CaptAin: Caption-driven Alignment for Bridging Modality Gaps in Partially Relevant Video Retrieval

Supplementary Material

We have organized the supplementary materials into the following sections. In Section 1, we provide more implementation details, as well as evaluation datasets, and the experimental environment. In Section 2, we further present the results of hyperparameter experiments and include additional visualization results.

1. More Implementation Details

1.1. Details of Prompts in GL-ACG

Fig. 1 depicts the three prompts employed in GL-ACG.

(1) Video Summary Generation Prompt. Uniformly sampling 72 frames from the video, this prompt synthesizes the full content to produce more detailed, comprehensive global captions than segment captions. When generating captions, 5 random training set queries are used as references to avoid style-inconsistent outputs, inconsistent data distribution, and heightened modality alignment challenges.

(2) Video Segment Description Prompt. For atomic event description, this prompt processes $U = 5$ equal temporal segments of the video, each uniformly sampled to extract 12 frames. It also incorporates in-context learning (as with Video Summary Generation Prompt) to mimic the native dataset's query style.

(3) Redundancy-Aware Merging Prompt. Since splitting videos into 5 equal segments may cause event overlap between adjacent segments (see Tab. 6 in the main paper), which impairs the construction of positive and negative samples in L_{ll} and accuracy of TTME, this prompt feeds the 5 initial local captions into the captioner to merge overlapping content between adjacent ones. It ensures each final local caption describes a relatively independent atomic event (enhancing subsequent utility) while requiring the retention of at least two captions to avoid conflation into a global summary.

1.2. Details of Video-Text Matching

Since the Video-Text (V-T) Matching differs from the baselines, we take the V-T Matching process of MS-SL [2] as an example. The V-T Matching score $\mathcal{F}_{vt}$ is:

$$\mathcal{F}_{vt}(V,T) = \sigma\mathcal{F}_s(V,T) + (1-\sigma)\mathcal{F}_a(V,T), \quad (1)$$

$$\mathcal{F}_s(V,T) = \max\{\cos(\mathbf{s}_i, \mathbf{T})\}_{i=1}^{K}, \quad (2)$$

$$\mathcal{F}_a(V,T) = \cos(\mathbf{a}, \mathbf{T}), \quad (3)$$

where $\mathcal{F}_s$ and $\mathcal{F}_a$ denote segment-query and frame-query matching score, respectively. $\{\mathbf{s}_i\}_{i=1}^{K}$ denote K segment video features. Here, K is set to 32. $\mathbf{T}$ is a text query feature. $\mathbf{a}$ is an aggregated frame feature. σ is set to 0.5 in MS-SL.

1.3. Details of $\mathcal{L}_{gg}$

Since our CaptAin is plug-and-play and applicable to various approaches, $\mathcal{L}_{gg}$ is actually inconsistent across different baselines. Here, as in the main paper, we use MS-SL [2] as an example to introduce $\mathcal{L}_{gg}$. Following MS-SL, we jointly use the triplet ranking loss [1, 3] and InfoNCE loss [6, 7]. The triplet ranking loss is defined as:

$$\mathcal{L}^t = \frac{1}{B}\sum_{i=1}^{B}[\max(0, m - \mathcal{F}_{s|a}(V_i,T_i) + \mathcal{F}_{s|a}(V_i,T_i^-)) + \max(0, m - \mathcal{F}_{s|a}(V_i,T_i) + \mathcal{F}_{s|a}(V_i^-,T_i))], \quad (4)$$

where m denotes the margin constant, we set it to 0.2 following [2]. Here, $\mathcal{F}_{s|a}(\cdot,\cdot)$ represents applying either $\mathcal{F}_s$ or $\mathcal{F}_a$. The losses $\mathcal{L}^{ts}$ and $\mathcal{L}^{ta}$ denote the use of $\mathcal{F}_s$ and $\mathcal{F}_a$ as the matching score, respectively. Besides, $\{V_i, T_i\}$ represents a positive video-text pair, while V_i^- and T_i^- denote negative video and text samples corresponding to V_i and T_i, respectively. Negative samples are initially selected randomly within each batch and become the hardest negative samples after 20 epochs. B is the batch size.

