Distilling Cross-Modal Knowledge via Feature Disentanglement

Junhong Liu[1], **Yuan Zhang**[2], **Tao Huang**[3], **Wenchao Xu**[4], **Renyu Yang**[1*]
[1]School of Software, Beihang University
[2]School of Computer Science, Peking University
[3]Shanghai Jiao Tong University
[4]Hong Kong University of Science and Technology
{junhongliu, renyuyang}@buaa.edu.cn, zhangyuan@alumni.pku.edu.cn, t.huang@sjtu.edu.cn, wenchaoxu@ust.hk

Abstract

Knowledge distillation (KD) has proven highly effective for compressing large models and enhancing the performance of smaller ones. However, its effectiveness diminishes in cross-modal scenarios, such as vision-to-language distillation, where inconsistencies in representation across modalities lead to difficult knowledge transfer. To address this challenge, we propose frequency-decoupled cross-modal knowledge distillation, a method designed to decouple and balance knowledge transfer across modalities by leveraging frequency-domain features. We observed that low-frequency features exhibit high consistency across different modalities, whereas high-frequency features demonstrate extremely low cross-modal similarity. Accordingly, we apply distinct losses to these features: enforcing strong alignment in the low-frequency domain and introducing relaxed alignment for high-frequency features. We also propose a scale consistency loss to address distributional shifts between modalities, and employ a shared classifier to unify feature spaces. Extensive experiments across multiple benchmark datasets show our method substantially outperforms traditional KD and state-of-the-art cross-modal KD approaches.

Code — https://github.com/Johumliu/FD-CMKD
Extended version — https://arxiv.org/abs/2511.19887

Introduction

Knowledge Distillation (KD) has emerged as a fundamental technique for model compression and performance improvement. The core concept of KD involves utilizing a large and high-capacity teacher model to mentor a smaller yet more efficient student model. Through this process, the student model learns to approximate the behavior of the teacher model, often achieving comparable or even superior performance despite its reduced complexity.

Despite the substantial success of traditional distillation methods in unimodal settings (Huang et al. 2022b; Zhang et al. 2023), such as image or text tasks, many real-world applications inherently involve multimodal data, including vision, language, and audio. In these cross-modal scenarios, effectively transferring knowledge among modalities presents unique challenges. As a result, researchers have increasingly turned their attention to devise cross-modal knowledge distillation (CMKD) framework to enhance the performance of a student model in one modality by leveraging the knowledge of a teacher model in a different modality.

CMKD presents more formidable challenges than its single-modal KD. The root cause of this lies in the inherent representational inconsistency among features from different modalities. Specifically, features from each modality concurrently encode both the cross-modal shared semantic content (the "*what*") and the modality-specific detailed characteristics (the "*how*"). Consequently, directly imposing a unified, strong alignment loss on these heterogeneous features often leads to "representational conflicts", compelling the student model to distort the unique characteristics of its native modality, thereby undermining its intrinsic expressive capabilities and impeding effective knowledge transfer.

While preliminary progress has been made in cross-modal distillation, existing methods (Gupta, Hoffman, and Malik 2016; Thoker and Gall 2019; Afouras, Chung, and Zisserman 2020; Liu, Jia, and Wang 2023; Cao et al. 2025) often suffer from several limitations: they are typically restricted to specific scenarios or primarily focus on distillation from a stronger modality to a weaker one. Recently, C2KD (Huo et al. 2024) introduced a cross-modal distillation technique based on logits to reduce the gap between modalities. However, it overlooks the distillation of challenging samples with misaligned soft labels, which are crucial for effective cross-modal knowledge transfer. Furthermore, C2KD exclusively emphasizes logit-level distillation, while intermediate features, which encapsulate richer modal details and semantic information in cross-modal settings, play a more pivotal role in facilitating complementary and effective knowledge transfer. This motivates us to explore cross-modal feature distillation, addressing the challenges posed by discrepant feature representations and misaligned feature spaces.

Inspired by (Williams and Li 2018; Xu et al. 2020; Pham et al. 2024; Zhang et al. 2024a), we investigate the frequency representations of multimodal features, and show that features decomposed into different frequency bands exhibit varying levels of effectiveness for representation. Specifically, low-frequency features tend to encode more modality-shared semantic information, whereas high-frequency features exhibit greater modality-specificity. As illustrated in

*Corresponding Author

Feature	CREMA-D	AVE
Original Features	0.84	0.74
Low Frequency	0.91	0.85
High Frequency	−0.02	−0.01

Table 1: Cross-Modal Similarity Results for CREMA-D and AVE Datasets across Different Features.

Table 1, low-frequency features consistently demonstrate higher inter-modal similarity compared to raw features across various datasets. This suggests that low-frequency features serve as effective carriers of modality-agnostic semantic information. In contrast, high-frequency features exhibit remarkably low inter-modal similarity, which further substantiates that high-frequency features predominantly encode modality-specific details that are challenging to directly align. This provides critical insights into how to separate and independently process these two distinct types of information, thereby enabling the formulation of effective distillation strategies.

