Supplementary Materials for "Beyond the Highlights: Video Retrieval with Salient and Surrounding Contexts"

Jaehun Bang[1] Moon Ye-Bin[2] Tae-Hyun Oh[3*] Kyungdon Joo[1*]
[1] UNIST [2] POSTECH [3] KAIST
{devappendcbangj, kyungdon}@unist.ac.kr ybmoon@postech.ac.kr taehyun.oh@kaist.ac.kr

A. Experimental Details

To ensure a fair comparison across all models and datasets, we conduct all experiments in a consistent computational environment and with standardized training protocols. Specifically, we use 8 NVIDIA A100 GPUs for model finetuning and 1 GPU for evaluation. The procedure follows the implementation details and setup described in UMT [4].

Model finetuning. We follow the training setup of UMT [4], with modifications only to the batch size and training dataset. For both video-caption and clip-caption finetuning, we use a batch size of 128 across all datasets. In the video-caption setting, each video is paired with multiple captions that reflect different aspects of the content, including the original, salient, and surrounding descriptions. For clip-caption finetuning, each segmented clip is similarly matched with a set of semantically diverse captions. This one-to-many pairing strategy enables the model to learn fine-grained spatial and temporal alignments across various caption types.

B. Dataset Details

B.1. Caption Generation

To generate high-quality captions, we first design a prompt instructing GPT-4o to describe both salient and surrounding contexts within a video clip, with five complete sentences for each type. This prompt, shown in Fig. S1, enables the model to extract not only prominent foreground objects and actions but also subtle background details such as color, weather, and viewpoint. While GPT-4o can process multiple images with a text prompt, we utilize this capability to handle a sequence of frames as a coherent video. This allows the model to perceive spatio-temporal continuity and produce richer, context-aware captions.

*Corresponding author.

Text Prompt for Caption Generation.

System Prompt
You are extracting visual information from a video.

User Prompt
Please describe both the salient and surrounding information in the video in five complete sentences each.
salient information is what people will pay the most attention to in the video, such as objects and their actions located at the center.
Surrounding information is something people would easily miss, such as detailed color, material, count, and location of the object or text, background, weather, or viewpoint.
All the information must not be imagined or guessed; only information that can be clearly seen and judged when watching the video must be described. The output format must be as follows: {"salient": ["sentence1", ..., "sentence5"], "surround": ["sentence1", ..., "sentence5"]}
These are the frames from the video. `[images tokens]`

Figure S1. **Example prompt used to instruct GPT-4o.**

B.2. Quality Analysis of Generated Captions

Although GPT-4o shows strong robustness to explicit biases in vision inputs, it still has certain limitations. Prior work reports occasional implicit biases [7] and emotion-related biases [3]. To assess potential biases in GPT-4o, we sample 50 videos and analyze 950 captions. Under strict evaluation criteria, 20 captions are found to contain subtle or ambiguous biases across categories such as object presence, action, direction, display effects, and gender.

Common issues include misinterpreting visual effects (e.g., water ripples) as real motion or extracting unclear spa-

tial details in low-visibility scenes. To address these cases, we regenerate the biased captions with GPT-4o for better alignment.

B.3. Dataset Statistics

We further analyze the semantic relationships between all caption types by computing pairwise distances using Sentence-BERT embeddings. As shown in Table S2, original captions are semantically close to salient captions yet exhibit lower diversity, as reflected in their smaller pairwise distances. In contrast, their substantially larger distances to surrounding captions indicate that they primarily describe overarching salient elements rather than the surrounding context. Consequently, they tend to overlook contextual details that are essential for fine-grained video understanding.

Moreover, the distances between salient and surrounding captions are substantial, even though they are temporally aligned. This demonstrates that our dataset distinguishes between salient elements and rich surrounding details, offering diverse and complementary information. These findings validate the richness of our annotations and highlight the suitability of our dataset for evaluating models across different levels of semantic granularity.

C. Additional Experiments

In this section, we provide additional analysis on the relationship between model performance and video properties, using the correlation matrix introduced in Sec. C.1. We also report extended evaluation results for UMT [4] and InternVideo2 [8] on our datasets in terms of Recall@1, 5, and 10, as detailed in Sec. C.2. Furthermore, Sec. C.3 presents baseline results for LSMDC-SS and DiDeMo-SS that were omitted from the main paper. Finally, Sec. C.4 provides qualitative comparisons between the pretrained model and our proposed baseline.

