Supplementary Material for CoreCaption: Core Caption based Text-to-Video Retrieval

This supplementary material provides additional experimental results and analysis that could not be included in the main paper due to space limitations. In the main paper, we primarily report the performance of the Student model, as it represents the practical deployment scenario of our framework. While the Teacher model achieves superior performance, it requires extracting and maintaining a database of core captions for all videos during deployment, which introduces additional computational and storage overhead. However, the analysis of the Teacher model is crucial because it directly validates our proposed methods, whereas the results of the Student model demonstrate the effectiveness of knowledge distillation.

In order to provide a comprehensive validation of our approach, we perform the same set of experiments on the Teacher model that were performed on the Student model in the main paper. This analysis is particularly important because the performance of the Student model, while practical, is derived through knowledge distillation from the Teacher model.

Our analysis focuses on four key aspects:

1. Comprehensive evaluation of the Teacher model's performance across MSR-VTT, VATEX, and MSVD datasets
2. Detailed ablation studies on the Guide Attention (GA) and CoreCaption Extraction (CCE) components of the Teacher model
3. Investigation of the Teacher model's performance under different caption availability scenarios
4. An ablation study on the reason for using only Distillation Loss instead of InfoNCE when training the Student Model of our model
5. Computational cost analysis comparing inference efficiency with existing state-of-the-art methods

Through these analyses, we demonstrate both the full capabilities of our Teacher model and validate the fundamentals of our proposed methods, thereby providing stronger evidence for the effectiveness of the CoreCaption framework.

A. Teacher Model Performance Analysis

In this section, we present a detailed performance comparison between our Teacher and Student models across three benchmark datasets: MSR-VTT, VATEX, and MSVD. While the main paper focused on the results of the Student model for practical deployment scenarios, here we demonstrate the remarkable capabilities of our Teacher model.

Dataset	Model	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MSR-VTT	Teacher	67.2	92.1	96.4	1.0	2.7
	Student	53.2	79.5	87.1	1.0	9.7
VATEX	Teacher	79.7	98.2	99.7	1.0	1.5
	Student	67.3	94.2	97.9	1.0	2.7
MSVD	Teacher	58.4	87.4	94.2	1.0	4.1
	Student	53.4	83.0	91.8	1.0	7.7

Table 1. Performance comparison between Teacher and Student models on multiple datasets. ↑ indicates higher is better, ↓ indicates lower is better.

As shown in Table 1, our Teacher model achieves exceptional performance across all datasets. This outstanding performance can be attributed to its ability to directly utilize core captions during the retrieval process, effectively leveraging the most representative textual descriptions for each video. Specifically, on the VATEX dataset, the Teacher model achieves an impressive R@1 of 79.2% , demonstrating the effectiveness of our core caption-based approach. Similar remarkable results are observed on MSR-VTT with R@1 of 67.0% and MSVD with R@1 of 58.2%. The direct incorporation of core captions into the retrieval process allows the Teacher model to capture and utilize the essential semantic content of videos more effectively.

The Student model, while being more efficient to deploy, successfully preserves much of this exceptional performance through knowledge distillation from the Teacher model. This preservation of the Teacher's capabilities enables the Student model to significantly outperform existing benchmarks, as demonstrated in the main paper, despite not having direct access to core captions during inference.

B. Ablation Studies on Teacher Model Components

Following the same methodology used in the main paper's ablation studies, we conduct extensive experiments to val-

idate the effectiveness of each component in our Teacher model on the MSR-VTT dataset. Specifically, we analyze the impact of different Guide Attention (GA) mechanisms and CoreCaption Extraction (CCE) methods.

As shown in Table 4, our proposed CCGA demonstrates remarkable effectiveness compared to other attention mechanisms. The significant performance gap between CCGA and conventional approaches such as Mean-Pool with R@1 improvement of 16.9% and Cross-Attention with R@1 improvement of 10.8% highlights the unique advantages of our method. In particular, CCGA's ability to preserve the original query intent while effectively incorporating core caption information proves to be crucial, as evidenced by the consistent performance improvements across all metrics.

