CMKD: Distilling Vision Language Models with a Student Model MobileVTL for Action Recognition

Huiting Li[1,2], Yao Yao[1,2], Lei Wang*[1,2], Xixia Xu[1], Dongchen Zhu[1,2], Jiamao Li[1,2]
[1]Bio-vision System Laboratory, Science and Technology on Micro-system Laboratory,
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China
[2]University of Chinese Academy of Sciences, Beijing, China
*wangl@mail.sim.ac.cn

***Abstract*—Pre-trained vision-language large models (VLMs) have achieved strong performance in action recognition after fine-tuning. However, Transformer-based VLMs entail substantial computational costs, which limits their deployment on edge devices. Although recent efficient hybrid architectures combining CNNs and Transformers strike a balance between performance and efficiency, they still underperform compared to fully fine-tuned VLMs. To narrow this gap, we propose a Cross-Modal Knowledge Distillation (CMKD) framework, which includes a lightweight hybrid-architecture student model called MobileVTL and three distillation strategies. CMKD distills knowledge from teacher into student through multi-level feature alignment including contrastive, temporal, and spatial feature alignment, and thus reduces the performance discrepancies between architectures. Additionally, we design a Hierarchical Video-Text Interaction Module (HVTI) to foster dynamic interaction between low-level and high-level semantic features of video and text, which yields more robust multimodal representations. Experiments on Kinetics-400, UCF-101, and HMDB-51 using VTR-B/16 as the teacher demonstrate that MobileVTL achieves high efficiency in both fully-supervised and few-shot video recognition, with minimal computational overhead.**

***Index Terms*—action recognition, knowledge distillation, CLIP**

I. Introduction

Numerous studies have focused on adapting large vision-language models (VLMs) such as CLIP [1] for action recognition through fine-tuning [3]–[7]. However, these Transformer-based models are inherently parameter-heavy and computationally intensive, which limits their applicability in resource-constrained environments. In recent years, many lightweight models have emerged. These models leverage the complementary strengths of CNNs and Transformers, and they achieve excellent performance at a lower computational cost [37]–[39]. Nevertheless, the pursuit of reduced model complexity in lightweight architectures often sacrifices certain intricate feature extraction abilities. Notably, these abilities include capturing long-range dependencies, fine-grained spatio-temporal dynamics, and nuanced cross-modal interactions—capabilities that larger VLMs typically handle well. This inherent limitation consequently leads to a significant performance gap between lightweight models and VLMs, especially pronounced in action recognition. To bridge this performance gap while maintaining the high efficiency demanded by lightweight models, we therefore explore knowledge distillation, a proven model compression technique that has been extensively studied and shown efficacy in both natural language processing and computer vision. We propose the Cross-Modal Knowledge Distillation framework(CMKD) designed for action recognition. This method addresses two key challenges: (1) how to design a lightweight student model with temporal modeling capabilities to achieve an optimal balance between recognition performance and computational efficiency; and (2) how to construct a flexible and efficient cross-modal distillation strategy to imbue the compact student model with the rich representational knowledge and missing capabilities absent in its lightweight design, transferring knowledge from a powerful teacher VLM.

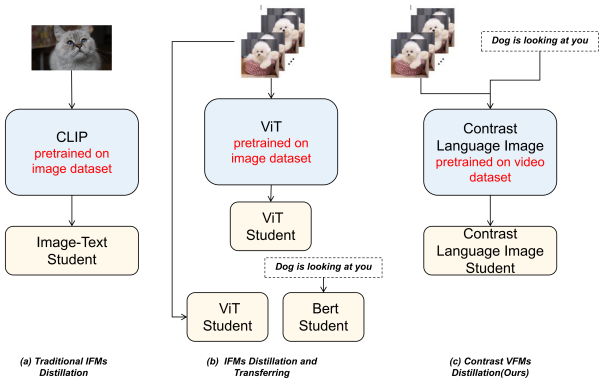

Fig. 1. Distillation Comparision. Figure (a) shows the traditional distillation on image datasets. Figure(b) shows the two-stage distillation for action recognition with image Models. Figure(c) briefly shows our proposed method. We directly distill student from video models on video datasets.

To address the challenge of designing an efficient yet accurate temporal-aware student model, we take advantage of two paradigms: unimodal and multimodal action recognition. On the one hand, existing lightweight models in action recognition predominantly adopt unimodal designs [9], [18], [19]. These models capture temporal information efficiently via sparse attention, separable 3D convolutions, or 2D approximations. However, their recognition accuracy is limited by sole reliance

This work was supported by the Natural Science Foundation of Shanghai (No.24ZR1476900), Pudong New Area Science and Technology Development Fund under Grant No. PKX2024-W02, Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233).