C.1. Video Properties and Performance

We analyze the correlation between model performance and video properties, in Figs. S2 and S3. We use the following key properties of the video: the number of clips in the video, duration per clip, number of frames in the video, frames per second, duration of the video, and file size. In MSRVTT-SS, as shown in Fig. S2, we observe that two video properties have a strong correlation with retrieval performance. This trend is consistent in LSMDC-SS and DiDeMo-SS, as shown in Fig. S3, where we again observe that two factors are particularly influential: the number of clips and the average clip duration. Based on these findings, we visualize their effects in Fig. 3 of the main paper.

Video-level Analysis. The overall results show that there is a negative correlation between the number of clips per video and model performance, *i.e.*, as the number of clips increases, performance decreases. These findings suggest that high temporal complexity reflected in a large number of clips reduces performance. In contrast, longer clip durations are positively correlated with performance, likely because short clips provide insufficient context for the model to understand. This is because the model struggles to capture the context in shorter clips.

Table S1. **Zero-shot text-to-video retrieval results of GRAM and CLIP4Clip on MSRVTT-SS.** The results show a consistent performance drop from original to salient and, more noticeably, to surround queries. This highlights the challenge of retrieving videos based on fine-grained surrounding details.

Model	Query	R@1	R@5	R@10
GRAM	Original	53.50	75.00	83.40
	Salient	49.60	67.20	72.70
	Surround	16.30	27.40	30.80
CLIP4Clip	Original	30.90	54.20	63.30
	Salient	20.20	38.20	47.50
	Surround	8.90	20.40	27.90

Clip-level Analysis. To further examine the impact of temporal complexity, we also analyze a correlation matrix based on the clip database, as shown in Fig. S2. Compared to the video-based analysis, the correlation between the number of clips and clip duration is much weaker in the clip-based analysis. This suggests that using clip-level representations helps reduce the negative effects of scene transitions on performance.

C.2. Additional Evaluation Results

We evaluate UMT [4] and InternVideo2 [8], regarding the salient and surrounding contexts in Sec. 4.1 of the main paper. The full version of Tables 6 and 7 of the main paper can be found in Table S3 and S4, respectively. These extended results are consistent with those presented in the main paper.

In addition, we report results of GRAM [2] and CLIP4Clip [5], two representative video retrieval models. GRAM is a recent model that achieves SOTA specifically on MSRVTT in the zero-shot setting, while InternVideo2 remains the leading model on LSMDC and DiDeMo. CLIP4Clip serves as a widely used standard baseline in video-text retrieval. As shown in Table S1, GRAM and CLIP4Clip exhibit the same trend observed in other models. This consistent pattern across UMT, InternVideo2, GRAM, and CLIP4Clip highlights that retrieving videos based on fine-grained surrounding context remains a significant challenge. It further supports the need for our dataset and evaluation setup to advance spatio-temporal understanding in video retrieval.

C.3. Additional Results of Baselines

Finetuning with video-caption pairs. The simplest baseline is finetuning the model with video-caption pairs that include both salient and surrounding captions. The result of MSRVTT-SS is in Table 8 of the main paper, and other datasets are in Table S5.

Finetuning with clip-caption pairs. We finetune the model using clip-caption pairs and evaluate its performance on the clip database. The results for MSRVTT-SS are in Table 9 of the main paper, while results for other datasets can be found in Table S6.

Most of the finetuned models exhibit improved performance after the finetuning process. However, when compared to the models finetuned with video-caption pairs, performance on the original query is lower for MSRVTT-SS and LSMDC-SS. This may be because DiDeMo [1] has multiple clip-wise captions per video, while MSRVTT [9] and LSMDC [6] datasets include only one query per video. As a result, finetuning with video pairs may be more effective for the original query. Overall, the results suggest that finetuning with clip-caption pairs helps the model deals with both temporal complexities and spatial details.

