ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464179

TEXT SEMANTICS-GUIDED DUAL-TEACHER KNOWLEDGE DISTILLATION FOR PARTIALLY RELEVANT VIDEO RETRIEVAL

Runhao Zeng[1], *Yirui Wang*[1,2], *Yishen Zhuo*[3], *Feng Liang*[1], *Haihan Duan*[1], *Hui Li*[3,*]

[1]Shenzhen MSU-BIT University [2]Beijing Institute of Technology [3]Shenzhen University

ABSTRACT

Partially Relevant Video Retrieval (PRVR) aims to retrieve untrimmed videos that contain segments partially relevant to a query, but the task is challenged by large amounts of irrelevant content that hinder precise cross-modal alignment. Existing single-teacher knowledge distillation approaches provide incomplete supervision and are vulnerable to noise. To address this, we propose a Text Semantics-Guided Dual-Teacher Knowledge Distillation framework, which integrates complementary knowledge from a multi-modal large language model and CLIP to enhance cross-modal alignment. We further design a cognition-driven distillation strategy that allows the student to adaptively adjust teacher guidance, and a text semantics-guided temporal enhancement module that aggregates multi-scale Gaussian mixture features for stronger temporal representation. Extensive experiments on ActivityNet Captions and Charades-STA demonstrate that our method achieves significant improvements over state-of-the-art approaches, validating the effectiveness of dual-teacher distillation for PRVR.

***Index Terms*—** Partially Relevant Video Retrieval, Dual-Teacher Knowledge Distillation

1. INTRODUCTION

Partially relevant video retrieval (PRVR) identifies untrimmed videos from a large corpus partially relevant to a textual query. It is crucial for applications like intelligent multimedia search and automated sports highlight generation [1]. This task, however, is challenging due to complex spatiotemporal information and severe background clutter, which complicate robust cross-modal alignment. Traditional methods for pre-trimmed clips [2], assuming full semantic match, perform poorly in this realistic scenario. The core bottleneck is effectively extracting and aligning relevant video features with query semantics amidst overwhelming noise.

Existing long-video retrieval has two categories: Non-distillation methods [3, 4, 5] often struggle without large-scale supervision. Distillation-based methods [6] utilize large pre-trained models, but are constrained by two key limitations: First, they typically employ a single teacher model, failing to fully leverage the complementarity of multi-source knowledge. Second, traditional knowledge distillation strategies often overlook the model's autonomous cognitive evolution, making dynamic knowledge absorption difficult.

To overcome these, we propose a **Text Semantics-Guided Dual-Teacher Knowledge Distillation** algorithm for video retrieval. Our approach introduces a dual-teacher framework for comprehensive knowledge transfer. A **cognition-driven distillation strategy** enables adaptive knowledge absorption, while a **text semantics-guided temporal enhancement module** explicitly models temporal relationships. This integrated method significantly improves cross-modal representations for accurate untrimmed video retrieval.

Our main contributions are: (1) A text semantics–guided dual-teacher distillation framework that enables comprehensive and robust knowledge transfer for video retrieval. (2) A cognition-driven strategy that adaptively modulates teacher guidance based on the student's learning state. (3) A text semantics–guided temporal module that leverages query cues for multi-scale feature aggregation, yielding superior results over state-of-the-art baselines.

2. RELATED WORK

Knowledge Distillation (KD) for Video Retrieval. KD transfers knowledge from large teachers to smaller students via soft labels [7, 8], evolving from response matching to feature and relation alignment [9, 10]. It is crucial for adapting large-scale models such as CLIP [11] to specialized tasks [12]. In video retrieval, KD alleviates annotation scarcity [6], yet single-teacher methods remain limited and noise-sensitive. We propose a dual-teacher framework that integrates complementary guidance for more robust transfer.

Video-Text Representation Learning. Effective video-text alignment relies on robust unimodal encoders. Visual modeling has advanced from CNNs (AlexNet [13], VGG [14], ResNet [15]) to Transformers (ViT [16]) and spatiotemporal 3D CNNs (I3D [17]). Text representation evolved from Word2Vec [18] to BERT and Transformer-based models [19, 20]. Cross-modal encoders such as CLIP [11] provide vision-grounded semantics, while LLMs (e.g., GPT [21], LLaMA [22]) contribute advanced language understanding. We leverage these encoders to build robust video-text representations.