Based on these insights, we propose a new approach of decomposing features into low-frequency and high-frequency components for distillation, applying distinct loss functions to each of them, respectively. For low-frequency features, we employ the traditional mean squared error (MSE) loss to ensure "strong consistency" such that modality-generic information from the teacher model can be better captured. Meanwhile, since high-frequency features contain modality-specific knowledge and exhibit greater variation, making the full alignment less suited, we introduce the logarithmic mean squared error (logMSE) loss to maintain "weak consistency". Furthermore, given that distribution differences are critical for effective knowledge transfer (Pan and Yang 2009; Li et al. 2019; Sun and Saenko 2016), we propose a scale consistency loss by the alignment of different modalities through feature standardization, to address the significant discrepancies between modalities. This allows the model to focus on intrinsic discriminative features and reduces the impact of scale variations. We also introduce a shared classifier to align feature spaces further to ensure consistent decision boundaries across modalities, enhancing the effectiveness of cross-modal distillation.

To summarize, our contributions are three-fold:

- By analyzing features across different modalities, we found that low-frequency features exhibit stronger cross-modal similarity compared to high-frequency features. Based on this crucial insight, we propose a novel frequency-domain decoupled Cross-Modal Knowledge Distillation framework, specifically designed to address the inherent conflict between semantic and detailed information during cross-modal distillation.
- We designed a set of differentiated distillation strategies tailored to process the decoupled features. Furthermore, we introduce a scale consistency loss and employ a shared classifier to further optimize the alignment of cross-modal feature spaces.
- We perform extensive experiments on diverse datasets, covering various modalities and tasks and employing different network architectures to demonstrate the effectiveness of our approach.

Related Work

Generic Knowledge Distillation

Traditional knowledge distillation methods fall into two main categories: **logit-based distillation** and **feature-based distillation**. Logit-based distillation, first introduced by (Hinton, Vinyals, and Dean 2015), transfers knowledge by minimizing the Kullback-Leibler (KL) divergence between the teacher and student model outputs, helping the student learn inter-class relationships. DKD (Zhao et al. 2022) refines this process by separating target and non-target class distillation, allowing better learning of category-specific information. DML (Zhang et al. 2018) enhances transfer by having two models train each other as teachers, while DIST (Huang et al. 2022a) uses correlation loss to improve logit distillation by capturing inter-class and intra-class relationships. Feature-based distillation uses intermediate features from the teacher model to supervise the student, aiding in better data representation. FitNet (Romero et al. 2014) was the first to have the student mimic the teacher's intermediate features. Review (Chen et al. 2021) introduced a mechanism that allows the student to learn teacher features layer by layer. Relational Knowledge Distillation (RKD) (Park et al. 2019) focuses on transferring relationships between samples, while PKD (Cao et al. 2022) preserves relational information using Pearson correlation. FreeKD (Zhang et al. 2024a) starts to use frequency information to represent visual features for distillation on dense prediction tasks. OFD (Heo et al. 2019) employs partial L2 loss, ignoring unhelpful features and focusing on beneficial ones. Generic Knowledge Distillation performs well in single-modal tasks, but due to differences in modality representation, it underperforms in cross-modal scenarios.

Cross-modal Knowledge Distillation

In cross-modal knowledge distillation (CMKD) research, various studies have explored effective methods for transferring knowledge across modalities. (Gupta, Hoffman, and Malik 2016) proposed an early CMKD framework that transferred labeled supervision from RGB images to depth and optical flow, enhancing the performance of these unlabeled modalities. For action recognition, studies such as (Thoker and Gall 2019; Dai, Das, and Bremond 2021; Zhang et al. 2025; Lee et al. 2023) leveraged RGB or optical flow to design CMKD frameworks that improved action detection accuracy. In medical image segmentation, (Wang et al. 2023) addressed missing modalities by selecting the most contributive one for cross-modal distillation in multi-modal learning. CMKD has also been applied to tasks like camera-radar object detection and visual place recognition, as seen in works like (Zhao, Song, and Skinner 2024; Zhang et al. 2024b; Wang et al. 2024). These works are limited to specific scenarios or focus on distillation for individual modalities. (Xue et al. 2022) introduced the Modality Focusing

Hypothesis (MFH), offering the first theoretical analysis of CMKD's effectiveness, highlighting modality-generic decisive features as crucial for knowledge transfer. More recently, (Huo et al. 2024) identified modality imbalance and soft label misalignment as major challenges for CMKD, and introduced the C2KD framework, which significantly improved performance through bidirectional distillation and dynamic selection. The works fall short in addressing the inconsistencies in specific modality information and fail to fully leverage modality-generic features for effective cross-modal transfer. Our work overcomes these limitations by introducing a frequency-domain feature decoupling approach.

How Cross-Modal Features Differ?

The major difference between conventional KD in single modality and cross-modal knowledge distillation (CMKD) is that, CMKD is designed to distill the knowledge from another different modality. This difference poses a significant challenge since the teacher and student are trained with data in different modalities, and therefore have more distinct feature representations. Therefore, to design our cross-modal feature distillation method, it is necessary to first analyze the difference of cross-modal features. In this section, we present two of our major findings: (i) modality-generic and modality-specific features act differently in frequency domain; (ii) The features of different modalities exhibit significant differences in scale.

