The Modality Focusing Hypothesis: On the Blink of Multimodal Knowledge Distillation

Zihui Xue\textsuperscript{*}  
UT Austin  
sherryxue@utexas.edu

Zhengqi Gao\textsuperscript{*}  
MIT  
zhengqi@mit.edu

Sucheng Ren\textsuperscript{*}  
South China University of Technology  
oliverrensu@gmail.com

Hang Zhao\textsuperscript{†}  
Tsinghua University, Shanghai Qi Zhi Institute  
hangzhao@mail.tsinghua.edu.cn

Abstract

Multimodal knowledge distillation (KD) extends traditional knowledge distillation to the area of multimodal learning. One common practice is to adopt a well-performed multimodal network as the teacher in the hope that it can transfer its full knowledge to a unimodal student for performance improvement. In this paper, we investigate the efficacy of multimodal KD. We begin by providing two failure cases of it and demonstrate that KD is not a universal cure in multimodal knowledge transfer. We present the modality Venn diagram to understand modality relationships and the modality focusing hypothesis revealing the decisive factor in the efficacy of multimodal KD. Experimental results on 6 multimodal datasets help justify our hypothesis, diagnose failure cases, and point directions to improve distillation performance.

1 Introduction

Knowledge distillation (KD) is an effective technique to transfer knowledge from one neural network to another [45, 12]. Its core mechanism is a teacher-student learning framework, where the student network is trained to mimic the teacher through a loss. The loss is realized as the KL divergence between teacher and student soft labels when first proposed by [17] and has been extended in many ways [49, 41, 31, 32, 39]. KD has been successfully applied to various fields and demonstrates its high practical value.

The wide applicability of KD stems from its generality: any student can learn from any teacher. To be more precise, the student and teacher network may differ in several ways. Three common scenarios are: (1) model capacity difference: Many works [49, 41, 31, 32] on model compression aim to learn a lightweight student matching the performance of its cumbersome teacher for deployment benefits. (2) architecture (inductive bias) difference: Recent works [40, 35, 47] propose to distill the inductive bias of a CNN teacher to train a transformer student for data efficiency. (3) modality difference: KD has been extended to the area of multimodal learning [13, 3, 51, 10, 38, 34, 1, 42, 48], where the teacher and student network have different input modalities (e.g., an audio teacher and a visual student).

Despite the great empirical success reported in prior works, the working mechanism of KD is still poorly understood [12]. This puts the efficacy of KD into question: Is KD always efficient? If not, what is a good indicator of KD performance? A few works [8, 37, 35] are in search for the answer in

\textsuperscript{*}Zihui, Zhengqi, and Sucheng contribute equally. Work done during internship at Shanghai Qi Zhi Institute.

\textsuperscript{†}Corresponding author: Hang Zhao.
the context of model capacity difference and architecture difference. However, the analysis for the third scenario, KD under modality difference or formally multimodal KD, remains an open problem. This work aims to fill this gap and for the first time provides a comprehensive analysis of multimodal KD. Our major contributions are the following:

- We present two failure cases of multimodal KD and find surprisingly that a more accurate multimodal teacher does not necessarily lead to a better student.
- To explore the cause of performance mismatch in multimodal KD, we propose the modality Venn diagram to understand modality relationships, where we define modality-general/specific decisive features.
- We present the modality focusing hypothesis that provides an explanation of when multimodal KD is effective. We hypothesize that modality-general decisive features are the crucial factor that determines the efficacy of multimodal KD. We further propose a permutation-based approach to identify modality-general decisive features for real multimodal data.
- We conduct experiments on 6 multimodal datasets (i.e., synthetic Gaussian, AV-MNIST, RAVDESS, VGGSound, NYU Depth and MM-IMDB) and demonstrate that our proposed hypothesis can guide multimodal KD and bring performance improvement.

2 Related Work

2.1 Unimodal Knowledge Distillation

KD represents a general technique that transfers information learned by a teacher network to a student network. It has been broadly applied to many vision tasks, such as image classification [41, 31, 32], semantic segmentation [16, 23] and video understanding [28, 30]. Model compression stands for one big application field of KD, where the goal is to obtain a small model (i.e., the student) with high inference efficiency through learning from a pretrained large model (i.e., the teacher). More recently, KD has been introduced to facilitate the training of a vision transformer [40, 35, 47]. The transformer student achieves high data efficiency with inductive biases distilled from CNN teachers. Despite the development towards better distillation techniques or new application fields, there is limited work [33, 8, 37, 35] on understanding the working mechanism of KD. Specifically, Cho et al. [8] and Mirzadeh et al. [27] investigate KD for model compression, i.e., when the student and teacher differ in model size. They point out that mismatched capacity between student and teacher network can lead to failure of KD. Ren et al. [35] provide an analysis of KD for vision transformers and demonstrate that teacher’s inductive bias matters more than its accuracy in improving performance of the transformer student. These works provide good insight into understanding KD, yet their discussion is limited to unimodality.

2.2 Multimodal Knowledge Distillation

With the accessibility of the Internet and the growing availability of multimodal sensors, multimodal learning has received increasing research attention [4]. Following this trend, KD has also been extended to achieve knowledge transfer from multimodal data and enjoys diverse applications, such as action recognition [10, 26, 38], lip reading [34, 1] and medical image segmentation [18, 21]. Vision models are often adopted as teachers to provide supervision to student models of other modalities, e.g., sound [3, 48], depth [13, 48], optical flow [10], thermal [19], wireless [51], etc. While these works demonstrate potentials of multimodal KD, they are often associated with a specific multimodal task and mostly empirical. An in-depth analysis of multimodal KD is notably lacking, which is the main focus of this paper.