Regarding caption extraction methods, while both Top-K and Random selection achieve reasonable performance when combined with our CCGA, our density-based approach further improves the results, achieving R@1 of 67.0%. This demonstrates that the combination of our sophisticated attention mechanism (CCGA) and effective caption extraction method creates a strong synergy, leading to optimal performance in text-video retrieval. These results are consistent with and further validate the results presented in the main paper's ablation studies, showing that the improvements from our proposed components are consistently significant in both the Teacher and Student models.

GA	CCE	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MP	Ours	50.1	81.5	90.7	1.0	6.0
CA	Ours	56.2	84.2	91.7	1.0	4.6
Ours	Top-K	65.2	90.8	95.4	1.0	3.1
Ours	Rand	64.7	90.2	95.3	1.0	3.4
Ours	Ours	**67.2**	**92.1**	**96.4**	**1.0**	**2.7**

Table 2. Ablation study of different Guide Attention (GA) and CoreCaption Extraction (CCE) methods on MSR-VTT dataset. MP: Mean-Pool, CA: Cross-ATT, Rand: Random selection.

C. Caption Availability Analysis

As detailed in the main paper, we investigate scenarios where the number of available captions per video is limited during training. While our framework is designed to take advantage of multiple captions per video, it is crucial to validate its effectiveness in more constrained settings that often occur in real-world applications.

Following the approaches described in the main paper, we evaluate our Teacher model using both the Global Caption Pool and Captioner Augmentation methods. As shown in Table 3, our Teacher model maintains impressive performance even in caption-constrained scenarios. Particularly noteworthy is that when we augment the limited caption set with a pre-trained captioner to create multiple captions, the model achieves remarkably strong performance with R@1 of 61.9%. This demonstrates that the core mechanisms of our framework remain effective even with synthetic captions, and that the strategy of generating multiple captions through augmentation can successfully compensate for limited caption availability. These results suggest that combining our core caption-based approach with caption augmentation techniques could be a promising solution for real-world applications where obtaining multiple human annotations per video is challenging.

Method	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Global Caption Pool	53.9	83.2	91.5	1.0	5.9
Captioner Augmentation	**61.9**	**89.6**	**95.3**	**1.0**	**3.3**

Table 3. Performance of our Teacher model when training with single caption per video on MSR-VTT dataset.

D. Ablation Study on Distillation-only Loss for Student Model Training

In our main paper, we adopt the strategy of training the student model exclusively with the distillation loss, L_d, to effectively transfer knowledge from the teacher model. Although it is common in other contexts to combine the distillation loss with the InfoNCE loss, L_{ce}, in our CoreCaption framework the teacher model's output is overwhelmingly strong. Consequently, pure knowledge transfer via L_d yields more stable and effective learning for the student model.

To verify this observation, we conducted an ablation study in which the total loss is defined as

$$L_{\text{total}} = L_d + \lambda L_{ce},$$

and we varied the weight λ over the set $\{0, 0.1, 0.2, 0.5, 1, 5, 10\}$. All experiments were performed on the MSR-VTT dataset using standard splits, and the student model was trained under identical architectural and hyperparameter settings (learning rate, batch size, number of epochs, etc.), differing only in the loss function formulation. We evaluated the models using the metrics Recall@1, Recall@5, Recall@10, Median Rank (MdR), and Mean Rank (MnR).

Table 4 summarizes the performance of the student model for different λ values.

As shown in Table 4, increasing the value of λ leads to a gradual degradation in performance. The best performance is achieved when $\lambda = 0$, i.e., when only the distillation loss is used. This result indicates that, contrary to typical scenarios, the strong performance of the teacher model in our framework makes pure knowledge transfer via L_d more effective and stable. In our case, the addition of the InfoNCE

λ	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
0 (Ours)	**53.2**	**79.5**	**87.1**	**1.0**	**9.7**
0.1	53.0	79.1	87.0	1.0	9.8
0.2	52.7	78.6	86.7	1.0	10.1
0.5	52.4	78.4	86.3	1.0	10.9
1	52.2	78.2	86.2	1.0	11.0
5	51.1	77.1	85.2	1.0	11.5
10	49.4	75.4	83.9	1.0	12.0

Table 4. Ablation study results for varying the weight λ in the total loss $L_{\text{total}} = L_d + \lambda L_{ce}$.

loss introduces unnecessary information that interferes with the optimization process, leading to a decrease in performance.