Decoupling Modality-specific and Modality-generic Features

Feature vectors learned by deep neural networks are not arbitrary arrangements, but rather encode learned, structurally meaningful information. Features originating from a single modality typically encode both modality-shared generic semantic information and modality-specific idiosyncratic details concurrently (Ngiam et al. 2011). Therefore, for improved knowledge transfer in Cross-Modal Knowledge Distillation (CMKD), the effective disentanglement of these two types of features is paramount.

Inspired by frequency-domain decomposition principles in signal processing (Bruna and Mallat 2013), we posit that the frequency composition of deep feature vectors naturally aligns with this semantic-detail hierarchy. Specifically, low-frequency components tend to capture smooth, macroscopic trends and patterns within feature vectors, which we hypothesize correspond to universal cross-modal semantic representations. Conversely, high-frequency components capture abrupt, fine-grained variations across feature vector dimensions, which we hypothesize primarily stem from modality-specific details or noise.

To validate the aforementioned hypothesis, we performed frequency-domain decomposition on feature vectors extracted from various modalities and quantified the inter-modal similarity of their resulting low- and high-frequency features. Cosine similarity was employed as the quantitative measure. As depicted in Table 1, low-frequency features exhibited significantly high inter-modal similarity, even surpassing that of the raw features, whereas high-frequency features displayed extremely low inter-modal similarity. These results further corroborate our hypothesis and strongly suggest that for effective cross-modal knowledge distillation, features must be decomposed into low- and high-frequency components, which should then be processed differentially based on their distinct representational characteristics. We will formally present our proposed strategy in following section.

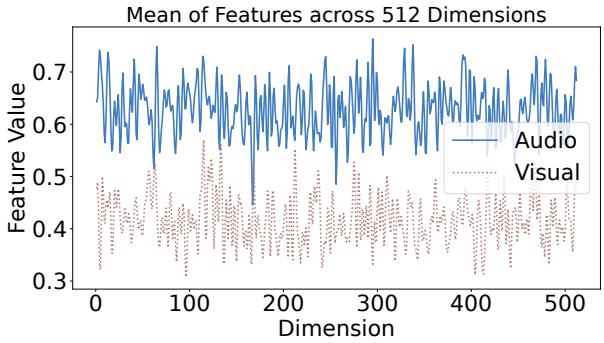


Figure 1: The comparison of feature mean value differences across modalities.

Scale Differences Across Modalities

In transfer learning, distribution differences play a crucial role in determining whether knowledge can be effectively transferred. If there are significant differences between the distributions of the source and target domains, the model may fail to capture useful information during the transfer process, leading to a substantial decrease in the effectiveness of the transfer.

We contend that inter-modal scale discrepancies constitute another contributing factor to suboptimal performance in cross-modal distillation. Through visualizing the feature means across 512 dimensions, it was revealed that significant scale differences exist among modalities. As shown in Figure 1, the feature means of the audio and visual modalities differ across dimensions, with the feature values of the audio modality being noticeably higher than those of the visual modality.

When MSE is used to force the alignment of the student model's features with those of the teacher model, the student's features may shift towards the mean of the teacher model's features. However, this may conflict with the optimal mean expected in the student's modality, leading to suboptimal performance in the student model.

Therefore, we should not directly use MSE loss to learn features from different modalities. Instead, we should design a loss function that respects the inherent scale differences between modalities to achieve more effective knowledge transfer.

Our Approach

As previously discussed, we found that modality-specific and modality-generic information can be effectively decoupled through frequency domain analysis, and there are sig-