3 On the Efficacy of Multimodal Knowledge Distillation

3.1 Problem Formulation

Consider a supervised $K$-class classification problem. We denote the input of the student and teacher networks by $x_S$ and $x_T$, respectively. We use $f_{\theta_S}(x_S) \in \mathbb{R}^K$ and $f_{\theta_T}(x_T) \in \mathbb{R}^K$ to respectively represent the output (i.e., class probabilities) of the student and teacher networks, where $\{\theta_S, \theta_T\}$
are learnable parameters. Without loss of generality, we limit our discussion within input data of two modalities, denoted by \( x_1 \) and \( x_2 \) for modality 1 and 2, respectively. Assume that we aim to learn a student network that takes \( x_1 \) as input. There are three typical ways of distillation: (i) unimodal KD: the student and teacher network receive input from the same modality (e.g., a visual teacher distills knowledge to a visual student), (ii) multimodal KD: we adopt a multimodal teacher for distillation (e.g., an audio-visual teacher and a visual student), and (iii) crossmodal KD: the teacher and student network receive input from distinct modalities (e.g., an audio teacher and a visual student).

**Unimodal KD.** \((x_T \leftarrow x_1, x_S \leftarrow x_1)\) Both the teacher and student take unimodal data \( x_1 \) as input. With an available teacher model, the training of student is achieved by minimizing:

\[
L = \rho L_{task} + (1 - \rho) L_{kd}
\]

\[
L_{task} = CE(y, f_{\theta_S}(x_1)) \quad \text{and} \quad L_{kd} = KL(f_{\theta_T}(x_1), f_{\theta_S}(x_1))
\]

where \( y \) represents the ground truth label, \( \rho \in [0, 1] \) weighs the importance of two terms \( L_{task} \) and \( L_{kd} \) (i.e., driving the student to true labels or teacher’s soft predictions), respectively. ‘CE’ denotes cross entropy, and ‘KL’ represents KL divergence.

**Multimodal KD.** \((x_T \leftarrow (x_1, x_2), x_S \leftarrow x_1)\) Multimodal KD resorts to a multimodal teacher in the distillation. The teacher takes multimodal data as input while the student input remains the same as in unimodal KD (i.e., \( x_1 \)). We still minimize Eq. (1) to train a student model, but now \( L_{kd} \) is given by:

\[
L_{kd} = KL(f_{\theta_T}(x_1, x_2), f_{\theta_S}(x_1))
\]

**Crossmodal KD.** \((x_T \leftarrow x_2, x_S \leftarrow x_1)\) Crossmodal KD refers to the case where teacher and student input modalities do not overlap. It could be regarded as a special case of multimodal KD when the teacher only takes input from modality 2. Eq. (1) is still valid with a slight correction to \( L_{kd} \):

\[
L_{kd} = KL(f_{\theta_T}(x_2), f_{\theta_S}(x_1))
\]

### 3.2 Is a Multimodal Teacher Better than a Unimodal One?

Benefiting from the great representation power of multimodal data, a multimodal network usually enjoys a higher accuracy than its unimodal counterpart [4]. This motivates many research works [26, 18, 42] to replace conventional unimodal KD with multimodal KD, in an attempt to improve student network performance. Their underlying motivation is intuitive: a more accurate teacher can lead to a better student. Also, the complementary modality-dependent information brought by the teacher network may further enhance student performance. However, in this paper, we reflect on this assumption and ask the question: *Is multimodal KD always more effective than unimodal KD?*

| Table 1: Evaluation of unimodal KD (UM-KD) and multimodal KD (MM-KD) on AV-MNIST and NYU Depth V2. ‘Mod.’ is short for modality, and A, I, RGB, D represents audio, grayscale images, RGB images and depth images, respectively. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Teacher Mod.** | **Student Mod.** | **Teacher Mod.** | **Student Mod.** |
| **AV-MNIST**    | **NYU Depth V2** |
| No-KD | - | - | A | 68.36 | - | - | RGB | 46.36 |
| UM-KD | A | 84.57 | A | 70.10 | RGB | 46.36 | RGB | 48.00 |
| MM-KD | I + A | 91.61 (↑) | A | 69.73 (↓) | RGB + D | 51.00 (↑) | RGB | 47.78 (↓) |

**Failure Cases.** We provide two example failure cases of multimodal KD in Table 1. AV-MNIST [43] is an audio-visual dataset containing hand-written digits (i.e., grayscale images) and human speaking digits (i.e., audio). Here we attempt to improve an audio student using KD. NYU Depth V2 [29] contains 1,449 pairs of aligned RGB and depth images. Similarly, the goal is to improve an RGB model performance with KD. Table 1 reports a somewhat counter-intuitive finding: a more accurate multimodal network does not serve as a good teacher. Despite the great increase in teacher accuracy, multimodal KD fails to outperform unimodal KD. In the case of AV-MNIST, while the audio-visual teacher itself has a much higher accuracy than the unimodal teacher (i.e., +7.04%), the
resulting student is worse (i.e., -0.37%) instead. Similarly, the great increase in teacher performance (i.e., +4.64%) does not translate to student improvement (i.e., -0.22%) on NYU Depth V2.

These results cast doubt on the efficacy of multimodal KD. Despite being more accurate, a multimodal network doesn’t always teach a better student. Thus, contradictory to the previous intuition, teacher performance seems not reflective of student performance. Inspired by this observation, our work targets on exploring the open problem: What is the fundamental factor deciding the efficacy of multimodal KD?

4 The Modality Focusing Hypothesis

4.1 Modality Venn Diagram

To study multimodal KD, it is critical to first establish an understanding of multimodal data. But before touching on multimodal data, let us first fall back and consider unimodal data. Following a causal perspective [36, 25] (i.e., features cause labels), we assume that the label $y$ is determined by a subset of features in $x_1$ (or $x_2$); this subset of features are referred to as decisive features for modality 1 (or modality 2) throughout the paper. For instance, colors of an image help identify some classes (e.g., distinguish between a zebra and a horse) and can be considered as decisive features.