on visual input. However, recent multimodal methods typically fine-tune large VLMs such as CLIP [3]–[7]. These approaches benefit from strong cross-modal representations and thus significantly outperform unimodal models in accuracy, but their efficiency is far lower than that of unimodal models. Since our teacher model is VTR [6], which builds upon CLIP, we propose a lightweight multimodal student network, named MobileVTL, designed to distill knowledge from VTR while inheriting the efficiency of unimodal architectures. MobileVTL employs Uniformer as the backbone of the visual branch and a scaled-down CLIP-Transformer as the text branch. As a representative achievement in lightweight design for unimodal action recognition, Uniformer adopts a hybrid architecture combining CNNs and Transformers, enabling simultaneous capture of local and global temporal features. Compared with pure CNN or pure Transformer architectures, it achieves a better trade-off between efficiency and performance. However, Uniformer lacks inherent multimodal fusion capability. To facilitate MobileVTL in learning the multi-modal representations of the teacher model, we further introduce a lightweight Hierarchical Video-Text Interaction Module (HVTI), which dynamically promotes interaction between textual and visual semantic features at both low and high levels. Specifically, lower layers contain higher resolution and abundant spatial information; thus, we employ Mixed-Modal Tokens to encourage information exchange between the text and visual branches. For higher layers, which are rich in high-level semantics, we apply a modality fusion module to further strengthen semantic interaction.

To tackle the second challenge, Cross-Modal Knowledge Distillation has been widely explored in the context of image-based VLMs [15], [16], [24]. However, as shown in Fig. 1, none of these methods consider the distillation of temporal features in video domain. Although UMT [8]trains a student model through two stages—masked distillation and transfer to the action recognition domain—it primarily focuses on accelerating training, without substantially reducing model complexity. Furthermore, it lacks distillation designs for cross-modal alignment and places strict requirements on the structural consistency between teacher and student models. Therefore, we propose a Cross-Modal Knowledge Distillation strategy (CMKD) to directly distill knowledge from video VLMs, comprising three complementary components: Spatial Base Knowledge Distillation (SBKD), Temporal Feature Knowledge Distillation (TFKD), and Contrastive Relation Knowledge Distillation (CRKD). These strategies respectively reduce feature discrepancies between teacher and student models in terms of spatial features, temporal features, and contrastive representations, enabling the student model to acquire stronger multi-modal representation capabilities without requiring structural consistency with the teacher model.

In summary, our principal contributions are fourfold:

- We propose the end-to-end video Cross-Modal Knowledge Distillation framework (CMKD) tailored for action recognition with Spatial Base Knowledge Distillation (SBKD), Temporal Feature Knowledge Distillation (TFKD), and Contrastive Relation Knowledge Distillation (CRKD). It significantly reduces the parameter count and computational cost of the model while retaining the strong multimodal modeling capabilities of the teacher VLM.
- We introduce a lightweight multi-modal student model, MobileVTL, which incorporates a Hierarchical Video-Text Interaction Module (HVTI). This module dynamically promotes interaction between textual and visual semantic features at both low and high levels, thereby enhancing the multimodal representation capability of the student model.
- When using VTR-B/14 pre-trained on the Kinetics-400M dataset as the teacher model, the proposed MobileVTL requires only 86M parameters (52.6% of the teacher model) and 69.37 GFLOPs (18.8% of the teacher), achieving 81.5% Top-1 accuracy on Kinetics-400. Furthermore, our smaller variant achieves 80.4% Top-1 accuracy with only 39.65M parameters (24.2% of the teacher) and 31.29 GFLOPs (8.5% of the teacher), demonstrating an excellent trade-off between performance and efficiency.

II. Related Work

A. Action Recognition

Action recognition tasks are mainly categorized into unimodal and multimodal methods. With the increasing demand for temporal modeling capabilities, the unimodal method has gradually shifted from traditional two-stream convolutional network architectures to Transformer-based large model paradigms. Early two-stream CNNs combined spatial features with optical flow-derived temporal features but struggled with complex temporal dependencies. While 3D convolutional methods like I3D [17], R(2+1)D [18], and TSM [19] improved accuracy, their receptive field limited long-video processing, prompting the shift to Transformer. Transformer-based method like Timesformer [10], VidTr [23] and ViViT [2] enhance the representation of dynamic actions from spatio-temporal separable attention modules, video sampling strategies, and data modeling approaches, but all incur considerable computational costs. Uniformer [9] combined CNNs with Transformers significantly reducing the number of model parameters while maintaining high accuracy. This provides a suitable backbone for our student model.

In recent years, following the breakthrough of the VLMs like CLIP [4], researchers have sought to transfer it to the video domain. Existing works mainly adopt two migration strategies: one leverages finetuning with adapters and temporal modeling modules [3], [6], [7]; the other introduces extra tokens to finetune the model [4], [13]. Experiments show that multimodal methods outperform unimodal ones.However, most of these methods focus on finetuning pretrained models to accelerate training. Yet, due to CLIP's large parameter size, the resulting action recognition models still suffer from high parameter counts and computational costs. To address this, we propose the Cross-Modal Knowledge Distillation framework

tailored for action recognition to balance performance and efficiency, and design a lightweight student model, MobileVTL.