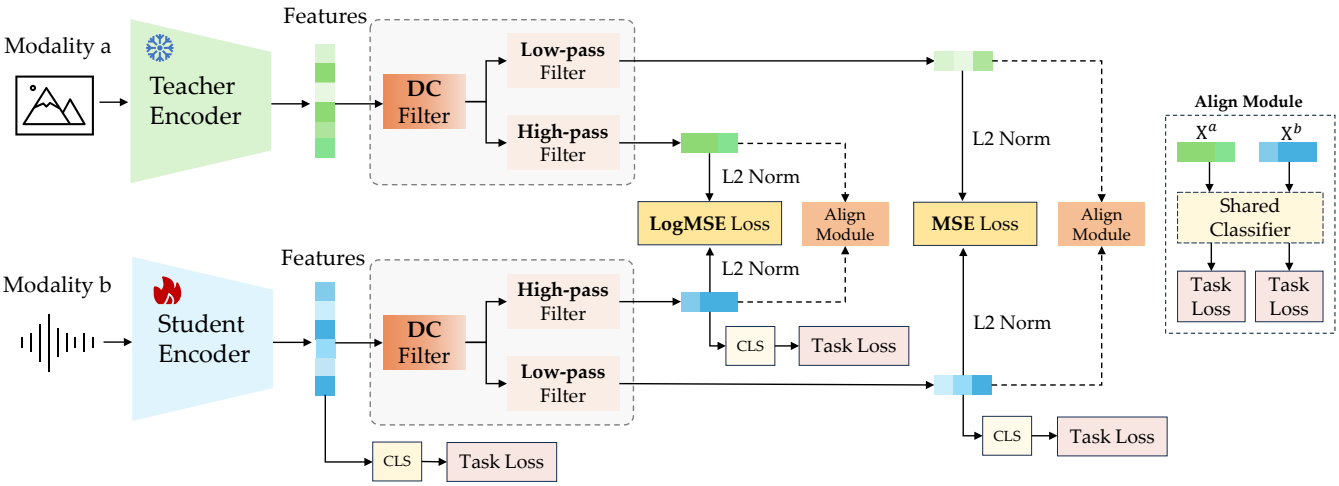


Figure 2: Framework of our method. We decouple the features of different modalities in the frequency domain into high-frequency and low-frequency components. For low-frequency features, MSE loss is applied, while logMSE loss is used for high-frequency features. Additionally, we ensure consistency in feature scale and feature space across modalities through feature normalization and alignment modules.

nificant differences in the feature distributions across different modalities. In this section, we will formally introduce our method to improve CMKD: (i) We decouple the features into low-frequency and high-frequency components and apply different loss functions for distillation accordingly. (ii) We ensure the features from different modalities are consistent in scale and feature space.

Frequency-Decoupled Distillation

We identified frequency decoupling of features as a effective way to disentangle the modality-generic and modality-specific information in the features. Formally, given the original feature $\mathbf{X}^m \in \mathbb{R}^D$ for a certain modality m, we compose the following three computation steps to decouple it into two features, namely low-frequency feature $\mathbf{X}^m_{\text{low}}$ and high-frequency feature $\mathbf{X}^m_{\text{high}}$.

Spatio-temporal domain to frequency domain. To decouple the original features, we first use Fourier transform to convert them into frequency domain, *i.e.*,

$$\mathbf{X}^m_f = \mathbf{DFT}(\mathbf{X}^m), \tag{1}$$

where $\mathbf{X}^m_f$ represents the corresponding Fourier-transformed feature in complex frequency domain.

High-pass and low-pass filtering. In the frequency domain, we decompose $\mathbf{X}^m_f$ into different frequency components by designing a low-pass filter $\mathbf{M}_{\text{low}}$ and a high-pass filter $\mathbf{M}_{\text{high}}$. $\mathbf{M}_{\text{low}}$ and $\mathbf{M}_{\text{high}}$ are fixed binary mask filters: $\mathbf{M}_{\text{low}}$ sets the first half of the frequency components to 1 (low-pass filter), and $\mathbf{M}_{\text{high}}$ sets the second half to 1 (high-pass filter). Then the low-frequency part $\mathbf{X}^m_{\text{low},f}$ and the high-frequency part $\mathbf{X}^m_{\text{high},f}$ are computed as follows:

$$\mathbf{X}^m_{f,\text{low}} = \mathbf{X}^m_f \cdot \mathbf{M}_{\text{low}}, \quad \mathbf{X}^m_{f,\text{high}} = \mathbf{X}^m_f \cdot \mathbf{M}_{\text{high}}. \tag{2}$$

Feature reconstruction with inverse Fourier transform. To obtain the reconstructed low-frequency and high-frequency features, we apply the Inverse Discrete Fourier Transform (IDFT) to transform the low-frequency and high-frequency components from the frequency domain back to the spatio-temporal domain, then we can obtain the decoupled features as

$$\mathbf{X}^m_{\text{low}} = \mathbf{IDFT}(\mathbf{X}^m_{f,\text{low}}), \quad \mathbf{X}^m_{\text{high}} = \mathbf{IDFT}(\mathbf{X}^m_{f,\text{high}}). \tag{3}$$

The next task is to pinpoint the most suitable design of distillation loss for each type of features, respectively. As analyzed in previous section, low-frequency features primarily encompass modality-generic information, highly shared across different modalities. Hence, it is imperative to maintain "strong consistency" for low-frequency features across different modalities so that their generality can be guaranteed. On the other hand, high-frequency features tend to capture modality-specific fine-grained information and are often accompanied by more noises. To preserve modality-specific details whilst reducing sensitivity to large errors stemming from the noises, we only require "weak consistency" for high-frequency features across different modalities.

As a result, for the low-frequency features on two different modalities a and b, we use the conventional mean square error (MSE) as the loss function, *i.e.*,

$$\mathcal{L}_{\text{low}} = \frac{1}{ND} \left\| \mathbf{X}^a_{\text{low}} - \mathbf{X}^b_{\text{low}} \right\|^2, \tag{4}$$

where N and D denote the batch size and dimension, respectively.