When considering multimodal data, input features of the two modalities will have logical relations such as intersection and union. To capture this relationship, we propose the modality Venn diagram in Fig. 1. The diagram reflects the common perception that multimodal data possess shared information and preserve information specific to each modality as well [44, 50, 14]. Decisive features of the two modalities are thus composed of two parts: (1) modality-general decisive features (colored in dark purple) and (2) modality-specific decisive features (colored in light purple). Consider a video-audio data pair, as shown in Fig. 1, where the camera only captures one person due to its position angle and the audio is mixed sounds of two instruments. The true label “playing instruments” can be correctly inferred with three sources of information: visual specific decisive features, audio specific decisive features and modality-general decisive features. Fig. 1 left illustrates how we interpret these three sources of information at the input level.

To quantify the relative importance of these three features, we define $\gamma$, $\alpha_1$ and $\alpha_2$ as follows. Let $\gamma$ be the relative importance of general decisive features among all decisive features. $\alpha_1$ and $\alpha_2$ denote the relative importance of specific decisive features in modality 1 and 2, respectively. There are multiple ways to calculate the relative importance, and we discrete input features by assigning a binary label for each feature dimension (i.e., this dimension belongs to general decisive features or specific decisive features). The ratio of general decisive feature channel number to decisive feature channel number approximates $\gamma$, $\alpha_1$ and $\alpha_2$ are calculated in a similar way, and we have $\alpha_1 + \alpha_2 + \gamma = 1$.

Now we are ready to consider crossmodal KD (i.e., $x_T \leftarrow x_2$, $x_S \leftarrow x_1$). Clearly, teacher performance is controlled by both general decisive features and specific decisive features in modality 2. However, we find that: (i) Although specific decisive features in modality 2 are meaningful for the teacher, they can not instruct the student since the student only sees modality 1; (ii) On the other hand, general decisive features are not specific to modality 2 and could be transferred to the student. This motivates our modality focusing hypothesis.

The Modality Focusing Hypothesis. In multimodal KD and crossmodal KD (defined in Sec. 3.1), distillation performance is dependent on general decisive features in the teacher network: with larger $\gamma$, the student network is expected to perform better.

The hypothesis states that in multimodal knowledge transfer, the student learns to “focus on” general decisive features. Multimodal KD is thus beneficial for the case where $\gamma$ is large (i.e., multimodal data share many label-relevant information). Moreover, this poses a valid explanation for the observation in Sec. 3.2. Since teacher performance is influenced by decisive features while student performance is only related to general decisive features, it is plausible that teacher performance fails to correlate with student performance in some scenarios. For instance, when $\alpha_2$ is large and $\gamma$ is small, the teacher network attains high accuracy primarily based on modality-specific information, which is not visible to the student. In such case, we may find that a high-accuracy teacher does not lead to a good student.

To have an intuitive and quick understanding of our hypothesis, here we show an experiment with synthetic Gaussian data. More details can be found in Sec. 5.1. As shown in Fig. 2, we start from the
Figure 1: Left: an input video-audio pair can be regarded as composed of modality general features and modality-specific features in the visual and audio modality. For instance, the man playing violin in the right is not captured by the camera and hence its sound (colored in pink) belongs to audio modality-specific information. Those three parts of features work together and contribute to the final label $y$. Right: the modality Venn diagram. For any given pair of multimodal features (input-level, middle-level, output-level), there exist general decisive features (colored in dark purple) and specific decisive features (colored in light purple).

Figure 2: An illustration of the modality focusing hypothesis with synthetic Gaussian data. Teacher modality is $x_2$ and student modality is $x_1$. We plot the confidence interval of one standard deviation for student accuracy. With increasing $\gamma$, crossmodal KD becomes more effective. In the extreme case where two modalities do not overlap, and gradually increase the proportion of general decisive features until all decisive features are shared by two modalities. We observe that crossmodal KD fails to work when $x_1$ and $x_2$ share few decisive features (i.e., $\gamma$ is small) as specific decisive features in modality 2 are not perceived by the student. As $\gamma$ gradually increases, crossmodal KD becomes more effective. For the case where all decisive features possess in both modalities, the student benefits from teacher’s knowledge on general decisive features and outperforms its baseline by 2.1%. Note that the teacher accuracy does not vary much during this process, yet student performance differs greatly. This further lends support to our hypothesis and demonstrates that teacher accuracy does not faithfully reflect the effectiveness of crossmodal KD.

The modality focusing hypothesis offers a new perspective to understand multimodal KD. Below we describe one important implication to showcase its practical value.

**Implication.** For multimodal KD (crossmodal KD), consider two teachers with identical architectures: Teacher (a) makes predictions primarily based on general decisive features while teacher (b) relies more on modality-specific information. In other words, teacher (a) has a larger $\gamma$ than teacher (b). We expect that the student taught by teacher (a) yields better performance than that by teacher (b).

By drawing inspiration from the implication, we can train a teacher network that “focuses on” general decisive features for prediction. Compared with a regularly-trained teacher, the new teacher is tailored for distillation task and can possibly lead to a better student. The modality focusing hypothesis equips us with a new type of approaches to improve multimodal KD.
4.2 Modality-General Decisive Feature Separation

In the synthetic Gaussian case, we set up the experiment by manually controlling general decisive features. However, in real multimodal datasets, modality-general decisive features are unknown and need to be identified. Here we propose a permutation-based approach to address this problem.

The major steps of our proposed method are demonstrated in Algorithm 1. The input of Algorithm 1 are \( X_1 \in \mathbb{R}^{N \times D_1}, X_2 \in \mathbb{R}^{N \times D_2}, \) and \( Y \in \mathbb{R}^N \), representing \( N \) paired features from modality 1 and 2, and \( N \) target labels, respectively. The output is a saliency vector \( p \in [0, 1]^{D_2} \) for general decisive features in modality 2, where its \( d \)-th entry \( p_d \in [0, 1] \) reflects the saliency of the \( d \)-th feature dimension. A larger saliency value indicates a more general decisive feature channel.