B. Knowledge Distillation

Knowledge distillation (KD), as a core technique for model compression and transfer learning, has significantly advanced the development of lightweight models. In the Vision Transformer domain, Tiny-ViT [14] pioneered knowledge distillation from teacher models but relied solely on soft labels generated by teachers while neglecting real labels and intermediate layer features, leading to limited generalization performance in real-world scenarios. To address this, Wang et al. [22] proposed an improved strategy combining attention probes with teacher intermediate features, enhancing student inputs through wild data augmentation, yet still suffering from cross-entropy loss-induced recognition accuracy degradation. Touvron et al. [21] introduced distillation tokens and real-label tokens into inputs, integrating cross-entropy constraints from real labels to improve knowledge transfer robustness.

The breakthrough of CLIP has driven research on dual-modal knowledge distillation. TinyClip [24] introduced two core technologies: affinity simulation to optimize modality interaction pathways and parameter inheritance to accelerate student training, though its strict structural consistency requirement limited model adaptability. Building on this, ClipKD [15] expanded multi-dimensional distillation strategies (relationship, feature, gradient, and contrast paradigms) and validated cross-modal knowledge transfer through experiments with lightweight MobileViT. Apple's MobileClip [16] further innovated by constructing an enhanced dataset using image captioning models and a strong CLIP encoder ensemble, offering a data-driven paradigm for knowledge distillation.

However, existing KD studies focus predominantly on image domain, with limited applications in downstream tasks like action recognition. Some works attempt two-stage training (distillation followed by fine-tuning) to adapt CLIP to video domains [8]. However, UMT [8] relies on complex adapter structures and only accelerates training, without achieving notable model lightweighting. Notably, direct distillation from VLMs to small models for action recognition remains unexplored. So, we propose the Cross-Modal Knowledge Distillation framework tailored for action recognition with four distillation losses targeting spatial, temporal, textual, and contrastive aspects to mimic the teacher without requiring structural consistency.

III. Method

A. MobileVTL

1) Architecture: The student model is a core component of the knowledge distillation framework. In this study, we select VTR [6] as teacher, which is an advanced large-scale multi-modal action recognition model with powerful visual-textual representation capabilities and temporal modeling performance. To fully transfer the knowledge from teacher, we design a lightweight dual-branch action recognition student model, and its overall architecture is illustrated in Fig. 2. The

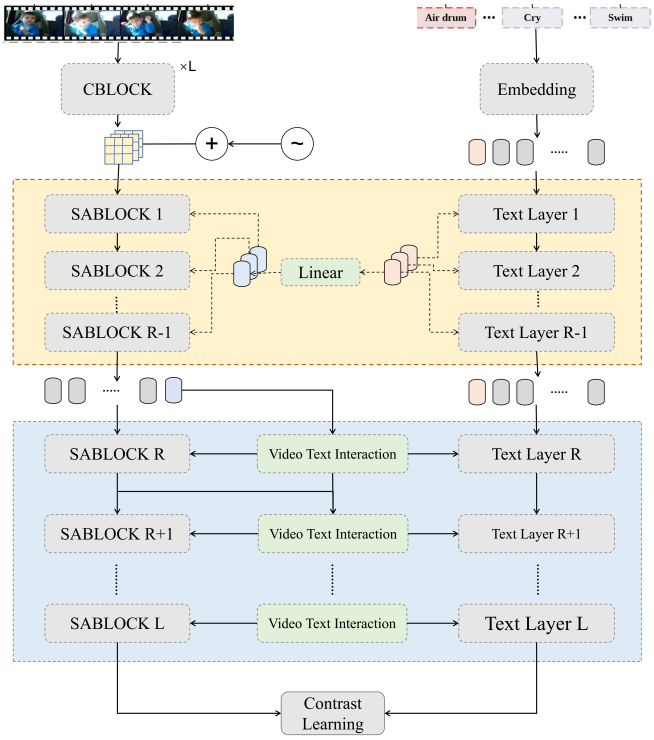

Fig. 2. MobileVTL architecture. Figure illustrates the overall framework of the student model MobileVTL, which consists of a visual branch, a textual branch, and an intermediate hierarchical cross-modal interaction mechanism.

visual branch is based on Uniformer [9], incorporating a hybrid design that fuses CNNs and Transformers, thereby enabling it to simultaneously capture both local and global temporal dynamics. Specifically, the first two layers use convolutional structures, while the last two layers adopt Transformer structures. The textual branch adopts a smaller CLIP-Transformer.

2) Hierarchical Video-Text Interaction Module (HVTI): Due to the reduced parameter size of the student model, its temporal modeling capability is relatively weak, leading to a robustness gap compared with large VLMs. The teacher model has a powerful cross-modal fusion mechanism that enhances robustness; however, the student model underutilizes intermediate-layer features and overlooks potential associations between text and visual modalities [3]. To comprehensively learn from the teacher, we propose a Hierarchical Video-Text Interaction (HVTI) module to improve representation quality. The core components of this module include learnable Mixed-Modal Tokens and a Video-Text Interaction unit. As shown in Fig. 2, at the shallow stages of the model, bidirectional cross-modal knowledge transfer between the visual encoder layers and textual encoder layers is achieved via Mixed-Modality Tokens. In contrast, at deeper stages where semantic information is more abundant, a more structurally complex video-text interaction mechanism is introduced to facilitate deep bidirectional interaction: on one hand, textual information guides and enhances the temporal modeling capability of the visual branch; on the other hand, visual information supervises

the representation learning of the textual branch.