While for the distillation of high-frequency features, a proper way is suppressing the significant gradient values caused by the noises and abnormally-large features. To this end, we leverage log mean square error (LogMSE) as the distillation loss, which has smoother gradients when the difference of two feature values is large, as shown in Figure 3. The distillation loss for high-frequency features is formulated as

$$\mathcal{L}_{\text{high}} = \frac{1}{ND} \left\| \sigma(\mathbf{X}^a_{\text{high}}) - \sigma(\mathbf{X}^b_{\text{high}}) \right\|^2 \tag{5}$$

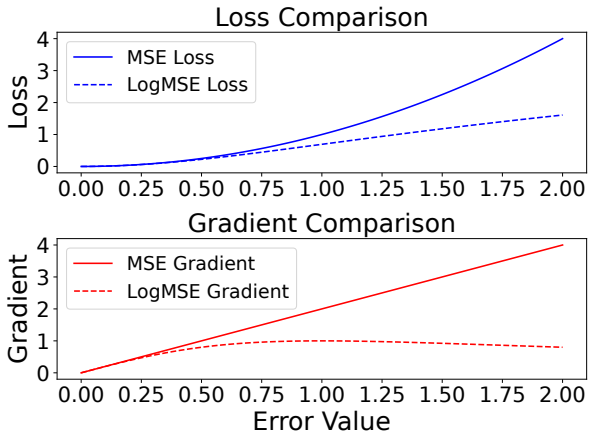


Figure 3: Comparison of value and gradient between MSE and logMSE losses.

$$\text{with} \quad \sigma(\mathbf{X}) = \begin{cases} \mathbf{log}(1+\mathbf{X}), & \mathbf{X} \geq 0 \\ -\mathbf{log}(1-\mathbf{X}), & \mathbf{X} < 0 \end{cases}, \tag{6}$$

where N and D denotes the batch size and dimension, respectively.

Alignment of Feature Scale and Feature Space

The consistency of feature distributions is pivotal for knowledge transfer. However, the significant differences in feature distributions between different modalities result in poor performance in cross-modal knowledge transfer. To mitigate the distribution discrepancies between modalities, we propose solutions from both the feature scale and feature space perspectives.

Feature scale alignment. The inconsistency in feature scales is typically reflected in the fact that feature vectors from different modalities may have varying numerical ranges, which can negatively impact the effectiveness of knowledge distillation. To achieve feature scale alignment, we employed a Feature Standardization strategy, which includes the following steps:

1. Mean Subtraction: First, mean subtraction is applied to the feature vectors to ensure that the mean of the features is zero, eliminating any bias in the features.
2. L2 Normalization: Next, L2 normalization is performed on the zero-centered feature vectors to ensure that the L2 norm of each feature vector is 1. This ensures that all feature vectors are compared on the same scale, avoiding computational biases caused by differences in the lengths of the feature vectors.

Herein is the formula for feature standardization:

$$\mathbf{Std}(\mathbf{X}) = \frac{\mathbf{X} - \bar{\mathbf{X}}}{\|\mathbf{X} - \bar{\mathbf{X}}\|_2}, \tag{7}$$

where $\mathbf{X}$ represents the input feature vector, $\bar{\mathbf{X}}$ represents the mean of the features, and $\|\cdot\|_2$ denotes the L2 norm. In practice, the mean subtraction operation can be directly implemented by using a DC filter in the frequency domain (as shown in Figure 2). By doing so, the previous distillation losses in Eq. 4 and Eq. 5 can be reformulated as follows:

$$\mathcal{L}_{\text{low}} = \frac{1}{ND} \left\| \mathbf{Std}(\mathbf{X}^a_{\text{low}}) - \mathbf{Std}(\mathbf{X}^b_{\text{low}}) \right\|^2, \tag{8}$$

$$\mathcal{L}_{\text{high}} = \frac{1}{ND} \left\| \sigma(\mathbf{Std}(\mathbf{X}^a_{\text{high}})) - \sigma(\mathbf{Std}(\mathbf{X}^b_{\text{high}})) \right\|^2. \tag{9}$$

Feature space alignment. Although feature scale alignment can alleviate the inconsistency in the numerical ranges of features from different modalities, solely relying on scale alignment is insufficient to address the fundamental differences in feature distributions across modalities. Features from different modalities not only differ in numerical scales but may also exhibit significant variations in the specific shapes of their distributions and the delineation of class boundaries.

To further enhance the effective transfer of cross-modal knowledge, we propose an alignment strategy from the perspective of feature space, ensuring that the features of the teacher model and the student model are comparable within the same space, thereby narrowing the distribution differences between modalities.

As shown in Figure 2, we designed an alignment module based on a shared classifier to achieve feature space alignment. Through the shared classifier, the features of both the teacher model and the student model can be aligned within the same decision space, thus reducing the distribution differences between modalities. Specifically, the features from the teacher model and the student model are fed into the same shared classifier, where they are classified through the shared classifier, and the classification alignment loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{align}} = \mathbf{CE}(\mathbf{\Phi}_h(\mathbf{X}^a_{\text{high}}), y) + \mathbf{CE}(\mathbf{\Phi}_h(\mathbf{X}^b_{\text{high}}), y) \\ + \mathbf{CE}(\Phi_l(\mathbf{X}^a_{\text{low}}), y) + \mathbf{CE}(\Phi_l(\mathbf{X}^b_{\text{low}}), y), \end{aligned} \tag{10}$$

where $\mathbf{CE}(\cdot)$ denotes the cross-entropy loss, $\mathbf{\Phi}_h$ and $\mathbf{\Phi}_l$ represent the shared classifiers for high-frequency and low-frequency features, respectively, and y denotes the ground truth labels.