*Corresponding author.

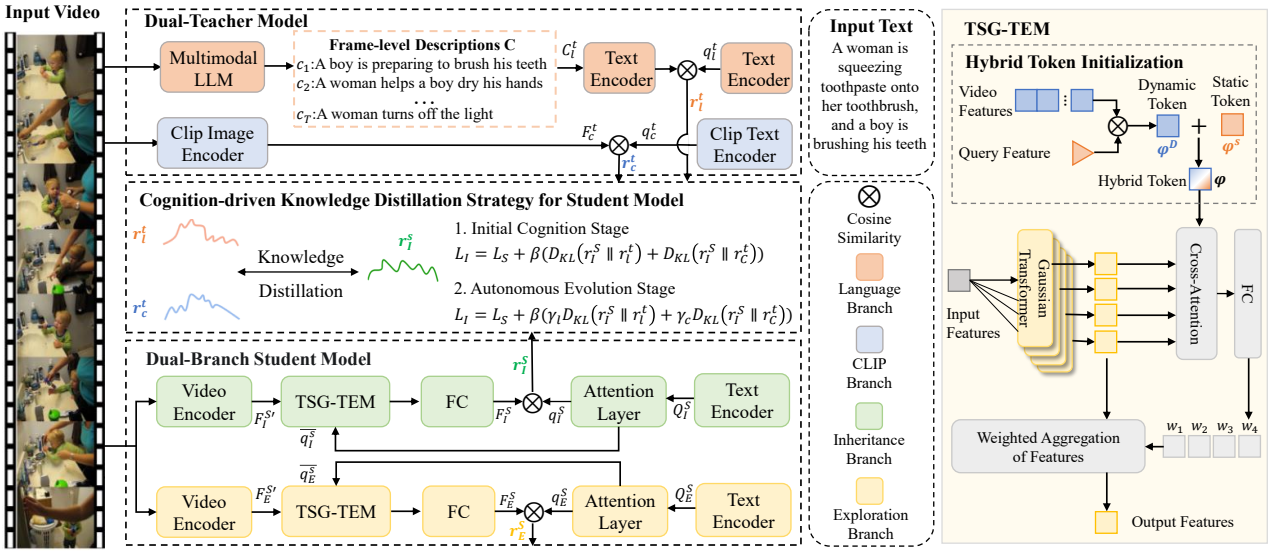

Fig. 1: An overview of the proposed framework. A **Dual-Teacher Model** produces supervisory distributions (r_l^t, r_c^t) to guide a **Dual-Branch Student** (Inheritance and Exploration) through a **Cognition-driven Distillation** Strategy, while the **Text Semantics-Guided Temporal Enhancement Module (TSG-TEM)** refines temporal features with query semantics.

3. PROPOSED METHOD

Partially Relevant Video Retrieval (PRVR) targets retrieving untrimmed videos that contain segments partially matching a natural language query Q. Given a query Q and a large corpus $\mathcal{U} = V_{i}{}_{i=1}^{B}$, the goal is to return a video $V \in \mathcal{U}$ that is semantically aligned with Q at least at the segment level.

3.1. General Scheme

The reliance on a single teacher model in prior distillation-based methods restricts comprehensive knowledge leverage. To overcome this, we propose a Text Semantics–Guided Dual-Teacher Knowledge Distillation algorithm comprising three components: a dual-teacher framework (Section 3.2), a cognition-driven strategy (Section 3.3), and a semantics-guided temporal module (Section 3.4). As shown in Figure 1, the dual teachers provide supervisory similarity distributions (r_l^t, r_c^t) to guide the student's inheritance (r_I^s) and exploration (r_E^s) branches. The cognition-driven strategy adaptively balances teacher guidance, while the temporal module refines video features with query semantics. These designs enable robust knowledge transfer and improved retrieval performance.