C.4. Qualitative Results

We compare the retrieval results of the existing zero-shot model with those of our baseline finetuned on video-caption pairs, as shown in Fig. S4. The results show that while the zero-shot model struggles to retrieve videos based on surrounding context, our baseline successfully identifies the correct video for such queries. For instance, although the zero-shot model fails to capture temporal details, our baseline correctly detects the red cushion, which appears only briefly in the video as a surrounding object. In addition, for salient object queries, our baseline retrieves highly relevant dog-related videos consistently within the top 2 to top 4 results, unlike the zero-shot model. These results demonstrate a clear improvement in spatio-temporal understanding by our baseline.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[2] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. *arXiv preprint arXiv:2412.11959*, 2024. 2

[3] Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. Visbias: Measuring explicit and implicit social biases in vision language models. *arXiv preprint arXiv:2503.07575*, 2025. 1

[4] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[5] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[6] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[7] Payam Saeedi, Mahsa Goodarzi, and M Abdullah Canbaz. Heuristics and biases in ai decision-making: Implications for responsible agi. In *2025 6th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pages 214–221. IEEE, 2025. 1

[8] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[9] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

Table S2. **Semantic distance between caption types.** We report semantic distances between different caption types (original, salient, surround) using Sentence-BERT embeddings. The results reveal that original captions are semantically closer to salient captions but distant from surround captions, indicating that original captions lack diverse surrounding details. Additionally, despite temporal alignment, salient and surround captions exhibit clear semantic separation, suggesting that our dataset captures complementary spatio-temporal information with fine-grained detail.

Diversity	Query	MSRVTT			LSMDC			DiDeMo		
		Original	Salient	Surround	Original	Salient	Surround	Original	Salient	Surround
Train	Original	0.720	0.937	1.062	-	1.218	1.300	0.632	0.977	1.105
	Salient	0.916	0.755	0.952	0.966	0.738	0.983	0.961	0.672	0.980
	Surround	1.012	0.929	0.789	1.051	0.967	0.783	1.048	0.933	0.745
Validation	Original	0.726	0.932	1.058	-	1.217	1.304	0.641	0.981	1.097
	Salient	0.920	0.748	0.959	0.964	0.742	0.974	0.958	0.686	0.976
	Surround	1.024	0.937	0.783	1.060	0.965	0.756	1.039	0.936	0.748
Test	Original	-	1.190	1.290	-	1.224	1.300	0.638	0.971	1.103
	Salient	0.920	0.754	0.955	0.966	0.750	0.982	0.958	0.671	0.990
	Surround	1.022	0.933	0.786	1.048	0.972	0.766	1.045	0.939	0.749