The InfoNCE loss in MS-SL is defined as:

$$\mathcal{L}^I = \frac{1}{B}\sum_{i=1}^{B}\Big[\log\frac{\exp(\mathcal{F}_{s|a}(V_i,T_i)/\tau)}{\sum_{j=1}^{B}\exp(\mathcal{F}_{s|a}(V_j,T_i)/\tau)} + \log\frac{\exp(\mathcal{F}_{s|a}(V_i,T_i)/\tau)}{\sum_{j=1}^{B}\exp(\mathcal{F}_{s|a}(V_i,T_j)/\tau)}\Big], \quad (5)$$

where τ is a learnable temperature coefficient. The losses $\mathcal{L}^{Is}$ and $\mathcal{L}^{Ia}$ denote the use of $\mathcal{F}_s$ and $\mathcal{F}_a$ as the matching score, respectively.

Video Summary Generation Prompt	Video Segment Description Prompt	Redundancy-Aware Merging Prompt
Your task is to generate a concise paragraph describing the main content of the entire video, based on the sampled frames. Use simple words and short sentences, following the style and length of the provided examples. Focus on the main actions, subjects, or events, and avoid complex vocabulary or unnecessary details. The description will be used in video retrieval. Here are reference examples: {examples} Please ensure that the pronouns used in the description are correct. Now, summarize the provided video frames in one paragraph. Output only the paragraph.	Your task is to generate a concise caption that describes the key content of these continuous video frames as a whole. Avoid overly detailed descriptions and focus on the core actions, subjects, and events. Each caption should be a short, straightforward sentence, similar to how users would search for a video. Here are examples of user search queries for video segments, which reflect the desired style and length of your output: {examples} Based on these examples, generate a caption for the provided video frames that adheres to the same simple, direct language and length. Do not include additional explanatory details or elaborate on the background elements.	Given a list of video caption segments (each with start_time, end_time, and caption) which is in chronological order, merge only adjacent segments whose captions are semantically overlapping or describe the same event. Do not merge segments that are just temporally adjacent but describe different content. After merging, ensure the total number of captions is at least 2. For each merged segment, combine all information from the original captions (must avoiding information loss!), avoid redundancy, and write fluent, simple English. Input segments (JSON array): {segments} Only return the merged captions in JSON format.

Figure 1. Prompts in GL-ACG: The Video Summary Generation Prompt is used to generate a global caption for each video; the Video Segment Description Prompt is used to generate 5 local captions for each video; and the Redundancy-Aware Merging Prompt is used to merge these 5 local captions based on the overlap between adjacent ones.

The overall L_{gg} of MS-SL is:

$$\mathcal{L}_{gg} = \mathcal{L}^{ts} + \mathcal{L}^{ta} + \omega_1 \mathcal{L}^{Is} + \omega_2 \mathcal{L}^{Ia}, \tag{6}$$

where ω_1 and ω_2 are set to 0.02 and 0.04, respectively.

1.4. More Details of Datasets.

We evaluate our CaptAin on three PRVR datasets: Charades-STA [4], ActivityNet-Caption [5] and YouCook2-PRVR. The details of these datasets are as follows:

Charades-STA consists of 6,670 untrimmed videos paired with 16,128 sentence descriptions. On average, each video contains about 2.4 temporal moments, and each sentence corresponds to a specific segment within the video. The average video duration is around 30 seconds, while each annotated moment typically covers only a small portion of the entire video, which aligns well with the partially relevant retrieval setting. Following the split protocol adopted by all previous PRVR methods, we use 5,538 videos with 12,408 captions for training, and 1,334 videos with 3,720 captions for testing.

ActivityNet-Captions contains approximately 20K untrimmed YouTube videos with a total duration exceeding 849 hours and an average video length of 118 seconds. Each video is associated with 3–4 natural language descriptions, each describing a semantically meaningful temporal segment. On average, each moment lasts about 36 seconds, covering roughly 30% of the video duration, which makes the dataset suitable for partially relevant video retrieval. Following the data splits used in all prior PRVR works, we use 9,043 videos with 33,721 captions for training and 4,430 videos with 15,753 captions for testing.

Table A. Effect of In-context Learning (ICL) in prompts.

GL-ACG	Charades-STA				
	R@1	R@5	R@10	R@100	SumR
w/o ICL	2.8	9.7	16.1	57.1	85.7
w/ ICL	**3.4**	**11.3**	**17.7**	**58.7**	**91.1**

YouCook2-PRVR is built from the YouCook2 [8] dataset, with an important difference in dataset utilization compared to traditional Text-to-Video Retrieval (T2VR) tasks. Previous T2VR tasks pair each text query with specific trimmed video segment timestamps for retrieval, whereas YouCook2-PRVR removes all timestamp annotations and requires models to retrieve the entire untrimmed video for each query, leading to a more challenging and realistic retrieval scenario. The dataset contains 2,000 instructional cooking videos across 89 recipes and 15,433 sentence descriptions. The videos are relatively long, with an average duration of 315.4 seconds, while the relevant content typically occupies only 6.3% of the video, further increasing the task difficulty. The original YouCook2 dataset is divided into training, validation, and testing subsets, containing 10,337, 3,492, and 1,604 sentence descriptions, respectively. In our experiments, only the training and validation subsets are used for model training and evaluation, as the testing set does not release corresponding sentence descriptions for retrieval evaluation.