3.2. Text Semantics-Guided Knowledge Distillation via Dual Teachers

To address single-teacher KD limitations, we introduce a dual-teacher framework. The first teacher, a Multimodal LLM (CogVLM2 [23]), generates frame descriptions $\mathcal{C}_l^t$. With query embeddings $\mathbf{q}_l^t$ (from Sentence-BERT [24]), this forms a language-language matching for similarity r_l^t:

$$r_l^t = \frac{\mathbf{q}_l^t \cdot \mathcal{C}_l^t}{\|\mathbf{q}_l^t\| \|\mathcal{C}_l^t\|}. \tag{1}$$

Here, r_l^t is the language teacher's similarity distribution, $\mathbf{q}_l^t$ is the language query embedding, and $\mathcal{C}_l^t$ represents the frame description embeddings. The second teacher, a pre-trained CLIP model [11], provides cross-modal priors. Its similarity r_c^t is computed via cosine similarity between video embedding $\mathbf{F}_c^t$ and query embedding $\mathbf{q}_c^t$:

$$r_c^t = \frac{\mathbf{q}_c^t \cdot \mathbf{F}_c^t}{\|\mathbf{q}_c^t\| \|\mathbf{F}_c^t\|}. \tag{2}$$

Here, r_c^t is the CLIP teacher's similarity distribution, $\mathbf{q}_c^t$ is the CLIP query embedding, and $\mathbf{F}_c^t$ is the CLIP video embedding.

Dual-Branch Student Model. Inspired by [6], the student consists of two branches. The *Inheritance Branch* learns generalizable knowledge from the teachers. Given video V and query Q, it encodes them via a CNN and RoBERTa [25], followed by our temporal enhancement module, producing representations $\mathbf{F}_I^s$ and $\mathbf{q}_I^s$ with similarity distribution:

$$r_I^s = \frac{\mathbf{q}_I^s \cdot \mathbf{F}_I^s}{\|\mathbf{q}_I^s\| \|\mathbf{F}_I^s\|}. \tag{3}$$

The *Exploration Branch* shares the same architecture but is trained without KD, enabling it to capture domain-specific cues and mitigate distribution shifts, and lastly outputs r_E^s.

3.3. Cognition-Driven Distillation Strategy for Student

To transcend predefined learning strategies, we propose a cognition-driven knowledge distillation (SC-KD) strategy.

This approach enables the student model to adaptively absorb knowledge based on its evolving state, operating in two phases demarcated by a threshold ρ. In the early stage ($t < \rho$), the inheritance branch is optimized by combining contrastive loss $\mathcal{L}_s$ with distillation losses from both LLM and CLIP teachers. This combined inheritance loss, $\mathcal{L}_I$, is defined as:

$$\mathcal{L}_I = \mathcal{L}_s + \beta(t)(\mathcal{L}_c + \mathcal{L}_l), \tag{4}$$

where $\mathcal{L}_l = D_{KL}(r_I^s \| r_l^t)$ and $\mathcal{L}_c = D_{KL}(r_I^s \| r_c^t)$ represent the Kullback-Leibler (KL) divergence distillation losses. $\beta(t) = \beta_0 k^t$ is an exponentially decaying weight.

In the autonomous evolution phase ($t \geq \rho$), the student leverages feedback from its exploration branch. The consistency between exploration outputs and each teacher is measured to derive adaptive weights γ_l and γ_c, which dynamically adjust the influence of the two teachers.

$$\gamma_l = 1 - \text{Sigmoid}(\mathcal{L}_l'), \quad \gamma_c = 1 - \text{Sigmoid}(\mathcal{L}_c'), \tag{5}$$

where $\mathcal{L}_l' = D_{KL}(r_E^s \| r_l^t)$ and $\mathcal{L}_c' = D_{KL}(r_E^s \| r_c^t)$. The final loss function for the inheritance branch becomes dynamically weighted, allowing the student to prioritize the more reliable teacher:

$$\mathcal{L}_I = \begin{cases} \mathcal{L}_s + \beta(t)(\mathcal{L}_l + \mathcal{L}_c), & \text{if } t < \rho \\ \mathcal{L}_s + \beta(t)(\gamma_l \mathcal{L}_l + \gamma_c \mathcal{L}_c), & \text{if } t \geq \rho. \end{cases} \tag{6}$$

This cognition-driven approach makes the knowledge transfer process more efficient and robust.