Mixed-Modal Tokens. We randomly initialize n learnable tokens as the Mixed-Modal Tokens for use by the textual branch, denoted as $[MT]_t \in R^{n \times D_t}$. Assume the text encoder has L layers; the aforementioned operation is only performed on the first R layers ($R < L$), ultimately yielding a sequence of textual token sequences $[MT_t^0, MT_t^1, ..., MT_t^{R-1}]$. To promote information interaction between the visual and textual branches, we map the textual token sequence into the visual feature space via a learnable weight matrix, obtaining an aligned visual-textual token sequence $[MT_{\mathrm{v}}^0, MT_{\mathrm{v}}^1, ..., MT_{\mathrm{v}}^{R-1}], [MT]_{\mathrm{v}}^j \in R^{n \times D_v}$. As depicted in Fig. 2, the initialized Mixed-Modal Tokens are concatenated with the video input and text input, respectively, before being fed into subsequent encoders. This operation realizes bidirectional cross-modal knowledge transfer between the two branches.

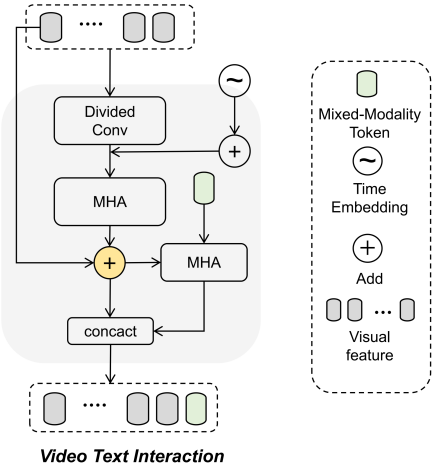

Fig. 3. Structure of Video Text Interaction Module.

Video-Text Interaction. To facilitate deep bidirectional interaction between the video and text modalities, this paper proposes a parallel multi-layer Video-Text Interaction (VTI) module. The structure of this module is illustrated in the Fig. 3, with inputs consisting of visual features and the Mixed-Modal Tokens. During processing, the visual features are subjected to local feature extraction via depthwise separable convolutions. They are then combined with added temporal embedding information to incorporate temporal context. The core innovation of the module is the introduction of Mixed-Modality Tokens MT_v^R as a supervisory signal. These tokens are jointly input with visual features into a multi-head attention mechanism, which achieves cross-modal feature interaction. The fused Mixed-Modal Tokens are then fed back into the textual branch, enabling the transfer of visual-to-textual semantic information. Additionally, the method incorporates skip connections and feature addition operations to enhance gradient flow and stabilize the training process. Through this multi-layer stacked interaction mechanism, the VTI module effectively achieves deep semantic alignment and information fusion between the video and text modalities.

B. Cross-Modal Knowledge Distillation

The introduction mentioned a significant performance gap between teacher and student. To reduce the performance discrepancy, as illustrated in Fig. 4, we propose a cross-modal distillation framework that assists student from three perspectives: Spatial Baseline Knowledge Distillation (SBKD), Temporal Feature Knowledge Distillation (TFKD) and Contrastive Relational Knowledge Distillation (CRKD).

1) Cross Relation Knowledge Distillation: The core idea of contrastive learning is to maximize the similarity between paired video-text embeddings based on a contrastive similarity distribution. Therefore, a straightforward type of knowledge to distill is the output-oriented contrastive relational contrast distribution. This concept has been utilized in prior works for image classification [25] and semantic segmentation [26]. We apply it directly to clip-based video recognition; experiments demonstrate that the contrastive distribution captures the structured relationships between feature embeddings. A good teacher often constructs a well-structured feature space. CRKD enables student to better imitate the structured semantic relationships from teacher, further improving the quality of the feature representations. Assuming a batch size of B, the final spatio-temporal representation of the student model is v_b^S, its textual representation is t_b^S, the final spatio-temporal representation of the teacher model is v_b^T, and its textual representation is t_b^T. The contrastive loss distributions for the student and teacher models are calculated as follows, respectively:

$$p_b^T[k] = \frac{exp(v_b^T \cdot t_k^T / \tau^T)}{\sum_{i=1}^{D} exp(v_b^T \cdot t_i^T / \tau^T)}, \tag{1}$$

$$p_b^S[k] = \frac{exp(v_b^S \cdot t_k^S / \tau^S)}{\sum_{i=1}^{C} exp(v_b^S \cdot t_i^S / \tau^S)} \tag{2}$$

where $k \in [1, 2, ..., B]$is the batch index.