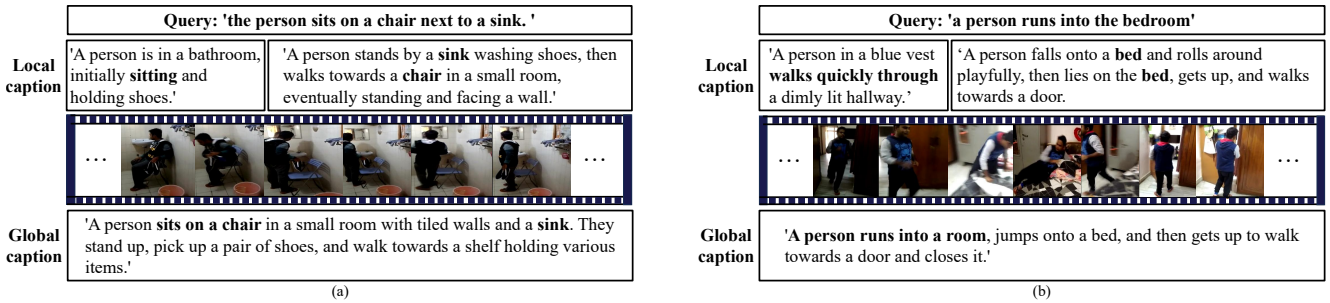

Figure 2. Visualizations show that query information is fragmented across local captions, reducing their matching scores. Global captions aggregate this information, addressing the issue and improving retrieval performance. (a): Rank improves from 12 (w/o global caption) to 1 (w/ global caption). (b): Rank improves from 18 to 1.

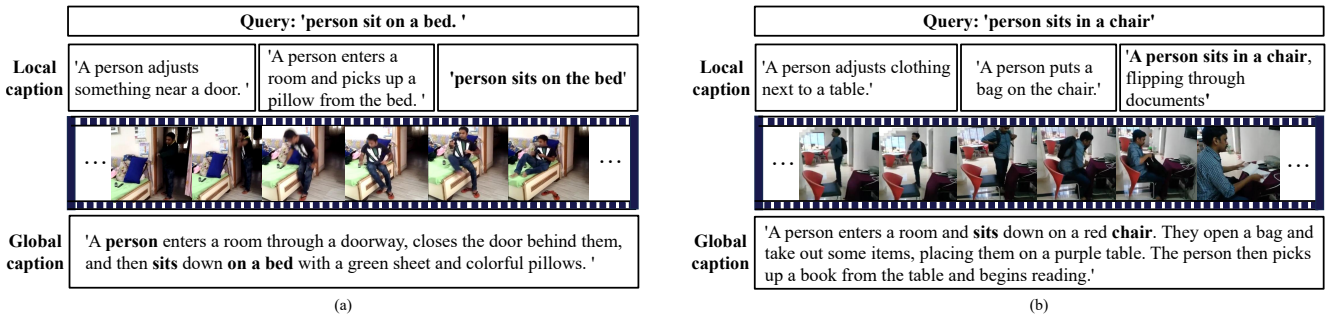

Figure 3. Visualizations show that global captions are often overly long and disperse query-relevant details across multiple sentences, reducing information density after global embedding compression and lowering matching scores. In contrast, local captions concisely capture the query content, addressing this issue and improving retrieval. (a): Rank improves from 13 (w/o local captions) to 1 (w/ local captions). (b): Rank improves from 12 to 1.

Table B. Effect of μ in TTME on Charades-STA.

μ	R@1	R@5	R@10	R@100	SumR
0.0	3.2	10.3	16.5	57.2	87.2
0.1	3.3	10.6	17.0	58.0	88.9
0.2	3.2	11.1	17.2	58.4	90.0
0.3	**3.4**	**11.3**	17.7	**58.7**	**91.1**
0.4	3.3	11.3	17.8	58.7	91.1
0.5	3.3	11.0	**17.9**	58.3	90.5
0.6	3.1	10.6	17.7	58.3	89.7
0.7	2.8	10.4	17.4	58.0	88.6
0.8	2.9	10.1	16.9	57.8	87.7
0.9	2.7	10.0	16.2	57.2	86.1
1.0	2.4	9.5	15.4	56.0	83.3

2. Additional Experiments and Visualizations

Similar to the ablation studies, we perform additional experiments on the Charades-STA dataset.