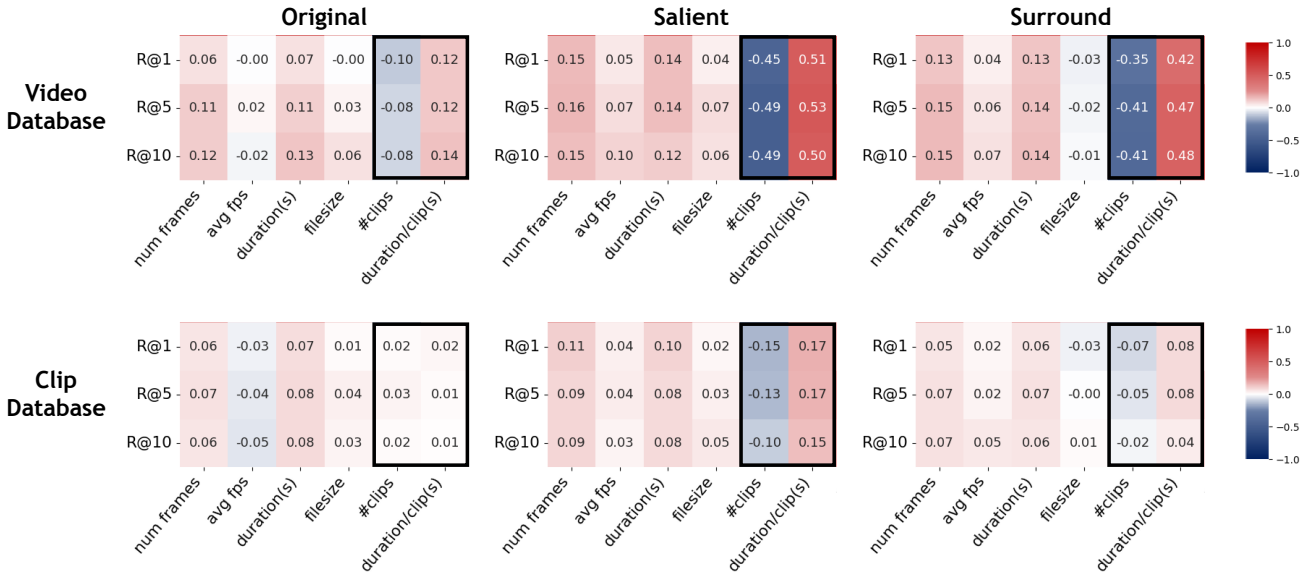

Figure S2. **Correlation matrix of MSRVTT-SS.** This is a correlation analysis using the pretrained UMT-L model on video and clip databases. The number of clips shows a negative correlation with Recall, while the duration per clip shows a positive correlation. These correlations are reduced when using the clip database, suggesting that clip-level retrieval mitigates the negative impact of scene transitions.

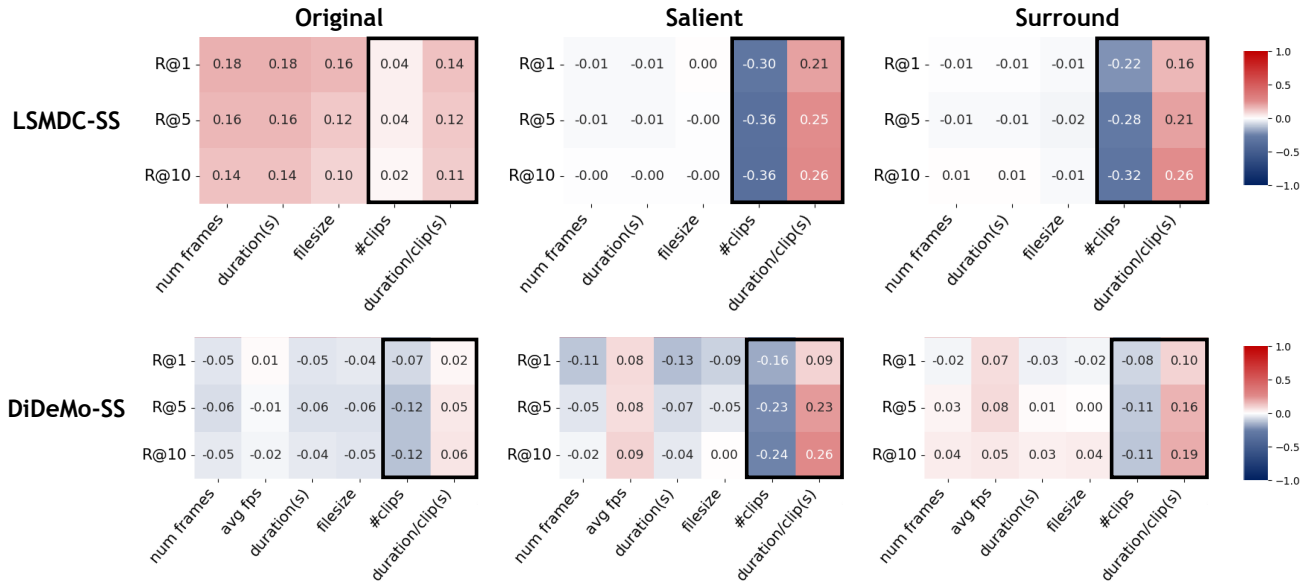

Figure S3. **Correlation matrix of LSMDC-SS and DiDeMo-SS.** This figure presents a correlation analysis using the pretrained UMT-L model. Similar to the MSRVTT-SS results, the number of clips shows a negative correlation with Recall, while the duration per clip shows a positive correlation.

Table S3. **Comparison of Recall (R) and Recall per Video (R/V).** This is the extension version of the Table 6 of the main paper. Based on the same retrieval results, the value of R/V is larger than the original R regardless of the model and datasets.