3.4. Semantics-Guided Temporal Enhancement Module

To enhance temporal representation, we propose the Text Semantics-guided Temporal Enhancement Module (TSG-TEM), which incorporates query semantics into temporal dependency modeling.

As depicted in the TSG-TEM architecture (Figure 1), for each branch, inheritance (I) and exploration (E), we initialize a hybrid token ϕ consisting of a shared static token ϕ^S and a branch-specific dynamic token ϕ^D derived from the visual feature most similar to the batch-averaged query embedding $\bar{\mathbf{q}}_s$:

$$\phi_I = \phi^S + \phi_I^D, \quad \phi_E = \phi^S + \phi_E^D. \tag{7}$$

The hybrid token ϕ is used in a cross-attention mechanism to generate adaptive weights for aggregating multi-scale features from Gaussian Transformers [4]. The output features $\mathbf{F}_I^s$ and $\mathbf{F}_E^s$ is obtained by concatenating the aggregated features and is applied in the final similarity computation.

3.5. Training and Inference Details

Training. The student model is optimized by minimizing the total loss $\mathcal{L}$, which is the sum of the Inheritance branch loss $\mathcal{L}_I$ and the Exploration branch loss $\mathcal{L}_E$:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_E. \tag{8}$$

Table 1: Performance comparison with SOTA methods on the Anet-Captions dataset using I3D features.

Method	R@1	R@5	R@10	R@100	SumR
Natural language-based short-video retrieval methods:					
HTM	3.7	13.7	22.3	66.2	105.9
HGR	4.0	15.0	24.8	63.2	107.0
RIVRL	5.2	18.0	28.2	66.4	117.8
VSE++	4.9	17.7	28.2	67.1	117.9
DE++	5.3	18.4	29.2	68.0	121.0
W2VV++	5.4	18.7	29.7	68.8	122.6
CE	5.5	19.1	29.9	71.1	125.6
CLIP4Clip	5.9	19.3	30.4	71.6	127.3
Cap4Video	6.3	20.4	30.9	72.6	130.2
Natural language-based long-video retrieval methods:					
MS-SL	7.1	22.5	34.7	75.8	140.1
T-D3N	7.3	23.8	36.0	76.6	143.6
GMMFormer	8.3	24.9	36.7	76.1	146.0
DL-DKD	8.0	25.0	37.5	77.1	147.6
Ours	8.9	26.6	39.2	78.9	153.6

Table 2: Performance comparison with SOTA methods on the Charades-STA dataset using I3D features.

Method	R@1	R@5	R@10	R@100	SumR
Natural language-based short-video retrieval methods:					
VSE++	0.8	3.9	7.2	31.7	43.6
W2VV++	0.9	3.5	6.6	34.3	45.3
HGR	1.2	3.8	7.3	33.4	45.7
CE	1.3	4.5	7.3	36.0	49.1
DE++	1.7	5.6	9.6	37.1	54.1
RIVRL	1.6	5.6	9.4	37.7	54.3
HTM	1.2	5.4	9.2	44.2	60.0
CLIP4Clip	1.8	6.5	10.9	44.2	63.4
Cap4Video	1.9	6.7	11.3	45.0	65.0
Natural language-based long-video retrieval methods:					
MS-SL	1.8	7.1	11.8	47.7	68.4
T-D3N	2.1	7.6	13.1	48.8	71.6
GMMFormer	2.1	7.8	12.5	50.6	72.9
DL-DKD	1.6	6.7	10.9	47.6	66.8
Ours	2.3	7.2	12.6	50.0	72.1

Here, $\mathcal{L}_E$ comprises triplet ranking [26] and InfoNCE [27] losses, while $\mathcal{L}_I$ is detailed in Section X.

Inference. Only the student model performs retrieval. For a video V and query Q, each branch calculates its matching score ($M_I(Q,V)$ for Inheritance, $M_E(Q,V)$ for Exploration) as the maximum cross-modal similarity across video frames. The final matching score $M(Q,V)$ is their weighted sum:

$$M(Q,V) = (1-\mu)M_I(Q,V) + \mu M_E(Q,V), \tag{9}$$

where μ is a hyperparameter balancing branch contributions.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets. We evaluate our method on two standard benchmarks: **ActivityNet Captions (ANet-Captions)** [28], a chal-

Table 3: Ablation study of our proposed modules.