Consequently, the contrastive relational distillation loss is calculated as follows:

$$Loss_{CRKD} = \frac{1}{|B|} \sum_{b=1}^{|B|} \sum_{k=1}^{|B|} p_b^T[k] log \frac{p_b^T[k]}{p_b^S[k]}. \tag{3}$$

2) Spatial Base Knowledge Distillation: High-level inputs in the visual branch have a large number of channels and rich semantic information but low resolution and minimal spatial details, while lower levels exhibit the opposite properties—high input resolution and rich spatial information but weaker semantics. These spatial details are also crucial for knowledge distillation [27]. Since student is CNN architecture at lower layers but teacher is Transformer-based architecture, feature difference between teacher and student is huge. Therefore, we propose Spatial Baseline Knowledge Distillation (SBKD) to align features and overcome the feature discrepancies caused by this heterogeneity, which specifically aims to capture the teacher's spatial encoding capabilities at

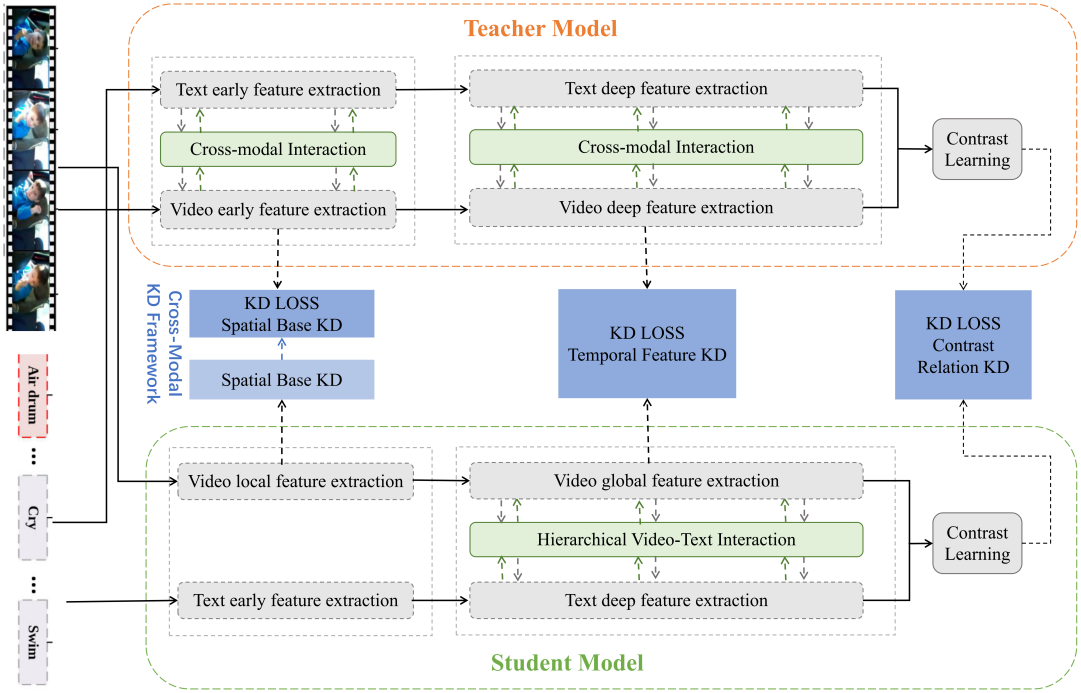

Fig. 4. Cross-Modal Knowledge Distillation Framework. CMKD consists of three distilation strategies: Spatial Base KD, Temporal Feature KD and Contrast Relation KD.

lower levels. SBKD prompts the student to learn from the teacher by treating each layer of the student model as a basis for the teacher's feature space. The mathematical principle is as follows:

$$L_i^T = \sum_{j=1}^{N} w_{ij} L_j^S, \qquad i \in (1, M) \tag{4}$$

where M represents the number of teacher layers, N represents the number of student layers, L_i^T is the feature of the i-th teacher layer, L_j^S is the feature of the j-th student layer, and w_{ij} is the weight matrix for each layer.

Based on the premise that each layer of the teacher model is a linear combination of multiple layers of the student model, SBKD performs alignment through the following operations.

Feature Rearrangement and Similarity Calculation. Since the input and output of convolutional layers are 3D, we convert them to 2D via pixel rearrangement. Suppose the feature L_j^S of j-th layer of the student model has a size of $[BT, C_j, H_j, W_j]$, and the feature L_i^T of the i-th layer of the teacher model has a size of $[BT, L, D]$. L_j^S is mapped via a linear transformation to $L_j^S = [BT, \frac{p^2 L}{H_j W_j}, H_j, W_j]$. After pixel rearrangement and flattening, it becomes $L_j^S = [BT, L, D]$. The flattened student features are concatenated to obtain $L^S = [BT, L, N, D]$. Similarly, the teacher features are concatenated to obtain $L^T = [BT, L, M, D]$. Finally, a similarity matrix $W = [BT, L, M, N]$ is computed.

Masking and Alignment. The lower layers of the student should learn from the lower layers of the teacher, while the higher layers of the student should learn from all layers of the teacher. Based on this principle, we designed a sequential mask. To ensure training convergence, we also applied a softmax operation after the sequential mask. The masking method is calculated as follows:

$$mask(i,j) = \begin{cases} 0, & \text{if } j > n_0 \text{ and } i \le m_0 \\ 0, & \text{if } j > n_1 \text{ and } m_0 < i \le m_1 \\ 1, & \text{else} \end{cases} \tag{5}$$

where n_0, m_0, n_i, m_i $(0 < n_0 < n_1 < N, 0 < m_0 < m_1 < M)$ are hyperparameters. These hyperparameters are set to 1, 6, 1, and 2 respectively during training.