### Algorithm 1 Modality-General Decisive Feature Separation

**Input:** multimodal data \( (X_1 \in \mathbb{R}^{N \times D_1}, X_2 \in \mathbb{R}^{N \times D_2}, Y \in \mathbb{R}^N) \)

**Output:** saliency vector \( p \in [0, 1]^{D_2} \) for features of modality 2

1. Jointly train two unimodal networks \( f_{\theta_1} \) and \( f_{\theta_2} \) using the following loss:

\[
\min_{\theta_1, \theta_2} L = Dist(f_{\theta_1}(X_1), f_{\theta_2}(X_2)) + CE(Y, f_{\theta_1}(X_1)) + CE(Y, f_{\theta_2}(X_2))
\]  

\( \text{\( Dist(\cdot, \cdot) \)} \) denotes a distance loss (e.g., mean squared error)

2. for \( d = 1 \) to \( D_2 \) do

3. \( p_d = 0 \)

4. for \( k = 1 \) to \( K \) do

5. permute the \( d \)-th column of \( X_2 \) yielding \( \tilde{X}_2 \)

6. \( p_d = p_d + \frac{1}{K} \times Dist(f_{\theta_1}(X_1), f_{\theta_2}(\tilde{X}_2)) \)

7. end for

8. end for

9. Perform normalization: \( p = \frac{P}{\max_{d \in [0, 1]^{D_2}} p_d} \)

**Clarifications.** Inspired by the fact that input-level features contain much label-irrelevant noise, as shown in the right of Fig. 1, our Algorithm 1 is designed following a trace-back thought starting from the output level. Namely, we drive two unimodal networks to the state of “feature alignment” at the output level using Eq. (5), and then use permutation to identify which input feature dimension has a larger impact to the state. Those more influential to the state (i.e., a large distance in step 6) will be assigned a larger saliency value.

In step 1, we jointly train two unimodal networks \( f_{\theta_1} \) and \( f_{\theta_2} \) that respectively take unimodal data \( X_1 \) and \( X_2 \) as input. The first loss term in Eq. (5) aims to align feature spaces learned by the two networks, and the remaining loss terms ensure that learned features are essential for a correct prediction. We believe that this training strategy aligns three sources of decisive features at the output level. In step 2, we follow the idea of permutation feature importance [5] to trace back general decisive features at the input level. For the \( d \)-th dimension in \( X_2 \), we randomly permute \( X_2 \) along this dimension and obtain a permuted \( \tilde{X}_2 \) in step 5. Next, we calculate the distance between \( f_{\theta_1}(X_1) \) and \( f_{\theta_2}(\tilde{X}_2) \) in step 6. A large distance indicates that the \( d \)-th dimension largely influences the state of “feature alignment”. Consequently, we are able to quantify the relative importance of each input feature channel and use the saliency vector \( p \) to represent it. We repeat the permutation process for \( K \) times and average the distance value for good stability. Finally, \( p \) is normalized to \([0, 1]^{D_2}\) in step 9.

**Remarks.** First, Algorithm 1 is not limited to feature separation at the input level. \( X_1 \) and \( X_2 \) can be features extracted from middle layers of the neural network as well. In such case, output \( p \) reflects the saliency of general decisive features for each middle-layer feature channel. Secondly, based on \( p \), we can decide whether each input dimension is “general decisive” or “non general decisive”. For instance, we can set a hard threshold or pick a certain ratio according to sorted \( p \). This provides us with a way to control \( \gamma \) for a given teacher network. In the experiments (Sec. 5), we nullify general decisive or non general decisive feature channels identified by Algorithm 1 to obtain teacher models with different \( \gamma \) to justify the hypothesis. Last but not least, Algorithm 1 could be equally applied to identify modality-general decisive features for modality 1 as long as we permute \( X_1 \).
5 Experimental Results

To justify the modality focusing hypothesis, we design two sets of experiments: (1) For synthetic data, we can control its inherent characteristics (i.e., $\alpha_1$, $\alpha_2$ and $\gamma$) and perform KD. We are curious whether teacher and student performance are as described in the modality focusing hypothesis. Meanwhile, we report the accuracy of our Algorithm 1 since whether a feature channel is general/species decisive is known. (2) For real-world data, we cannot know the ground truth $\gamma$. To evaluate the implication in Sec. 4.1, we experiment with a modality-general teacher and a modality-specific teacher (i.e., one teacher with a large $\gamma$ and one teacher with a small $\gamma$). The two teachers can be obtained by applying Algorithm 1 to estimate $p$ and then removing corresponding feature channels in the teacher (Sec. 5.2 and Sec. 5.3). Alternatively, we can apply specific training techniques to obtain a modality-general teacher (Sec. 5.4). In all, we provide experiments on 6 datasets (synthetic Gaussian, AV-MNIST, RAVDESS, VGGSound, NYU Depth and MM-IMDB) with diverse modalities (i.e., RGB images, depth images, audio and text). Due to space limit, some results (e.g., experiments on MM-IMDB [2] movie genre classification) and ablation studies are presented in the Appendix.

5.1 Synthetic Gaussian Data

We have presented the experiment with increasing $\gamma$ in Sec. 4.1. Below we show results with increasing $\alpha_2$ (i.e., decreasing $\gamma$). Data generation details are provided in the Appendix. Fig. 3 illustrates the process where modality-specific decisive features in modality 2 gradually dominate. With increasing $\alpha_2$, the teacher gradually improves since it receives more modality-specific decisive features for prediction. However, the student network fails to benefit from the improved teacher and performs slightly worse instead. Clearly, teacher performance is not reflective of student performance in this case. It is the amount of modality-general decisive information that really matters. This set of experiment offers a valid explanation about our empirical finding in Sec. 3.2 that a more accurate model does not necessarily serve as a good teacher.

Figure 3: An illustration of the modality focusing hypothesis with synthetic Gaussian data. Teacher modality is $x_2$ and student modality is $x_1$. With increasing $\alpha_2$ (i.e., decreasing $\gamma$), the teacher improves its prediction accuracy but the student network fails to benefit from KD.