Overall loss function. In addition to the aforementioned losses, we also compute the cross-entropy loss on the raw features, low-frequency features, and high-frequency features of the student model, respectively, to ensure that these features are discriminative (Mao, Mohri, and Zhong 2023). We denote this loss as $\mathcal{L}_{\text{task}}$. See Figure 2 for an easier reference of all the losses we conduct. As a result, the total loss function can be expressed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}} + \lambda_1 \mathcal{L}_{\text{low}} + \lambda_2 \mathcal{L}_{\text{high}}, \tag{11}$$

where λ_1 and λ_2 represent the weighting parameters for the distillation losses of low-frequency and high-frequency features, respectively.

Experiments

We evaluate our method on classification and semantic segmentation tasks across various multimodal datasets. We provide experimental settings before detailing the result analysis.

Category	Method	CREMA-D		AVE		VGGSound		CrisisMMD	
		A	V	A	V	A	V	T	V
Uni-Modal	w/o KD	62.4	66.8	63.7	38.8	68.9	44.9	77.4	70.2
Logits	Logit	61.7	62.6	60.0	39.1	65.7	45.4	78.5	70.5
	DIST	62.2	64.0	62.4	40.3	66.4	45.5	78.3	71.3
	DML	52.7	61.2	60.2	43.3	57.8	44.9	78.2	71.2
	NKD	62.4	61.8	60.7	38.1	65.6	44.9	78.1	71.2
	DKD	61.0	61.4	60.5	38.1	64.4	44.5	79.0	70.7
Feature	Feat	60.9	64.3	58.7	39.6	67.7	45.5	77.7	70.8
	PKD	60.4	64.8	58.0	41.0	62.9	46.9	77.5	70.9
	OFD	60.6	61.6	58.0	39.6	68.5	45.8	78.1	71.2
	AFD	61.2	59.5	62.7	38.8	68.7	45.8	69.8	72.3
Cross-Modal	C2KD	57.5	59.8	62.7	39.3	67.0	47.9	77.9	71.4
	Ours	**64.1**	**71.0**	**64.9**	**47.8**	**70.0**	**48.1**	**79.1**	**72.7**

Table 2: The comparison of methods on Audio-Visual and Image-Text classification tasks. The metric is the top-1 accuracy(%). 'A', 'V', and 'T' represent Audio, Visual, and Text modalities, respectively. "Uni" refers to unimodal models without distillation. "Logit" and "Feat" correspond to the original logit-based and feature-based distillation methods. "C2KD" represents the cross-modal distillation method mentioned in (Huo et al. 2024). The best is in **bold**, and the second is underlined.

Method	Uni	Logit	DIST	DKD	Feat	PKD	AFD	C2KD	Ours
Depth	30.9	29.7	32.3	32.5	29.4	31.0	30.2	31.8	**33.2**
RGB	34.1	32.8	34.9	35.3	32.8	33.7	32.7	34.8	**36.9**

Table 3: The comparison on the semantic segmentation task. The metric denotes the mean Intersection over Union (mIoU).

Classification Task

Dataset. CREMA-D (Cao et al. 2014) is an emotion recognition dataset with audio and vision, featuring six emotions: happy, sad, angry, fear, disgust, neutral. AVE (Tian et al. 2018) is an audio-visual event localization dataset with $4,143$ videos across 28 event categories. While VGGSound (Chen et al. 2020) is a large-scale audio-visual dataset with 210K ten-second videos, a subset of 50 categories for our experiments. CrisisMMD (Alam, Ofli, and Imran 2018) is a multimodal dataset for natural disaster research, including annotated tweets and images from Twitter in image and text formats. For more detailed information about the dataset, please refer to Appendix.

Experimental Settings. Our experimental settings follow (Huo et al. 2024; Fan et al. 2024; Wei et al. 2024). We use the ResNet-18 (He et al. 2016) as the backbone for audio-visual datasets and train them for 100 epochs in total. In the CrisisMMD dataset, we employ BERT-base (Devlin 2018) and MobileNetV2 (Sandler et al. 2018) to extract text and visual features, respectively. We only train text modality for 20 epochs. We utilize the SGD optimizer with a momentum of 0.9, and the batch size for training is set to 64. For detailed training information, see Appendix.

Results Analysis. In Table 2, we present the performance of our method on classification benchmarks. We compare the logit-based, feature-based, and cross-modal state-of-the-art distillation methods. that our proposed method consistently achieves the best performance across all datasets and modalities. For example, on the AVE dataset's visual modality, our method improves performance by 9%, reaching 47.8%, compared to the unimodal baseline. This highlights the effectiveness of our approach in transferring knowledge across modalities. Notably, our method excels in transferring knowledge from low-performing modalities to high-performing ones, where other methods fail. For instance, on CREMA-D's visual modality, AVE's audio modality, and VGGSound's audio modality, most methods underperform compared to the unimodal baseline, while our approach consistently improves performance by effectively transferring knowledge from weaker modalities. Additionally, our method is stable in bidirectional cross-modal transfer. On CrisisMMD, while DKD works well for text but not visual, and AFD succeeds for visual but fails for text, our method performs consistently across both modalities, achieving 79.1% on text and 72.7% on visual. This outstanding performance is attributed to our method's ability to capture both modality-specific and modality-agnostic information through frequency decomposition and customized loss functions, as well as mitigating inherent feature distribution differences through feature alignment. This ensures robust results across various modality pairs (A-V, T-V) and network architectures (ResNet-ResNet, BERT-MobileNet).