The masked similarity matrix W' is then obtained by performing an element-wise multiplication of the mask matrix and the similarity matrix W. Finally, W' is multiplied by $L^S = [BT, L, N, D]$ to yield the masked student features. The loss value is calculated using the Kullback-Leibler (KL) divergence between these features and the teacher features $L^T = [BT, L, M, D]$, computed as follows:

$$Loss_{SBKD} = KL((mask \odot W) \times L^S,\ L^T). \tag{6}$$

3) Temporal Feature Knowledge Distillation: Temporal modeling capability is crucial for action recognition. Since the selected teacher model inherently possesses strong temporal modeling capabilities, directly distilling the teacher's temporal features can maximally reduce the temporal feature distance between the student and teacher. Assuming a batch size of B, the final spatio-temporal representation of the student model is

TABLE I
KINETICS-400 RESULTS.

Method	Pretrain	frames	views	Top1	GFLOPs
Uniformer-S [9]	ImageNet-21K	8	1×4	78.4	17.5
Uniformer-S [9]	ImageNet-21K	16	1×4	80.8	41.8
DSFNet [11]	ImageNet-21K	16	-	78.6	-
Mformer-HR [34]	ImageNet-21K	16	3×10	81.1	959.0
MVFNet-en [41]	ImageNet-21K	24	3×10	79.1	359
VideoSwin-T [31]	ImageNet-21K	32	4×3	78.8	88.0
VideoSwin-S [31]	ImageNet-21K	32	4×3	80.6	166.0
ViViT-L/16 [2]	ImageNet-21K	32	1×1	80.6	3980
MViTv2-B [35]	ImageNet-21K	32	1×5	81.2	255.0
VLLM [40]	CLIP-400M	8	1×1	77.3	-
ActionCLIP-ViT-B/32 [7]	CLIP-400M	8	1×1	78.4	35.4
ActionCLIP-ViT-B/16 [7]	CLIP-400M	8	1×1	81.1	140.8
Vita-CLIP-B/16 [32]	CLIP-400M	8	1×1	80.5	97.0
BIKE ViT-B/32 [5]	CLIP-400M	8	1×1	81.4	-
X-CLIP-B/32 [3]	CLIP-400M	8	4×3	80.4	39.0
X-CLIP-B/32 [3]	CLIP-400M	16	4×3	81.1	75.0
Ours					
MobileVTL-S	ImageNet-21K	8	4×3	79.2	**15.7**
MobileVTL-S	ImageNet-21K	16	4×3	80.8	31.3
MobileVTL-B	ImageNet-21K	16	4×3	**81.5**	69.4

TABLE II
INFERENCE SPEED OF THE PROPOSED MOBILEVTL.

Model	Params	GPU Pytorch	GPU ONNX
Teacher-16	163.5M	122.2 ± 3.3ms	-
Teacher-8	163.5M	117.4 ± 1.4ms	-
MobileVTL-B-16	89.7M	81.2 ± 1.3ms	47.5 ± 0.4ms
MobileVTL-S-16	**40.6**M	**39.3** $\pm$ **0.5**ms	22.3 ± 0.1ms
MobileVTL-S-8	**40.6**M	41.2 ± 0.9ms	**11.7** $\pm$ **1.0**ms

v_b^S, and that of the teacher model is v_b^T. The temporal feature distillation loss is calculated as follows:

$$Loss_{TFKD} = \frac{1}{|B|} \sum_{b=1}^{|B|} \sum_{k=1}^{D} v_b^T[k] log \frac{v_b^T[k]}{v_b^S[k]}. \quad (7)$$

IV. EXPERIMENTS

A. Implementation

Datasets. We conduct experiments on scene-related Kinetics400 (K400) [28], HMDB-51 [29] and UCF-101 [30] datasets.

Architectures. For MobileVTL-S, we set the number of Video-Text Interaction(VTI) blocks to 4 for all datasets. That means the number of initialized Mixed-Modal Tokens will be set to 8. For MobileVTL-B, we set the number of VTI blocks to 8 for all datasets. That means the number of initialized Mixed-Modal Tokens will be set to 20. The hyperparameter of CMKD is set to 1, 1, and 0.1 for CRKD, TFKD and SBKD respectively.

B. Fully-supervised Experiments

Settings. We conduct the fully-supervised experiments on K400. We sample 8 or 16 frames with a sparse sampling method in fully-supervised experiments. The experiments were conducted using the Adam optimizer with a cosine learning rate scheduler. The initial learning rate was set to 8×10^{-6}, and a warm-up phase of 5 epochs was applied at the beginning of training. Weight decay was configured as 0.001. All experiments were performed on four NVIDIA GeForce RTX 3090 GPUs.