Next, based on the saliency vector $p$ given by Algorithm 1, we identify feature channels that correspond to non modality-general decisive features and remove them in $X_2$ to form a “clean” version of data that only contains general decisive features, $\hat{X}_2$. We train a new teacher model with $\hat{X}_2$ and apply KD subsequently. The new teacher is more modality-general compared with the regularly-trained teacher. We experiment with different values of $\gamma$ and summarize the results in Table 2. Results are averaged over 10 runs.

As shown in Table 2, our proposed feature separation method can correctly identify modality-general features in input data (accuracy close to 1.0). Moreover, by removing non modality-general decisive feature channels, we can train a teacher that is specialized for distillation and obtain an improved student network. Since the modified teacher only relies on modality-general decisive features for prediction, we observe a accuracy loss in the teacher. However, the student network does not get affected and demonstrates increased accuracy for all choices of $\gamma$ instead. These results validate our proposed modality focusing hypothesis and shed light on how to utilize the hypothesis to facilitate the student network’s learning process.
Table 2: Evaluation on synthetic Gaussian data. ‘FS Acc.’ denotes the accuracy of our feature separation method for $X_2$. By removing non modality-general feature channels identified by Algorithm 1, we obtain a modality-general teacher with downgraded performance and its student with increasing accuracy; this validates our modality focusing hypothesis.

| $\gamma$ | FS Acc. | Teacher Accuracy (%) | Student Accuracy (%) |
|----------|---------|----------------------|----------------------|
|          |         | Regular | Modality-general | No-KD | Regular | Modality-general |
| 0.33     | 0.98    | 89.41   | 68.42 (↓) | 61.98 | 62.16   | 64.50 (↑)   |
| 0.50     | 0.99    | 89.62   | 73.41 (↓) | 67.64 | 67.86   | 70.25 (↑)   |
| 0.66     | 0.99    | 89.82   | 77.98 (↓) | 73.48 | 73.60   | 75.43 (↑)   |
| 0.75     | 0.99    | 89.70   | 79.41 (↓) | 76.46 | 76.53   | 77.70 (↑)   |

5.2 Emotion Recognition on RAVDESS

RAVDESS [24] is an audio-visual dataset containing 1,440 emotional utterances with 8 different emotion classes. We consider images as modality 1 and audio as modality 2. Teacher and student network are 3-layer CNNs. More details are provided in the Appendix. Since permutation is time-consuming at the model input level, we take middle-layer features (i.e., 128-dimensional features after the first linear layer) as input to Algorithm 1. After we obtain the saliency vector $p \in \mathbb{R}^{128}$, we evaluate KD under four settings: (1) regular KD; (2) random: we randomly nullify 50% of audio feature dimension, i.e., set the corresponding feature dimension to be its batch mean, which is identical for all samples in this batch. The student is trained to match predictions given by the “randomly masked” teacher; (3) modality-specific: we sort $p$ in descending order, nullify the first 50% feature dimension in the teacher and train a student subsequently; (4) modality-general: we sort $p$ in ascending order, nullify the first 50% feature dimension and perform KD. In this way, we remove modality-general decisive features in teacher’s decision process in the third setting and remove non modality-general decisive features in the last setting. While we can not know the true value of $\gamma$ for each teacher, we know that $\gamma$ increases in the order of modality-specific, random and modality-general teacher. This allows us to justify the proposed modality focusing hypothesis. Results are averaged over 5 runs.

Table 3: Test Accuracy (%) on RAVDESS Emotion Recognition. Teacher modality is audio and student modality is images. With identical model architecture, a modality-general teacher improves regular KD while a modality-specific teacher leads to a significantly worse student.

| Teacher | Regular | Modality-specific | Random | Modality-general |
|---------|---------|-------------------|--------|------------------|
| 77.60   | 77.22   | 42.66 (↓)        | 64.75  | 78.28 (↑)        |

As demonstrated in Table 3, a modality-specific teacher leads to a significantly downgraded student (i.e., accuracy drops from 77.22% to 42.66%). On the other hand, we observe an improved student (i.e., test accuracy increases to 78.28%) after nullifying non general decisive features in the teacher (modality-general setting). This indicates the critical role of modality-general decisive features in distillation. A modality-general audio teacher can better transfer knowledge to the visual student as it discards audio-specific information not understandable by the student.

5.3 Event Classification on VGGSound

VGGSound [6] is a large-scale audio-visual event classification dataset including over 200,000 video and 310 classes. We set modality 1 (i.e., student modality) as video and modality 2 (i.e., teacher modality) as audio. More experimental details are provided in the Appendix. We take features before the last fully connected layer as input to Algorithm 1 and obtain a 512-dimensional saliency vector $p$. Similar to the previous experiment, we obtain a random, modality-specific and modality-general teacher by nullifying 75% audio feature dimension according to $p$. We experiment with two backbone architectures, ResNet-18 and ResNet-50 [15], and report results in Table 4.
As shown in the Table 4, modality-general decisive features play a critical role in multimodal KD. The student learning from a modality-specific teacher suffers from a major performance degradation. On the contrary, nullifying non general decisive features in the teacher improves the student performance from 30.62% to 31.88% (for ResNet-18) and from 38.78% to 39.81% (for ResNet-50).

Moreover, we provide an ablation study on the ratio of feature dimension that gets nullified in Fig. 4. We can see that there exists a sweet spot for modality-general KD. As the ratio increases, the student performance improves in the beginning. The improvement indicates that non general decisive features in the teacher are gradually discarded, which in turn results in a better student. Later, after all non general decisive information are removed, the nullifying process starts to hinder student performance as general decisive features get nullified as well. These results align well with our understanding of the modality Venn diagram and demonstrate the efficacy of our proposed Algorithm 1.