Effect of in-context learning in prompts of GL-ACG. As shown in Tab. A, without in-context learning, generated captions fail to match the style of the original dataset's query sentences. This mismatch hinders the model's ability to accurately grasp and follow the intended language patterns during training, increasing learning difficulty as the model must additionally adapt to semantic biases caused by style differences. This led to a SumR drop from 91.1 to 85.7 on Charades-STA.

Effect of μ in TTME. As shown in Tab. B, we conducted hyperparameter experiments on the weight coefficient μ in the Text-Text Matching Score. The best performance was achieved when $\mu = 0.3$, with SumR reaching 91.1. In contrast, when $\mu = 0.0$ and $\mu = 1.0$, the SumR values were 87.2 (-4.28%) and 83.3 (-8.56%), respectively, highlighting the complementary roles of local and global captions. To further demonstrate the necessity of both captions, we provide visual examples (see Fig. 2 and Fig. 3).

As shown in Fig. 2(a), for the query "the person sits on a chair next to a sink," local captions only partially match the query (*e.g.*,"sitting" or "sink & chair"), whereas the global caption captures the full context, improving the rank from 12 to 1. In (b), for the query "person runs into the bedroom," local captions offer partial matches (*e.g.*, "walks quickly through" or "bed"), while the global caption boosts the rank from 18 to 1. However, global captions are not always superior to local captions. In Fig. 3(a), for the query "person sits

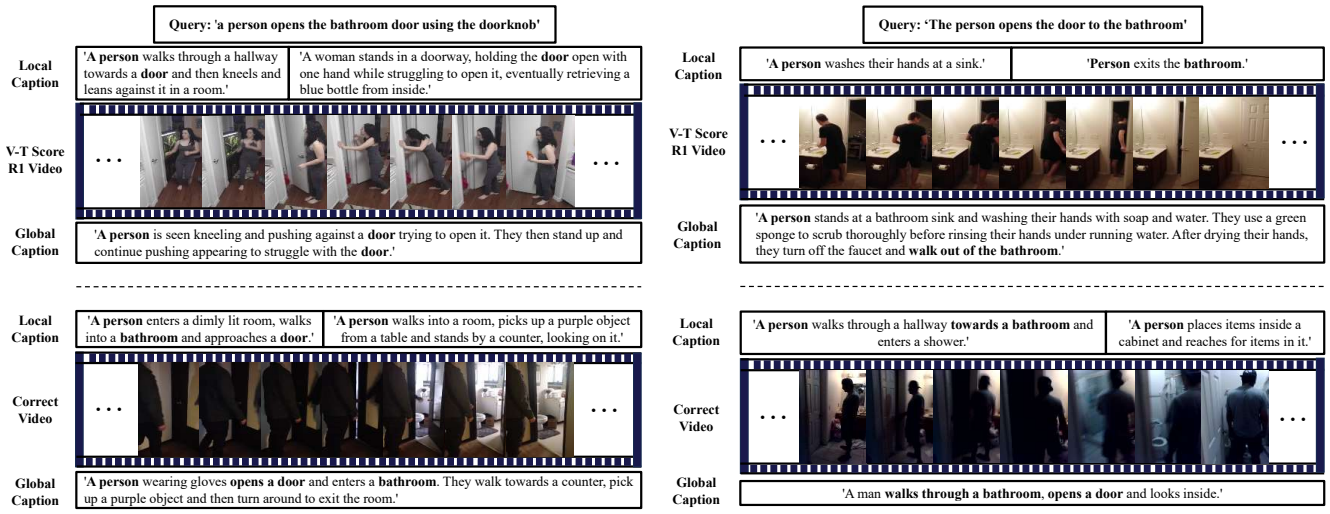

Figure 4. Visualizations demonstrate the role of TTME. Without TTME, V-T matching fails to align with the queries, leading to low retrieval ranks—27 in (a) and 24 in (b). With TTME, the combined information resolves these mismatches, improving all ranks to 1.

Table C. Effect of η in $\mathcal{F}$ on Charades-STA.