Results. As shown in Table. I, when using 8-frame input, MobileVTL-S reaches a Top1 accuracy of 79.2% with a computation amount of 15.7 GFLOPs. Compared with the sub-optimal model MViTv2-B's Top1 accuracy of 81.2%, it only loses 2.0 percentage points, but the computation amount is reduced by 93.8%; compared with VideoSwin, MobileVTL-S achieves a 0.4% increase in Top1 accuracy with only 17.8% GFLOPs. When extended to the 16-frame architecture , MobileVTL-B achieves a Top1 accuracy of 81.5% at a computational cost of 69.4 GFLOPs, which is 0.4 percentage points higher than that of ActionCLIP-ViT-B/16 model, and the computation amount is reduced by 50.7%. MobileVTL achieves performance comparable to or even better than that of mainstream large models with fewer parameters and lower computational costs. The inference speed after ONNX acceleration is also reported in Table. II. Speed experimentation is implemented on NVIDIA TITAN RTX and Intel(R) Xeon(R) CPU E5-2682 V4.

C. Few-shot Experiments

TABLE III
PERFORMANCE COMPARISON ON THE HMDB-51 DATASET

Method	K=2	K=4	K=8	K=16
TSM [19]	17.5	20.9	18.9	30.0
TimeSFormer [10]	19.6	40.6	49.4	55.4
Video-Swin-B [31]	20.9	41.3	47.9	56.1
X-CLIP-B/16 [3]	53.0	57.3	62.8	64.0
VicTR ViT-B/16 [36]	**60.0**	63.2	66.6	70.7
Ours				
MobileVTL-S	57.0	**64.5**	**70.3**	**72.1**

TABLE IV
PERFORMANCE COMPARISON ON THE UCF-101 DATASET

Method	K=2	K=4	K=8	K=16
TSM [19]	25.3	47.0	64.4	61.0
TimeSFormer [10]	48.5	75.6	83.7	89.4
Video-Swin-B [31]	53.3	74.1	85.8	88.7
X-CLIP-B/16 [3]	76.4	83.4	88.3	91.4
VicTR ViT-B/16 [36]	87.7	92.3	93.6	95.8
VTR-B/16 [6]	**90.4**	91.4	93.7	95.7
Ours				
MobileVTL-S	88.1	**94.1**	**95.7**	**97.0**

Settings. We conduct a general K-shot setting. We sample 32 frames with a sparse sampling method and finetune 25 epochs in few-shot experiments. The other settings are the same as fully-supervised experiments. We use HMDB-51 and UCF-101 to evaluate the few-shot performance of MobileVTL-S. Following X-CLIP [3], we randomly sample 2, 4, 8, or 16 examples per class to create our few-shot training sets.

Results. We report few-shot transfer performance on HMDB51 and UCF-101 in Table III and Table IV. Our method significantly outperforms VicTR Vit-B/16 [36]. For the extreme case where K=2, our MobileVTL-S outperforms X-CLIP by 4% on HMDB-51 and by 11.7% on UCF-101. It indicates that our MobileVTL exhibits strong generalization capabilities for few-shot video tasks.

D. Ablation Experiments

TABLE V
DISTILLATION STRATEGY ABLATION

CRKD	TFKD	SBKD	Top1	Top5
×	×	×	74.9	90.7
√	×	×	77.0	92.8
×	√	×	75.6	91.7
×	×	√	76.0	92.2
√	√	×	77.2	92.8
√	√	√	**77.4**	**93.0**

The effects of the proposed distillation strategies. To validate the effectiveness of the distillation strategy, we utilize testing strategies of 1×1 views with 8 frames input on MobileVTL-S for K400. As shown in Table V, ablation studies demonstrate that the four distillation strategies are complementary: applying CRKD alone produces a 2.1% improvement in Top1 accuracy, while combining CRKD, TFKD and SBKD further increases the accuracy to 77.4% (+2.5% over the baseline). Contrastive relation distillation (CRKD) contributes markedly to temporal modeling, while spatial baseline distillation (SBKD) enhances spatial detail capture.

TABLE VI
MOBILEVTL MODULE ABLATION

Mixed-Modal Tokens	VT_interaction	Top1	Top5
×	×	78.6	93.5
√	×	78.8	93.8
√	√	**79.6**	**94.1**

The effects of the proposed components. Table VI presents the performance improvements brought by the proposed modules based on the backbone. We utilize testing strategies of 1×1 views with 16 frames input on MobileVTL-B for K400. As shown in Table VI, incorporating the Mixed-Modal Tokens leads to a 0.2% increase in Top1 accuracy, while combining it with Video-Text Interaction yields a cumulative gain of 1.0%. The Mixed-Modal Tokens addresses shallow semantic alignment via bidirectional cross-modal propagation, whereas the Video-Text Interaction Module facilitates spatio-temporal feature fusion in deeper layers.