**Table 4:** Student mean Average Precision (%) on VGGSound Event Classification. Teacher modality is audio and student modality is video. The student learning from a modality-general teacher demonstrates best performance.

| Model Architecture | ResNet-18 | ResNet-50 |
|--------------------|-----------|-----------|
| Teacher            | 29.86     | 37.51     |
| Regular            | 30.62     | 38.78     |
| Modality-specific  | 24.98 (↓) | 28.65 (↓) |
| Random             | 28.99     | 35.95     |
| Modality-general   | 31.88 (↑) | 39.81 (↑) |

5.4 **Semantic Segmentation on NYU Depth V2**

Finally, we revisit the example of NYU Depth V2 [29] in Sec. 3.2. Here the two modalities (i.e., RGB and depth images) have large similarity and can be taken as input for a single model. Therefore, we design a modality-agnostic teacher model for distillation. This can be seen as an alternative way to demonstrate the modality focusing hypothesis apart from our proposed feature separation method.

We adopt a teacher network that takes depth images as input to transfer knowledge to an RGB student. As shown in Table 5, regular KD (i.e., using the modality-specific teacher) does not bring many advantages: the student achieves a similar mIoU compared with a RGB model without KD. Based on this result, one might easily blame the failure of KD on teacher accuracy and assume that KD is not effective because the depth teacher itself yields poor performance (i.e., has an mIoU of 37.33%).

**Table 5:** Mean Intersection over Union (mIoU) on NYU Depth RGB-D Semantic Segmentation. Teacher modality is depth images and student modality is RGB images. Compared with regular crossmodal KD (i.e., using a modality-specific depth teacher), a modality-general teacher is more suitable for distillation and leads to a better student.

| Teacher Modality | mIoU (%) | Student Modality | mIoU (%) |
|------------------|----------|------------------|----------|
| No-KD            | -        | RGB              | 46.36    |
| Modality-specific-KD | D 37.33 | RGB              | 46.89    |
| Modality-general-KD | RGB/D 37.47 | RGB              | 47.93    |

The modality focusing hypothesis motivates us to think otherwise. We train a modality-general teacher that receives both RGB and depth images as input. Similar to [11], we use one single model to predict labels for two modalities with exactly same parameters. Note that the model architecture is identical to the previous modality-specific teacher to ensure a fair comparison. In this way, the teacher is forced to extract modality-general decisive features for prediction rather than rely on depth-specific features as it also takes RGB images as input. While we do not observe difference in teacher performance, the new teacher is categorized for distillation and leads to a better depth
student (i.e., student mIoU increases from 46.36% to 47.93%). These results show how our modality focusing hypothesis can indeed help diagnose failure of crossmodal KD and improve it.

6 Conclusion

In this work, we present a thorough investigation of multimodal KD. We find that modality-general decisive features are the key in multimodal KD performance. The proposed modality Venn diagram and modality focusing hypothesis characterize multimodal relationships and reveal when multimodal KD is effective. We further propose an approach to separate modality-general decisive features for real-world multimodal data and conduct extensive experiments to justify our proposed hypothesis. Finally, we hope the modality focusing hypothesis shed light on applications of multimodal KD and will raise interest for general understanding of multimodal learning in the community. As future work, we would like to develop corresponding theory to gain a theoretical understanding of the problem.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 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)*, pages 2143–2147. IEEE, 2020.

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[7] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *European Conference on Computer Vision*, pages 561–577. Springer, 2020.

[8] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[9] Zhengqi Gao, Sucheng Ren, Zihui Xue, Siting Li, and Hang Zhao. Training-free robust multimodal learning via sample-wise jacobian regularization. *arXiv preprint arXiv:2204.02485*, 2022.

[10] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *European Conference on Computer Vision*, pages 103–118, 2018.

[11] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. *arXiv preprint arXiv:2201.08377*, 2022.

[12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[14] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.
[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*, pages 770–778, 2016.

[16] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 578–587, 2019.

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[18] Minhao Hu, Matthias Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 772–781. Springer, 2020.

[19] Srinivas SS Kruthiventi, Pratyush Sahay, and Rajesh Biswal. Low-light pedestrian detection from rgb images using multi-modal knowledge distillation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4207–4211. IEEE, 2017.

[20] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[21] Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 775–783, 2020.

[22] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[23] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019.

[24] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[25] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[26] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018.

[27] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.
[33] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151. PMLR, 2019.

[34] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13325–13333, June 2021.

[35] Sucheng Ren, Zhengqi Gao, Tianyu Hua, Zihui Xue, Yonglong Tian, Shengfeng He, and Hang Zhao. Co-advice: Cross inductive bias distillation. *arXiv preprint arXiv:2106.12378*, 2021.

[36] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

[37] Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020.

[38] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE, 2019.

[39] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[41] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[42] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11612–11621, 2021.

[43] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multi-layer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[44] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, pages 664–679. Springer, 2016.

[45] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[46] Chen Xianing, Cao Qiong, Zhong Yujie, Zhang Jing, Gao Shenghua, and Tao Dacheng. Dearkd: Data-efficient early knowledge distillation for vision transformers. *arXiv preprint arXiv:2204.12997*, 2022.

[47] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021.

[48] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[49] Wei Zhang, Weixi Gu, Fei Ma, Shiguang Ni, Lin Zhang, and Shao-Lun Huang. Multimodal emotion recognition by extracting common and modality-specific information. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 396–397, 2018.

[50] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.
A Synthetic Gaussian Data

Assume two vectors $x_1 \in \mathbb{R}^{D_1}$ and $x_2 \in \mathbb{R}^{D_2}$ compose one multimodal data pair $(x_1, x_2)$. We select a subset of input features as decisive features, denoted by $x^* \in \mathbb{R}^D$. We assume that $x^*$ exist in both $x_1$ and $x_2$, and denote the corresponding decisive feature index set of $x_1, x_2$ as $J_1, J_2$. The separating hyperplanes are denoted by $\beta \in \mathbb{R}^D$. Formally, we generate one feature-label pair $(x_1, x_2, y)$ by:

$$x^* \sim N(0, I_D), \quad y \leftarrow \mathbb{1}(\langle \beta, x^* \rangle > 0)$$
$$x_1 \sim N(0, I_{D_1}), \quad x_{1, J_1} \leftarrow x_{J_1}$$
$$x_2 \sim N(0, I_{D_2}), \quad x_{2, J_2} \leftarrow x_{J_2}$$