η	R@1	R@5	R@10	R@100	SumR
0.0	2.5	8.6	14.2	52.1	77.3
0.1	2.7	9.7	15.6	54.3	82.3
0.2	2.8	10.2	16.6	55.8	85.4
0.3	3.0	10.1	17.0	57.0	87.2
0.4	3.1	10.5	17.5	57.4	88.6
0.5	3.3	10.9	17.5	58.2	89.8
0.6	3.3	10.9	**17.8**	58.5	90.5
0.7	**3.4**	11.3	17.7	**58.7**	**91.1**
0.8	3.4	**11.6**	17.4	58.5	91.0
0.9	3.3	11.5	17.5	58.3	90.7
1.0	3.3	11.4	17.1	58.4	90.2

Table D. Effect of loss weight λ_1 in $\mathcal{L}$.

λ_1	R@1	R@5	R@10	R@100	SumR
0.30	2.0	7.8	13.3	51.6	74.7
0.35	2.3	7.8	13.8	51.4	75.3
0.40	2.4	8.1	13.7	51.2	75.3
0.45	**2.5**	8.6	**14.2**	**52.1**	**77.3**
0.50	2.4	**8.8**	14.2	51.9	77.2

Table E. Effect of loss weight λ_2 in $\mathcal{L}$.

λ_2	R@1	R@5	R@10	R@100	SumR
0.6	2.3	7.6	13.8	52.3	75.9
0.7	2.2	8.0	14.0	51.5	75.7
0.8	2.1	8.6	14.1	51.0	75.8
0.9	**2.5**	**8.6**	**14.2**	**52.1**	**77.3**
1.0	2.4	8.1	14.2	51.4	76.0

on a bed", the global caption is lengthy and yields sparse information after compression, resulting in lower performance. In contrast, the concise local caption ("person sits on the bed") fully aligns with the query, improving the rank from 13 to 1. In (b), for the query "person sits in a chair", the overlong global caption dilutes key information during compression, while the local caption ("A person sits in a chair") matches precisely, boosting the rank from 12 to 1.

Effect of η in TTME. As shown in Tab. C, moderately incorporating Text-Text (T-T) matching scores significantly boosts performance. The best result is at $\eta = 0.7$, with SumR reaching 91.1—17.8% higher than without T-T ($\eta = 0.0$, SumR=77.3), demonstrating TTME's effectiveness. Visual examples in Fig. 4 illustrate this: limited cross-modal alignment causes poor Video-Text (V-T) matching, while captions describe videos better. In (a), V-T alone ranks the correct video 27th; adding T-T scores lowers irrelevant videos' ranks (their captions lack "bathroom") and raises the correct one to Rank 1. In (b), V-T mismatches "opens the door to the bathroom" with an "walk out of the bathroom" video, but T-T corrects this, moving the true match from Rank 24 to 1.

Effect of λ_1 and λ_2 in $\mathcal{L}$. As shown in Tables D and E, SumR peaked at $\lambda_1 = 0.45$ before slightly declining. For λ_2, SumR increased steadily, reaching its highest value at 0.9 before a minor drop. The optimal performance was achieved at $\lambda_1 = 0.45$ and $\lambda_2 = 0.9$.

Table F. Effect of $\mathcal{W}$ in L_{gl} on Charades-STA.

$\mathcal{W}$	R@1	R@5	R@10	R@100	SumR
2	2.2	8.0	13.8	51.8	75.7
5	**2.5**	**8.6**	14.2	**52.1**	**77.3**
8	2.4	8.5	**14.4**	51.9	77.2

Effect of $\mathcal{W}$ in $\mathcal{L}_{gl}$. As shown in Tab. F, when the maximum window length $\mathcal{W} = 2$, the number of constructible negative samples is much lower than at $\mathcal{W} = 5$. Increasing to $\mathcal{W} = 8$ yields no further improvement. Additionally, larger window sizes strengthen the correlation between positive and negative samples, leading to lower-quality negatives. Therefore, $\mathcal{W} = 5$ is selected as optimal.

Effect of α and β in HCA. Since positive and negative samples in the intra-contrastive loss share inherent homology (*i.e.*, they are not entirely unrelated), we adopt small margins in the triplet loss: $\alpha = 0.2$ and $\beta = 0.1$, to avoid overly harsh supervision. We find that the intra-contrastive loss does not excessively separate these samples—an expected outcome, as their intrinsic similarity calls for moderate, not extreme, separation. Consequently, the gap between positive and negative samples already falls within the range defined by α and β, and increasing these values further yields no additional effect.

References

[1] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021. 1

[2] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022. 1

[3] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 1

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2

[5] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 2

[6] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 1

[7] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021. 1

[8] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2