Semantic Segmentation Task

Dataset. NYU-Depth V2 (Wofk et al. 2019) is a multimodal dataset for indoor scene understanding research. It provides two modalities of depth information and RGB image information. There are a total of 40 categories. It contains $1,449$ densely labeled RGB and depth image alignment pairs.

Experimental Settings. Following C2KD (Huo et al. 2024), the DeepLab V3+ (Chen et al. 2018) model is utilized with ResNet-18 as the backbone, which is initialized with the pre-trained weights on ImageNet (Deng et al. 2009). We train the

Method				CREMAD		AVE	
Freq	Align	Scale	Log	A	V	A	V
				60.9	64.3	58.7	39.6
✓				60.8	68.7	61.0	43.3
	✓			60.9	67.9	63.2	41.3
✓	✓			61.8	68.7	62.4	45.8
✓		✓		62.2	70.0	62.4	44.8
✓	✓	✓		62.2	70.6	62.4	46.0
✓	✓	✓	✓	64.1	71.0	64.9	47.8

Table 4: Ablation study of our components. Freq: frequency decomposition; Align: feature space alignment; Scale: feature standardization; Log: logMSE loss on high-frequency features; Baseline: original feature distillation.

student for 50 epochs in total and the batch size is 16.

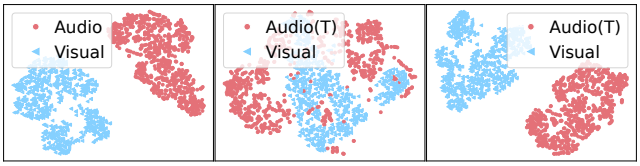
Results Analysis. Regarding segmentation task, Table 3 shows the performance of various KD methods on NYU-Depth V2. Our method still consistently outperforms all other methods, with 33.2% mIoU for Depth and 36.9% mIoU for RGB. These results surpass the next best method (DIST for Depth, DKD for RGB) by a notable margin of 0.9% and 1.6%, respectively. As highlighted earlier in the classification tasks, our method is also stable in bidirectional cross-modal transfer in segmentation tasks.

Analysis

In this section, we first evaluate the effectiveness of the key components in our CMKD method, including frequency decomposition, feature alignment, and loss functions, through ablation studies. Then, we visualize the feature distributions using t-SNE, showcasing the improved feature separation and cross-modal knowledge transfer achieved by our proposed method, when compared with traditional techniques. For further analysis, please refer to the Appendix.

Effectiveness of components in CMKD

We perform experiments to show the effectiveness of each proposed component in CMKD in Table 4. Firstly, it is evident that each individual component contributes positively to the overall performance. Frequency decomposition distillation provides improvements for most modalities as it helps to separate modality-specific information from modality-generic information. However, these improvements are not always consistent; for example, on the audio (A) modality of the CREMA-D dataset, there is a 0.1% performance drop. This inconsistency may stem from significant differences in feature distributions across modalities. When we add the Feature space alignment and Feature standardization modules, the cross-modal performance improves significantly, highlighting the importance of reducing feature distribution discrepancies between modalities. Moreover, applying logMSE loss to high-frequency features enhances the transmission of modality-specific information, indicating that it is not necessary to fully align modality-specific information, and maintaining a weak consistency is more effective for transferring such information. Finally, the comprehensive integration of all components ensures more robust cross-modal knowledge transfer, thereby achieving more stable performance across different modalities.



(a) Without KD (b) Feat KD (c) Ours

Figure 4: t-SNE visualization comparison between the conventional feature distillation method and our proposed approach. We visualize the features of different modalities on the CREMAD test set. T represents the teacher modality.

Visualization

In Figure 4, we present a t-SNE (Van der Maaten and Hinton 2008) visualization to compare the performance of the original feature distillation method and our proposed approach in cross-modal knowledge distillation. Figure 4a shows the result without any distillation, where the feature distributions of different modalities are clearly distinguishable. Figure 4b illustrates the result of traditional feature distillation, where there is significant overlap between the features of the visual modality (teacher) and audio modality (student). This indicates that the method fails to effectively differentiate between modality-specific features, leading to insufficient retention of modality-specific information, disrupting the original distribution of the student's features, and reducing the student model's discriminative ability. In contrast, Figure 4c demonstrates the visualization of our approach, where the features from different modalities are clearly separated, forming two distinct clusters. This separation suggests that our method successfully disentangles modality-generic and modality-specific information, improving feature discrimination. These results validate our frequency decomposition strategy, which preserves modality-specific characteristics while enhancing cross-modal knowledge transfer.