As depicted in the modality Venn diagram, modality-general decisive features are decisive features shared by both modalities and thus indexed by $J_1 \cap J_2$. $J_1 \cup J_2$ represents the index set of decisive features from both modalities. Therefore, $\alpha_1 = 1 - \frac{|J_2|}{|J_1 \cup J_2|}$, $\alpha_2 = 1 - \frac{|J_1|}{|J_1 \cup J_2|}$ and $\gamma = \frac{|J_1 \cap J_2|}{|J_1 \cup J_2|}$. By changing $J_1$ and $J_2$, we can generate multimodal data with different inherent characteristics (i.e., different $\alpha_1, \alpha_2$ and $\gamma$). We consider 2 settings: (1) varying $\gamma$ (Sec. 4.1 in the main paper). Let $D_1 = 50, D_2 = 25$ and $D = 20$, we gradually increase $|J_1 \cap J_2|$ from 0 to 10, with a step size of 2 and perform KD on every step. (2) varying $\alpha_2$ (Sec. 5.1 in the main paper). Let $D_1 = D_2 = 50$ and $D = |J_1 \cup J_2|$ increase from 10 to 50, with a step size of 10.

Following [25], teacher and student are implemented as logistic regression models, and we use 200 samples for training and 1,000 samples for testing. $\beta$ is sampled from the standard normal distribution. $\rho$ in Eq. (1) in the main paper is set as 0.5. Results are averaged over 10 runs.

B Digit Classification on AV-MNIST

AV-MNIST [43] is an audio-visual dataset created by pairing audio and image features. The two modalities are MNIST images with 75% energy removed by principal component analysis and audio spectrograms with random natural noise injected. There are 50,000 pairs for training, 5,000 pairs for validation and 10,000 pairs for testing. Following [43, 9], we adopt a 6-layer CNN as the audio teacher network. The audio student network is implemented as a 3-layer CNN, and the multimodal teacher is a late fusion network. The multimodal teacher uses LeNet5 [20] as the image backbone and a 5-layer CNN as the audio backbone; Audio and image features are then concatenated and passed to fully-connected layers for the final prediction. We experiment with both $\rho = 0$ and $\rho = 0.5$, and repeat the experiments for 10 times. We have provided the results of $\rho = 0.5$ in Table 1 in the main paper, and a more detailed version can be found in Table 6.

Table 6: Evaluation of unimodal KD (UM-KD) and multimodal KD (MM-KD) on AV-MNIST. ‘Mod.’ is short for modality and ‘Params’ denotes model parameters. A and T represent audio and grayscale images, respectively.

| Teacher       | Mod. | Params | Acc. (%) | Student       | Mod. | Acc. (%) | Mod. | Acc. (%) |
|---------------|------|--------|----------|---------------|------|----------|------|----------|
| No-KD         | A    | 0.25M  | 84.57    | A             | 0.5  | 69.36 ± 0.79 | 68.36 ± 0.79 |
| UM-KD         | I + A| 0.24M  | 91.61    | A             | 0.5  | 69.46 ± 1.12 | 69.73 ± 0.73 |
| MM-KD         | I + A| 0.24M  | 91.61    | A             | 0.5  | 69.46 ± 1.12 | 69.73 ± 0.73 |

From the table, we can see that multimodal KD does not have advantages over unimodal KD for both values of $\rho$. Note that we design the unimodal teacher and multimodal teacher model to be roughly of the same size to rule out the factor of model capacity. The results indicate that multimodal KD is not effective in this case. We hypothesize that $\gamma$ is small for this dataset since a multimodal data pair is assembled by randomly pairing an image with an audio that belongs to the same class. Thus the two modalities are not naturally correlated and there may be little modality-general information.
Our proposed modality focusing hypothesis provides a plausible explanation for this failure case of multimodal KD.

C Emotion Recognition on RA VDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RA VDESS) [24] is released under a Creative Commons Attribution License. It contains videos and audios of 24 professional actors vocalizing two lexically-matched statements. For modality 1 (i.e., student modality), we uniformly sample single-frame images every 0.5 second from each video. For modality 2 (i.e., teacher modality), we adopt Kaiser best sampling and take mel-frequency cepstral coefficients (MFCCs) features from corresponding audio. We randomly split image-audio pairs, and have 7,943 data for training, 2,364 data for validation and 1,001 data for testing. Similar to [48], the teacher and student architecture are 3-layer CNNs followed by 3 fully-connected layers. We set $\rho$ in Eq. (1) in the main paper as 0 (i.e., only use $L_{kd}$ for distillation) to fully observe the teacher’s influence on student performance. We report results with three nullified feature dimension ratio in Table 7, which is a detailed version of Table 3 in the main paper. Results are averaged over 5 runs.

Table 7: Test Accuracy (%) on RA VDESS Emotion Recognition. Teacher modality is audio and student modality is images. Modality-general KD can improve student performance for all nullified feature dimension ratio.

| Nullified Ratio (%) | Regular     | Modality-specific | Random        | Modality-general |
|---------------------|-------------|------------------|---------------|-----------------|
| 25                  | 77.22       | 75.26 (↓)        | 77.24         | 78.12 (↑)       |
| 50                  | 77.22       | 42.66 (↓)        | 64.75         | 78.28 (↑)       |
| 75                  | 77.22       | 12.00 (↓)        | 63.40         | 77.82 (↑)       |

As shown in the table, with more nullified feature channels, the random and modality-specific version both suffer from a heavy performance degradation. On the contrary, the modality-general setting (i.e., nullifying “non modality-general decisive” feature channels) still attains satisfactory performance even when the nullifying ratio goes to 75% and outperforms the regular KD baseline. This demonstrates the efficacy of our proposed feature separation method as well as the practical value of the modality focusing hypothesis.