Model	Query	Metric	MSRVTT-SS			LSMDC-SS			DiDeMo-SS		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UMT-L	Salient	R	26.09	42.77	50.02	27.99	48.74	56.46	35.45	56.48	64.49
		R/V	37.16	54.66	61.34	32.34	54.03	61.52	48.08	72.20	79.35
	Surround	R	12.47	23.69	29.71	13.19	27.60	35.41	16.58	31.96	39.15
		R/V	18.48	32.12	38.77	15.64	31.59	40.24	21.05	39.19	46.88
InternVideo2	Salient	R	34.23	52.97	60.78	32.20	54.34	61.84	41.80	63.56	71.83
		R/V	44.55	62.78	69.45	36.35	59.13	66.55	54.34	77.47	84.42
	Surround	R	17.31	32.46	39.91	16.00	33.90	41.83	20.24	38.14	46.94
		R/V	23.28	39.78	47.11	18.43	37.87	46.24	25.95	46.47	55.13

Table S4. **Zero-shot performance comparison of video database and clip database.** This is the extended version of Table 7 from the main paper. We color the result gray when switching from the video database to the clip database if the Recall increases by less than 5 or decreases. If the Recall increases by more than 5, we highlight the result in blue.

Model	Query	Database	MSRVTT-SS			LSMDC-SS			DiDeMo-SS		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UMT-L	Original	Video	40.14	63.76	71.87	25.70	42.50	51.70	22.18	42.30	50.29
		Clip	39.17	59.39	66.27	26.81	47.01	53.85	25.29	39.42	48.18
		Δ	(−0.97)	(−4.37)	(−5.60)	(+1.11)	(+4.51)	(+2.15)	(+3.11)	(−2.88)	(−2.11)
	Salient	Video	26.09	42.77	50.02	27.99	48.74	56.46	35.45	56.48	64.49
		Clip	39.18	58.68	66.38	35.68	57.32	65.34	47.36	67.47	74.77
		Δ	(+13.09)	(+15.91)	(+16.36)	(+7.69)	(+8.58)	(+8.88)	(+11.91)	(+10.99)	(+10.28)
	Surround	Video	12.47	23.69	29.71	13.19	27.60	35.41	16.58	31.96	39.15
		Clip	18.23	33.57	40.69	15.62	32.25	40.62	24.04	38.86	45.76
		Δ	(+5.76)	(+9.88)	(+10.98)	(+2.43)	(+4.65)	(+5.21)	(+7.46)	(+6.90)	(+6.61)
InternVideo2	Original	Video	52.25	74.47	82.18	31.80	52.40	59.10	20.46	41.25	51.26
		Clip	47.81	68.40	77.16	32.80	53.96	60.13	25.07	44.29	51.71
		Δ	(−4.44)	(−6.07)	(−5.02)	(+1.00)	(+1.56)	(+1.03)	(+4.61)	(+3.04)	(+0.45)
	Salient	Video	34.23	52.97	60.78	32.20	54.34	61.84	41.80	63.56	71.83
		Clip	47.44	67.38	74.46	40.55	62.91	71.00	55.10	74.02	80.38
		Δ	(+13.21)	(+14.41)	(+13.68)	(+8.35)	(+8.57)	(+9.16)	(+13.3)	(+10.46)	(+8.55)
	Surround	Video	17.31	32.46	39.91	16.00	33.90	41.83	20.24	38.14	46.94
		Clip	23.58	40.11	47.75	19.89	39.14	47.67	28.70	45.00	52.31
		Δ	(+6.27)	(+7.65)	(+7.84)	(+3.89)	(+5.24)	(+5.84)	(+8.46)	(+6.86)	(+5.37)

Table S5. **Video-caption finetuning results.** We finetune UMT-L on the train set of LSMDC-SS and DiDeMo-SS and evaluate the text-to-video retrieval performance on the test set of each dataset with a video database.

Query	Model	LSMDC-SS			DiDeMo-SS		
		R@1	R@5	R@10	R@1	R@5	R@10
Original	Zero-shot	25.70	42.50	51.70	22.18	42.30	50.29
	Video ft.	30.90	50.20	59.70	23.43	45.53	56.56
		(+5.20)	(+7.70)	(+8.00)	(+1.25)	(+3.23)	(+6.27)
Salient	Zero-shot	27.99	48.74	56.46	35.45	56.48	64.49
	Video ft.	40.51	64.99	72.73	36.64	61.06	70.34
		(+12.52)	(+16.25)	(+16.27)	(+1.19)	(+4.58)	(+5.85)
Surround	Zero-shot	13.19	27.60	35.41	16.58	31.96	39.15
	Video ft.	28.35	54.68	64.47	18.60	37.32	47.17
		(+15.16)	(+27.08)	(+29.06)	(+2.02)	(+5.36)	(+8.02)

Table S6. **Clip-caption finetuning results.** We finetune UMT-L on the train set of LSMDC-SS and DiDeMo-SS and evaluate the text-to-clip retrieval performance on the test set of each dataset with a clip database.