Conclusion

In this paper, we investigate the non-negligible challenges faced by cross-modal knowledge distillation, particularly focusing on the discrepancies between modality-specific and modality-generic information, and the differences in feature distributions across modalities. Based on the observation and analysis of the failure of feature distillation in cross-modal scenarios, we propose a novel distillation framework. It decouples these types of information through frequency-based feature analysis and introduces a differentiated distillation strategy for different frequency components. Additionally, we address feature distribution discrepancies by incorporating a scale consistency loss and using a shared classifier for feature space alignment. Comprehensive experiments demonstrate the effectiveness of our approach.

Acknowledgments

We would very much like to thank the anonymous reviewers for their valuable comments. This work is supported in part by National Key R&D Program of China (Grant No. 2024YFB4505901), in part by the National Natural Science Foundation of China (Grant No. 62402024), in part by the Beijing Natural Science Foundation (Grant No. L241050), and in part by the Fundamental Research Funds for the Central Universities. For any correspondence, please refer to Dr. Renyu Yang (renyuyang@buaa.edu.cn).

References

Afouras, T.; Chung, J. S.; and Zisserman, A. 2020. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2143–2147. IEEE.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Bruna, J.; and Mallat, S. 2013. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1872–1886.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390.

Cao, J.; Zhang, Y.; Huang, T.; Lu, M.; Zhang, Q.; An, R.; Ma, N.; and Zhang, S. 2025. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Cao, W.; Zhang, Y.; Gao, J.; Cheng, A.; Cheng, K.; and Cheng, J. 2022. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35: 15394–15406.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5008–5017.

Dai, R.; Das, S.; and Bremond, F. 2021. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13053–13064.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fan, Y.; Xu, W.; Wang, H.; Liu, J.; and Guo, S. 2024. Detached and Interactive Multimodal Learning. *arXiv preprint arXiv:2407.19514*.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1921–1930.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022a. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Huang, T.; Zhang, Y.; You, S.; Wang, F.; Qian, C.; Cao, J.; and Xu, C. 2022b. Masked distillation with receptive tokens. *International conference on learning representations*.

Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16006–16015.

Lee, P.; Kim, T.; Shim, M.; Wee, D.; and Byun, H. 2023. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2373–2383.

Li, X.; Xiong, H.; Wang, H.; Rao, Y.; Liu, L.; Chen, Z.; and Huan, J. 2019. Delta: Deep learning transfer using feature map with attention for convolutional networks. *arXiv preprint arXiv:1901.09229*.

Liu, Y.; Jia, Z.; and Wang, H. 2023. Emotionkd: a cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6122–6131.

Mao, A.; Mohri, M.; and Zhong, Y. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, 23803–23828. PMLR.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.

Pham, C.; Nguyen, V.-A.; Le, T.; Phung, D.; Carneiro, G.; and Do, T.-T. 2024. Frequency attention for knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2277–2286.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Sun, B.; and Saenko, K. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, 443–450. Springer.

Thoker, F. M.; and Gall, J. 2019. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, 6–10. IEEE.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Ma, C.; Zhang, J.; Zhang, Y.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 216–226. Springer.

Wang, S.; She, R.; Kang, Q.; Jian, X.; Zhao, K.; Song, Y.; and Tay, W. P. 2024. DistilVPR: Cross-Modal Knowledge Distillation for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10377–10385.

Wei, Y.; Feng, R.; Wang, Z.; and Hu, D. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27338–27347.

Williams, T.; and Li, R. 2018. Wavelet pooling for convolutional neural networks. In *International conference on learning representations*.

Wofk, D.; Ma, F.; Yang, T.-J.; Karaman, S.; and Sze, V. 2019. FastDepth: Fast Monocular Depth Estimation on Embedded Systems. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; and Ren, F. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1740–1749.

Xue, Z.; Gao, Z.; Ren, S.; and Zhao, H. 2022. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*.

Zhang, Y.; Chen, W.; Lu, Y.; Huang, T.; Sun, X.; and Cao, J. 2023. Avatar knowledge distillation: self-ensemble teacher paradigm with uncertainty. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5272–5280.

Zhang, Y.; Fan, C.-K.; Huang, T.; Lu, M.; Yu, S.; Pan, J.; Cheng, K.; She, Q.; and Zhang, S. 2025. AutoV: Learning to Retrieve Visual Prompt for Large Vision-Language Models. *arXiv preprint arXiv:2506.16112*.

Zhang, Y.; Huang, T.; Liu, J.; Jiang, T.; Cheng, K.; and Zhang, S. 2024a. FreeKD: Knowledge Distillation via Semantic Frequency Prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15931–15940.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4320–4328.

Zhang, Y.; Xiao, F.; Huang, T.; Fan, C.-K.; Dong, H.; Li, J.; Wang, J.; Cheng, K.; Zhang, S.; and Guo, H. 2024b. Unveiling the tapestry of consistency in large vision-language models. *Advances in Neural Information Processing Systems*.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 11943–11952. IEEE.

Zhao, L.; Song, J.; and Skinner, K. A. 2024. CRKD: Enhanced Camera-Radar Object Detection with Cross-modality Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15470–15480.