Based on these findings, our final CoreCaption framework adopts the strategy of training the student model solely with the distillation loss. These experimental results provide the quantitative support for the statement of the main paper, demonstrating that, for our approach, pure distillation results in more stable and consistent performance.

E. About computational cost

To address potential concerns about computational efficiency, our teacher-student architecture is specifically designed to maintain practical inference costs. During inference, the student model operates without core caption extraction, achieving computational efficiency comparable to existing single-model approaches while preserving the performance benefits gained from teacher-student distillation.

Method	R@1(%)	Params(M)	Mem(G)	Infer
UCOFIA	49.4	166	3.47	4.11
DiffusionRet	49.0	-	3.25	62.04
Cap4Video++	50.3	104	2.84	4.39
Ours	**53.2**	128	**2.32**	**3.25**

Table 5. Cost Comparison; Infer:Inference time(s/gallery size)

Our method shows the lowest memory usage (2.32G) and fastest inference time (3.25s) among compared methods. While core caption extraction introduces additional complexity during training, this overhead is eliminated during inference through knowledge distillation.

F. Experiments Setting

Dataset and Evaluation Metric. We utilize four major benchmark datasets to evaluate the performance of video-to-text retrieval and text-to-video retrieval tasks. First, MSR-VTT consists of a total of 10k video clips, with each clip assigned 20 captions. We train our model using relevant captions from the Training-9K set and report results on the test 1K-A set. Second, VATEX has collected approximately 35k videos, each containing multiple annotations. This dataset consists of about 26k videos for training, 1,500 for validation, and 1,500 for testing. Lastly, MSVD includes 1,970 videos and 80k captions, with an average of about 40 captions assigned per video. The training, validation, and test sets of this dataset consist of 1,200, 100, and 670 videos, respectively. For evaluating retrieval performance, we select recall at rank K (R@K) where K is 1,5,10, median rank (MdR), and mean rank (MnR).

Implementation Details. In all our experiments, we use CLIP's VIT-32 encoder for video frames and CLIP's transformer encoder for text queries. The maximum text length is set to 32, and for video frames, we sample 12 frames from video clips across all datasets, resizing all frames to 224×224. We set the batch size to 32 and use an initial learning rate of 1e-5 for non-CLIP parameters and 1e-6 for CLIP parameters, following a cosine schedule over 5 epochs of training. Experiments were conducted on an RTX4090 GPU. Consistently across all datasets, we used a single transformer layer for the joint encoding function ψ, adaptively determined the number of core captions per video clip with a maximum limit of 8 captions, and maintained a model dropout of 0.2 with a dimension of 512.

G. Captioner Augmentation for Single Caption Scenarios

Pre-trained (zero-shot) Video Captioning Augmentation To address single caption scenarios, we employ ZeroCap [1], a zero-shot video captioning method that generates captions without requiring video-caption training pairs. ZeroCap utilizes CLIP to assess visual-text similarity and GPT-2 for language generation, optimizing pseudo-tokens during auto-regression to guide the language model toward generating coherent video descriptions. The method starts with an initial prompt ("Video shows") and iteratively generates tokens while using CLIP loss to maintain alignment with video content and a language modeling loss to preserve linguistic coherence.

For each video in single-caption datasets, we generate 20 additional captions using ZeroCap to create a sufficient caption pool for our core caption extraction process. This augmentation strategy enables our framework to operate effectively even when original datasets provide only one caption per video, as demonstrated in Tables 6-7.

MLLM-based Caption Augmentation We explore MLLM-based augmentation using LLaVA-OneVision and InstructBLIP-Video [7]. Both models require 4 input frames per inference for video understanding. We divide each video into 16 equal segments and create 4 different frame combinations: (0,4,8,12), (1,5,9,13), (2,6,10,14), and (3,7,11,15), where numbers indicate segment indices. For each frame combination, we generate 5 captions using different temperature settings (0.2, 0.4, 0.6, 0.8, 1.0), resulting in 20 diverse captions per video. Both models are excellent MLLM approaches capable of generating captions using

video content. These MLLM-generated captions provide richer semantic information compared to traditional captioning methods, as demonstrated by superior performance in Tables 6-7.

References

[1] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17918–17928, 2022. 3