Query	Model	LSMDC-SS			DiDeMo-SS		
		R@1	R@5	R@10	R@1	R@5	R@10
Original	Zero-shot	26.81	47.01	53.85	25.29	39.42	48.18
	Video ft.	**31.32**	**50.54**	**58.76**	25.18	42.19	49.85
	Clip ft.	27.10	47.06	54.65	**25.94**	**44.63**	**54.45**
Salient	Zero-shot	35.68	57.32	65.34	**47.36**	67.47	74.77
	Video ft.	**46.10**	**69.57**	**77.17**	46.57	69.08	76.68
	Clip ft.	45.85	69.50	76.98	45.53	**69.20**	**77.00**
Surround	Zero-shot	15.62	32.25	40.62	24.04	38.86	45.76
	Video ft.	32.17	58.92	68.58	25.03	42.97	51.29
	Clip ft.	**33.42**	**60.41**	**69.97**	**26.97**	**45.92**	**54.51**

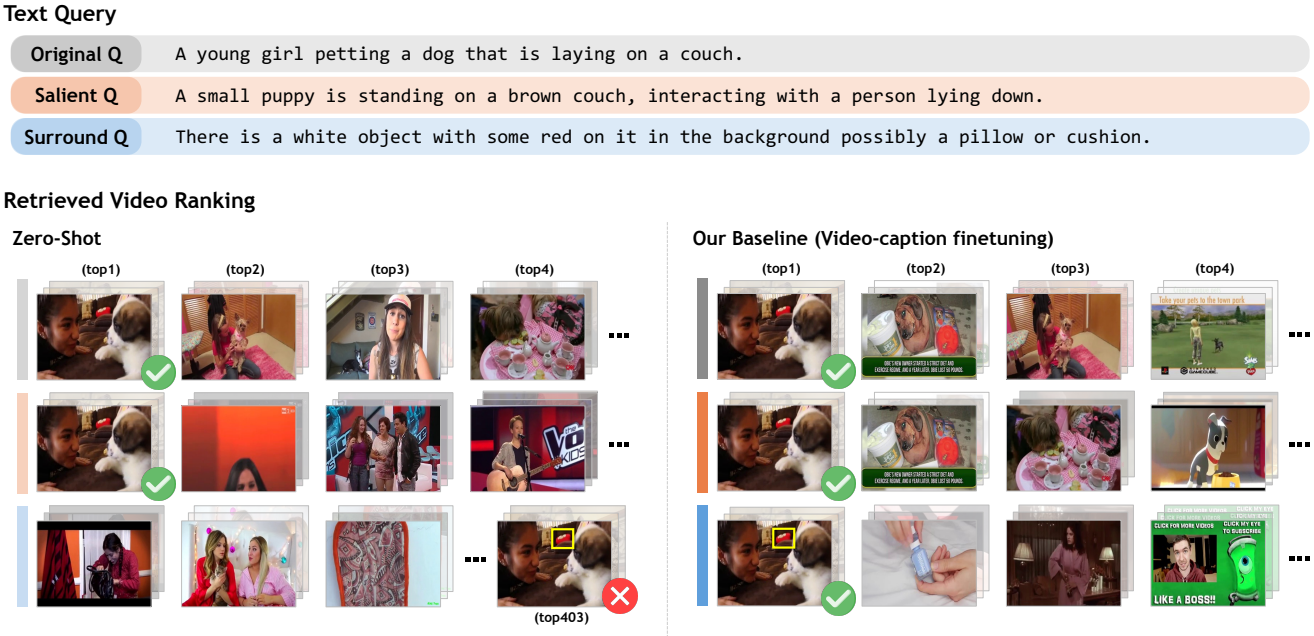

Figure S4. **Qualitative comparison between zero-shot and our baseline.** We compare the retrieval results across original, salient, and surrounding context queries. The left side shows the retrieval results from the existing model, while the right side shows the results from our baseline. For each query, we rank the retrieved videos and show whether the ground-truth video is successfully retrieved.