D Event Classification on VGGSound

VGGSound is a large-scale audio-visual correspondent dataset, under a Creative Commons Attribution 4.0 International License. We randomly choose 100 class from its 310 classes and obtain 56,614 audio-video pairs for training and 4,501 audio-video pairs for testing. Videos clips and audio spectrograms are modality 1 and modality 2 input features, respectively. The audio teacher is implemented as a ResNet-18/ResNet-50 backbone followed by linear layers and the video student network is the same architecture with 2D convolution replaced by 3D convolution. For Table 4 in the main paper, we set $\rho$ in Eq. (1) to be 0 and experiment with both ResNet-18 and ResNet-50. In Table 8, we report results of both $\rho = 0$ and $\rho = 0.5$ with the ResNet-18 backbone. The conclusion is consistent: a modality-general KD improves student performance while a modality-specific teacher results in performance degradation. These results help validate our proposed modality focusing hypothesis.

Table 8: Student mean Average Precision (%) on VGGSound Event Classification. Teacher modality is audio and student modality is video.

| $\rho$ | 0    | 0.5  |
|--------|------|------|
| Teacher| 29.86| 29.86|
| Regular| 30.62| 30.70|
| Modality-specific| 24.98 (↓)| 28.14 (↓)|
| Random | 28.99| 29.56|
| Modality-general | 31.88 (↑) | 31.98 (↑) |

Figure 5: Ablation study on permutation number.
Number of permutation times. In Algorithm 1 in the main paper, we repeat permutation $K$ times for a better estimation of each feature dimension’s saliency value. Fig. 5 provides an ablation study on the number of permutation times (i.e., $K$). From the figure, as $K$ increases, the modality-general KD performs better. This indicates that a larger $K$ leads to a more accurate estimation of $p$. Consequently, we obtain a better student by removing non modality-general feature channels based on $p$.

E Semantic Segmentation on NYU Depth V2

NYU Depth V2 [29] contains 1,449 aligned RGB and depth images with 40-class labels, where 795 images are used for training and 654 images are for testing. $\rho$ is set as 0.5. We implement two model architectures for the multimodal teacher: (1) Channel Exchanging Networks (CEN) [46] and (2) Separation-and-Aggregation Gate (SA-Gate) [7]. The unimodal teacher and student are adopted as the RGB branch of the corresponding multimodal network. The results are shown in Table 9, which corresponds to the right part of Table 1 in the main paper.

Table 9: Evaluation of unimodal KD (UM-KD) and multimodal KD (MM-KD) on NYU Depth V2. ‘Mod.’ is short for modality. RGB and D represent RGB images and depth images, respectively.

| Teacher Mod. | Student Mod. | CEN     |     |     | SA-Gate |     |     |
|--------------|--------------|---------|-----|-----|---------|-----|-----|
| RGB          | 45.69        | RGB     | 45.69 | RGB | 46.36  |
| RGB          | 45.69        | RGB     | 46.23 | RGB | 46.36  |
| RGB + D      | 51.14        | RGB     | 46.70 | RGB | 47.78  |

Table 9 demonstrates that multimodal KD is not effective in both cases. The great advantages in teacher performance does not enhance student performance. Adopting CEN as the multimodal teacher seems better than SA-Gate, but the improvement compared with unimodal KD is still marginal (i.e., from 46.23% to 46.70%). According to the modality focusing hypothesis, different teacher networks utilize different amount of modality-general decisive information for prediction, which results in different distillation performance. We hypothesize that CEN has a larger $\gamma$ than SA-Gate due to their model design: CEN shares all parameters for the RGB and depth input except for Batch Normalization layer while SA-Gate has separate encoders for the two modalities. This indicates that CEN is more “modality-general” than SA-Gate and may explain their differences. There may be other factors lying behind, and one future direction is to develop methods to compare existing model architectures to decide which one is best for distillation.

For results reported in Sec. 5.2 in the main paper, we adopt the depth teacher and RGB student as the depth and RGB branch of SA-Gate, respectively. $\rho$ is set as 0.5.

F Movie Genre Classification on MM-IMDB

MM-IMDB [2] is the largest publicly available multimodal dataset for genre prediction on movies. It contains 25,959 movie titles and posters that belong to 27 movie genres. We pick two movie genres (i.e., drama and comedy) for multi-label classification. There is 15,552 data for training, 2,608 for validation and 7,799 for testing. We adopt the same pre-processing method as [2] to extract image and text features. Modality 1 is text and modality 2 is image. We aim to improve an unimodal text network with multimodal KD and experiment with two multimodal teacher networks: (1) Following [48], we train a multimodal network that receives pseudo labels from an unimodal image network; (2) We regularly train a multimodal network with labels. The first teacher only has access to pseudo labels from the image modality and is thus more “modality-specific” (i.e., has a smaller $\gamma$) compared with the second teacher. We term the two teacher networks as “specific” and “general”, respectively. The unimodal and multimodal architecture are identical to [22]. We randomly split training data as 1:1, and use the first half to train the unimodal teacher and the general multimodal teacher. The other part of data is used for training the student network, and we set $\rho = 0$. 

15
Table 10: F1 score (%) on MM-IMDB movie genre classification. T and I represent text and images, respectively. While MM-KD (general) has a lower teacher score than MM-KD (specific), it leads to a student with better performance.

| Modality   | Teacher | Teacher | Student |
|------------|---------|---------|---------|
|            | micro F1| macro F1| micro F1| macro F1 |
| UM-KD      | T       | 61.76   | T       | 61.04   |
| MM-KD (specific) | T+I    | 62.01   | T       | 61.77   |
| MM-KD (general) | T+I    | 61.01   | T       | 65.09   |

Table 10 shows the teacher and student performance for both unimodal KD and multimodal KD. We select three teacher models that have similar performance on test data, and use them for distillation to detach the influence of teacher performance. Clearly, the three teachers transfer different knowledge to the student. The unimodal teacher comes from the image modality and the specific multimodal teacher is also biased towards the image modality due to its training strategy. Finally, the general multimodal teacher utilizes more modality-general information compared with the previous two teachers (i.e., has a larger $\gamma$). As can be seen from the table, MM-KD (general) results in the best unimodal text student. This lends support to our proposed modality-focusing hypothesis.