022)	5.43	2.85	4.77	2.09	5.22	2.35	4.57	2.42
	UniVTG (2023)	21.93	13.00	23.89	**13.64**	22.78	13.19	22.52	12.79
	UVCOM (2023)	17.08	9.77	16.78	10.97	17.36	11.68	17.10	11.23

Table 10. Model performance (Recall) of MQ-Cap without refinement text and VQ-Enc on ICQ. We highlight the best score in **bold** for both methods and reference image style.

D_{gen} and D_{ret}. We notice that generation artifacts usually do not change the image semantics and thus do not influence the caption dramatically, as shown in Fig. 11.

While collecting this subset, we noticed that AI-generated images become more prevalent on the Internet. This indicates that our generated dataset has a more realistic application and reflects the practical scenarios when users aim to locate events with generated images online. In addition, we find that generation artifacts do not pose significant issues in scribble and cartoon styles since the images are already simple.



MQ-Cap: A line drawing of a woman with a white top, holding a brush in her hand

MQ-Sum: A woman applying makeup with a brush, wearing a white top

MQ-Cap: Two men playing instruments in front of a statue

MQ-Sum: Two musicians sharing a moment of camaraderie, one playing a guitar, the other a violin, in front of a grand statue on a city street

MQ-Cap: An illustration of a group of women sitting around a table drinking coffee.

MQ-Sum: Friends gather around a table, now filled with coffee cups and forks, sharing a moment of relaxation and conversation

MQ-Cap: Cars are submerged in a flooded street in the middle of a city.

MQ-Sum: A city submerged in floodwaters, with cars and people struggling to navigate the deluge.

MQ-Cap: A line drawing of a woman with a white top, holding a brush to her face

MQ-Sum: A woman with a white top, applying makeup with a brush.

MQ-Cap: Three men sitting on a bench with an instrument

MQ-Sum: Three musicians sharing a moment on a park bench, each with their own instrument

MQ-Cap: An illustration of an African American family enjoying coffee after their Thanksgiving dinner.

MQ-Sum: A warm gathering of friends enjoying a meal together, now transitioning to a moment of relaxation with cups of coffee.

MQ-Cap: Flooding wreaks havoc in the city of Istanbul, Turkey

MQ-Sum: A city street transformed by a flood, with buildings partially submerged and cars stranded in the rising waters.

Figure 11. We showcase four examples in our subsets D_{gen} and D_{ret}. We notice that image generation artifacts usually do not change the image semantics dramatically and thus do not influence the caption directly. *Please note that the retrieved images provided are for research purposes only. Distribution or sharing of these images without proper authorization is strictly prohibited.*

Type	Prompt
Object	In this task, you are given an input sentence. Your job is to generate a sentence with a different meaning by only changing the main entities (subject, object, people, animal, ...) in the input sentence, the others remain unchanged, make sure the modified sentence are still reasonable. Only output the modified sentence, do not include explanations. Input sentence: "{}". Output:
Attributes	In this task, you are given an input sentence. Your job is to generate a sentence with a different meaning by only changing the attributes (such as color, size, shape, texture, ...) of the objects in the input sentence, the others remain unchanged, make sure the modified sentence are still reasonable. Only output the modified sentence, do not include explanations. Input sentence: "{}". Output:
Actions	In this task, you are given an input sentence. Your job is to generate a sentence with a different meaning by only changing the action verbs in the input sentence, the others remain unchanged, make sure the modified sentence are still reasonable. Only output the modified sentence, do not include explanations. Input sentence: "{}". Output:
Environment	In this task, you are given an input sentence. Your job is to generate a sentence with a different meaning by only changing the environment (focus on 'where', such as the background, location, atmosphere, settings...) in the input sentence, the others remain unchanged, make sure the modified sentence are still reasonable. Only output the modified sentence, do not include explanations. Input sentence: "{}". Output:
Relations	In this task, you are given an input sentence. Your job is to generate a sentence with a different meaning by only changing the relationships (focus on the relationship between different entities, such as spatial, temporal, interaction-based connections, ...) in the input sentence, make sure the modified sentence are still reasonable. Only output the modified sentence, do not include explanations. Input sentence:"{}". Output:

Table 11. Examples of prompt templates used to generate forged captions with GPT3.5.