TABLE VII
VT INTERACTION BLOCK NUMBER ABLATION

Video-Text Interaction Block Number	Top1	Top5
4	**77.3**	**92.9**
6	76.8	92.7
8	77.2	92.8
12	77.0	92.7

The number of Video-Text Interaction blocks. We utilize testing strategies of 1×1 views with 8 frames input on MobileVTL-S for K400. As shown in Table VII, experiments on the number of VTI blocks indicate that the configuration with 4 blocks achieves the best performance (Top1 77.3%), outperforming both the other setups. An excessive number of interaction blocks may induce over-smoothing issues, thereby weakening the diversity of temporal features.

V. CONCLUSION

This paper proposes a Cross-Modal Knowledge Distillation(CMKD) mechanism tailored for action recognition, along with a family of lightweight student models named MobileVTL. The framework incorporates Spatial Base Knowledge Distillation (SBKD), Temporal Feature Knowledge Distillation (TFKD), and Contrastive Relation Knowledge Distillation (CRKD), which are jointly employed to reduce the representational discrepancy between teacher and student networks without requiring structural consistency between them. Furthermore, MobileVTL enhances accuracy through a hierarchical video–text interaction mechanism. Extensive experiments under two distinct learning scenarios on three widely-used video benchmarks demonstrate the competitive performance of our method.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *ICCV*, 2021, pp. 6836–6846.

[3] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," in *ECCV*, 2022, pp. 1–18.

[4] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," in *ECCV*, 2022, pp. 388–404.

[5] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," in *CVPR*, 2023, pp. 6620–6630.

[6] S. Yu, L. Chen, X. Zhang, and J. Li, "VTR: Bidirectional video-textual transmission rail for CLIP-based video recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2024, pp. 1–6.

[7] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," arXiv preprint arXiv:2109.08472, 2021.

[8] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19948–19960.

[9] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y.J. Qiao, "UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning," arXiv preprint arXiv:2201.04676, 2022.

[10] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in ICML, vol. 2, no. 3, 2021, p. 4.

[11] B. Sun, X. Ye, T. Yan, Z. Wang, H. Li, and Z. Wang, "Discriminative segment focus network for fine-grained video action recognition," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 7, pp. 1–20, 2024.

[12] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "AIM: Adapting Image Models for Efficient Video Action Recognition," arXiv preprint arXiv:2302.03024, 2023.

[13] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, "Promptkd: Unsupervised prompt distillation for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26617–26626.

[14] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2022, pp. 68–85.

[15] C. Yang, Z. An, L. Huang, J. Bi, X. Yu, H. Yang, … and Y. Xu, "Clip-kd: An empirical study of clip model distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15952–15962.

[16] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel, "Mobileclip: Fast image-text models through multi-modal reinforced training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15963–15974.

[17] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *European Conference on Computer Vision*, 2018, pp. 318–335.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[19] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *International Conference on Computer Vision*, 2019, pp. 7082–7092.

[20] S. Ren, Z. Gao, T. Hua, Z. Xue, Y. Tian, S. He, and H. Zhao, "Co-advise: Cross inductive bias distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16773–16782.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers distillation through attention," in *International Conference on Machine Learning*, PMLR, 2021, pp. 10347–10357.

[22] J. Wang, M. Cao, S. Shi, B. Wu, and Y. Yang, "Attention probe: Vision transformer distillation in the wild," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 2220–2224.

[23] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, … and J. Tighe, "Vidtr: Video transformer without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13577–13587.

[24] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang, et al., "Tinyclip: Clip distillation via affinity mimicking and weight inheritance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21970–21980.

[25] C. Yang, Z. An, H. Zhou, et al., "Online knowledge distillation via mutual contrastive learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, 2023, pp. 10212–10227.

[26] C. Yang, H. Zhou, Z. An, et al., "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12319–12328.

[27] G. Ghiasi, T. Y. Lin, and Q. V. Le, "Nas-fpn: Learning scalable feature pyramid architecture for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7036–7045.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[29] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *ICCV*, 2011, pp. 2556–2563.

[30] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[31] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *CVPR*, 2022, pp. 3202–3211.

[32] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23034–23044.

[33] S. Tu, Q. Dai, Z. Wu, Z. Q. Cheng, H. Hu, and Y. G. Jiang, "Implicit temporal modeling with learnable alignment for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19936–19947.

[34] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, … and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12493–12506, 2021.

[35] Y. Li, C. Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[36] K. Kahatapitiya, A. Arnab, A. Nagrani, and M. S. Ryoo, "VidTR: Video-conditioned text representations for activity recognition," arXiv preprint arXiv:2304.02560, 2023.

[37] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *CVPR*, 2022.

[38] S. Mehta and M. Rastegari, "Separable selfattention for mobile vision transformers," in *ICLR*, 2022.

[39] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "MobileFormer: Bridging MobileNet and transformer," in *CVPR*, 2022.

[40] J. Duan and Z. Jin, "Enhancing Video Recognition Through Cross-Attention," in *Proceedings of the 2025 2nd International Conference on Intelligent Computing and Data Mining (ICDM)*, Guangzhou, China, 2025, pp. 183–187, doi: 10.1109/ICDM68174.2025.11309618.

[41] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding, "Mvfnet: Multi-view fusion network for efficient video recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 2943–2951.