TSK	SC-KD	TEM	TSG-TEM	R@1	R@5	R@10	R@100	SumR
-	-	-	-	1.6	6.7	10.9	47.6	66.8
✓	-	-	-	1.6	6.5	11.1	49.0	68.2
✓	✓	-	-	1.7	7.2	11.6	48.9	69.3
✓	✓	✓	-	1.8	7.1	12.7	49.1	70.8
✓	✓	✓	✓	**2.3**	**7.2**	**12.6**	**50.0**	**72.1**

Table 4: Effect of different student branches for guiding KD.

Setup	R@1	R@5	R@10	R@100	SumR
Baseline	1.6	6.7	10.9	47.6	66.8
Inheritance Branch	1.9	6.9	12.3	49.0	70.1
Exploration Branch (Ours)	**2.3**	**7.2**	**12.6**	**50.0**	**72.1**

lenging dataset of 20k untrimmed videos (avg. 118s), and **Charades-STA** [28], with 6,670 activity videos (avg. 29.8s).

Evaluation Metrics. We follow [3] to use Recall at Rank n (R@n, for n=1, 5, 10, 100) and the sum of recalls (SumR).

Implementation Details. The dual-teacher model comprises a language teacher (CogVLM2 [23] + Sentence-BETR [24]) and a CLIP teacher (ViT-B/32 [11]). The student model uses I3D [17] and RoBERTa [25] encoders, both yielding 1024-D features. We train on NVIDIA RTX 4090 GPUs with a batch size of 128 and an initial learning rate of 0.00025, using the scheduling from [6] and early stopping based on validation SumR. During inference, scores are fused with $\mu = 0.3$.

4.2. Comparison with State-of-the-Art Baselines

To validate our method's effectiveness, we compare it against a set of state-of-the-art (SOTA) baselines. Specifically, these include eleven short-video retrieval methods (VSE++ [29], W2VV++ [30], HGR [2], CE [31], DE++ [26], RIVRL [32], HTM [33], CLIP4Clip [34], Cap4Video [35]) and four long-video retrieval methods (MS-SL [3], DL-DKD [6], GMMFormer [4], T-D3N [36]). For fair comparison, all methods utilize identical pre-extracted I3D visual features.

Results on ANet-Captions. As Table 1 shows, our method significantly surpasses all baselines. Short-video methods underperform in this long-form context. Even against strong long-video baselines like DL-DKD [6], our approach yields a 6.0% SumR and 0.9% R@1 gain. This validates our dual-teacher framework's superiority in temporal and cross-modal correlation modeling.

Results on Charades-STA. Table 2 confirms our method's effectiveness on Charades-STA, achieving the highest R@1. We improve SumR by 5.3% over the strong DL-DKD baseline. These consistent inter-dataset improvements demonstrate our approach mitigates single-teacher distillation limitations via complementary knowledge, leading to accurate cross-modal alignment for long videos.

Table 5: Effect of sharing the static token.

Setup	R@1	R@5	R@10	R@100	SumR
Baseline	1.6	6.7	10.9	47.6	66.8
Non-shared Static Token	2.1	7.6	12.2	47.8	69.7
Shared Static Token (Ours)	**2.3**	**7.2**	**12.6**	**50.0**	**72.1**

4.3. Ablation Studies

Ablation studies on the Charades-STA dataset, using DL-DKD [6] as baseline, analyze each component's contribution.

Effect of Model Components. To validate the effectiveness of each proposed module, we incrementally add them to the baseline. The modules are: Textual Semantic Knowledge (TSK), Student Model Cognition-driven Knowledge Distillation (SC-KD), the standard Temporal Enhancement Module (TEM), and our Text Semantics-guided Temporal Enhancement Module (TSG-TEM). As shown in Table 3, each module provides a progressive performance gain. The full model achieves a 5.3% improvement in SumR over the baseline. Specifically, introducing TSK brings a 1.4% boost by leveraging complementary knowledge from a dual-teacher setup. The SC-KD strategy further improves performance by 1.1% through adaptive knowledge fusion. TEM enhances temporal modeling, yielding a 1.5% gain, while our query-aware TSG-TEM adds another 1.3%. This demonstrates the strong synergistic effect of our components.

Impact of Cognition-Driven Distillation Strategy. Our SC-KD strategy is validated by comparing distillation guided by exploration vs. inheritance branches. Table 4 shows exploration guidance is superior (2.1% higher SumR). This gain stems from its independent target-domain knowledge acquisition, providing a more reliable distillation signal and mitigating teacher bias.

Impact of Shared Static Token between the Two Student Branches. Finally, we verify sharing the static token between student branches. Table 5 shows the shared design significantly outperforms its non-shared alternative, improving SumR by 2.4%. Shared static tokens capture global semantic commonalities, enhancing cross-modal consistency. Independent dynamic tokens model fine-grained, branch-specific features, creating a powerful "static-shared, dynamic-independent" design.

5. CONCLUSION

We have addressed video retrieval in untrimmed videos by proposing a text semantics–guided dual-teacher distillation framework that integrates complementary supervision, a cognition-driven strategy for adaptive transfer, and a temporal module for cross-modal alignment. Extensive experiments on ActivityNet Captions and Charades-STA have demonstrated clear improvements over state-of-the-art baselines, confirming the effectiveness of our approach.

Acknowledgements. This work was partially supported by Shenzhen Science and Technology Foundation under Grant JCYJ20250604173210013, Key Scientific Research Project of the Department of Education of Guangdong Province under Grant 2024ZDZX3012.

6. REFERENCES

[1] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou, "Temporal sentence grounding in videos: A survey and future directions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10443–10465, 2023.

[2] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10638–10647.

[3] 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*, 2022, pp. 246–257.

[4] Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia, "Gmmformer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, vol. 38, pp. 5767–5775.

[5] Yuting Wang, Jinpeng Wang, Bin Chen, Tao Dai, Ruisheng Luo, and Shu-Tao Xia, "Gmmformer v2: An uncertainty-aware framework for partially relevant video retrieval," *arXiv preprint arXiv:2405.13824*, 2024.

[6] Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu, Xun Wang, and Baolong Liu, "Dual learning with dynamic knowledge distillation for partially relevant video retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11302–11312.

[7] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[8] Jimmy Ba and Rich Caruana, "Do deep nets really need to be deep?," *Advances in neural information processing systems*, vol. 27, 2014.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[10] Yi Xie, Yihong Lin, Wenjie Cai, Xuemiao Xu, Huaidong Zhang, Yong Du, and Shengfeng He, "D3still: Decoupled differential distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17181–17190.

[11] 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*. PMLR, 2021, pp. 8748–8763.

[12] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu, "Distilling clip with dual guidance for learning discriminative human body shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 256–266.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[18] Kenneth Ward Church, "Word2vec," *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[22] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth'ee Lacroix, Baptiste Rozi'ere, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[23] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al., "Cogvlm2: Visual language models for image and video understanding," *arXiv preprint arXiv:2408.16500*, 2024.

[24] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[26] 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*, vol. 44, no. 8, pp. 4065–4080, 2021.

[27] 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*, 2020, pp. 9879–9889.

[28] 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*, 2017, pp. 5267–5275.

[29] 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.

[30] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong, "W2vv++ fully deep learning for ad-hoc video search," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1786–1794.

[31] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," *arXiv preprint arXiv:1907.13487*, 2019.

[32] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang, "Reading-strategy inspired visual representation learning for text-to-video retrieval," *IEEE transactions on circuits and systems for video technology*, vol. 32, no. 8, pp. 5680–5694, 2022.

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640.

[34] 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*, vol. 508, pp. 293–304, 2022.

[35] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang, "Cap4video: What can auxiliary captions do for text-video retrieval?," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10704–10713.

[36] Qun Zhang, Chao Yang, Bin Jiang, and Bolin Zhang, "Multi-grained alignment with knowledge distillation for partially relevant video retrieval," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.