Towards Out-of-Modal Generalization without Instance-level Modal Correspondence

Anonymous authors
Paper under double-blind review

Abstract

The world is understood from various *modalities*, such as appearance, sound, and language. Since each modality only partially represents objects in a certain meaning, leveraging additional ones is beneficial in both theory and practice. However, exploiting novel modalities normally requires cross-modal pairs corresponding to the same instance, which is extremely resource-consuming and sometimes even impossible, making knowledge exploration of novel modalities largely restricted. To seek practical multi-modal learning, here we study *Out-of-Modal (OOM) Generalization* as an initial attempt to generalize to an unknown modality without given instance-level modal correspondence. Specifically, we consider Semi-Supervised and Unsupervised scenarios of OOM Generalization, where the first has scarce correspondences and the second has none, and propose *connect & explore* (COX) to solve these problems. COX first connects OOM data and known In-Modal (IM) data through a variational information bottleneck framework to extract shared information. Then, COX leverages the shared knowledge to create emergent correspondences, which is theoretically justified from an information-theoretic perspective. As a result, the label information on OOM data emerges along with the correspondences, which help explore the OOM data with unknown knowledge, thus benefiting generalization results. We carefully evaluate the proposed COX method under various OOM generalization scenarios, verifying its effectiveness and extensibility.

1 Introduction

To understand the world, we use various data *modalities*, such as image data (He et al., 2016; 2017; Ren et al., 2015) and text data (Devlin et al., 2018; Vaswani et al., 2017). Each modality describes objects through a certain physical perspective, and thus contributing to understanding objects. Therefore, *multi-modal learning* (MML) (Alayrac et al., 2022; Ngiam et al., 2011; Radford et al., 2021; Socher et al., 2013) which learns from multiple modality data has been a core research topic in AI. Thanks to the utilization of various modalities, the learning performance has shown to be beneficial on various tasks compared to uni-modal learning (Huang et al., 2021; Lu, 2024; Radford et al., 2021; Sun et al., 2020), such as cross-modal retrieval and generation (Yasunaga et al., 2023; Zhang et al., 2021; Zhen et al., 2019), human-computer interaction (Pantic & Rothkrantz, 2003; Rahman et al., 2022), and robotics (Jiang et al., 2023; Yu et al., 2023).

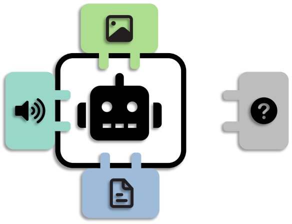

Figure 1: AI is enhanced as more modalities are incorporated, so how can we teach AI to learn from novel modalities based on the ones it already know?

However, existing states of the art performance are not satisfactory, and emerging modalities need to be explored and leveraged effectively just like the relatively new data modalities of the geomagnetic fields (Hashimoto, 1926), sound waves (Harley et al., 2003), and electromagnetic waves (Weinstein, 1988). Therefore, emerging technologies have been constantly leveraging new sensors to enhance their performance. For example, Embodied AIs (Savva et al., 2019) already possess abilities like 3D vision and language, but they are still exploring novel skills, such as tactile sensing and bio-sensing. Since it is hard to leverage such uncommon and inexperienced skills in practice, adapting the knowledge from common modalities to better understand the novel ones could be potentially beneficial, as shown in Figure 1. In practice, most existing MML investigations (Radford et al., 2021; Girdhar et al., 2023; Wang et al., 2024; Zhu et al., 2023) require *instance-level modal correspondence*, *i.e.,* multi-modal data are paired with the same instance, which is often hard to satisfy in real-world scenarios when facing novel modalities (Liang et al., 2023; 2021; Sun et al., 2020; Xia et al., 2024). For a robotic example, some modalities are common and easy to acquire, *e.g.*, vision and language, but others like tactile data might need special sensors to resample from the same objects seen or spoken. Unfortunately, the resample could no longer access the same objects in real-world applications. As a result, the new modalities usually have incomplete or even no correspondence, which could seriously block the knowledge interaction across modalities and hinder the benefits brought by MML. Hence, a question naturally occurs: *Do we really need instance-level modal correspondence to explore novel modalities?*

This paper studies a practical yet unexplored problem named *Out-of-Modal (OOM) Generalization*. Particularly, we are given several modalities, *i.e.,* In-Modal (IM) data, and then our goal is to generalize to an unknown modality that has no correspondence to any of the known ones, or in some cases only scarcely paired. Such a problem setting implies the utilization of novel modalities in realistic situations: Even though our knowledge is limited to certain modalities, *e.g.,* human perceptions only have touch, sight, sound, smell, and taste, but we can still understand unperceivable ones such as magnetism by utilizing inherently-possessed senses, *e.g.*, feel the force when pulling two magnets together; or see the magnetic field by observing the alignment of iron filings around a magnet.

Based on this insight, we utilize IM perceptors that contain prior knowledge to encode known IM data, which can be implemented using existing MML models (Radford et al., 2021; Girdhar et al., 2023; Zhu et al., 2023; Wang et al., 2024), and an OOM learner which learns novel modalities without any prior knowledge. By analyzing the interactions between the extracted latent features, we show theoretically and empirically that the knowledge from OOM data can be gradually discovered, allowing us to train the OOM learner to enhance its understanding of the novel modality, as shown in Figure 2. First, we consider *semi-supervised OOM generalization* where few correspondences are given. Based on the correspondence, we can capture the prior probability distribution and learn mappings that connect OOM data and IM data. Through an information-theoretic perspective, we propose connect & explore (COX), which encourages the agreement on mappings across modalities, further the cross-modal knowledge can be shared and novel information can be explored. Then, we extend COX to an *unsupervised OOM generalization* scenario where there is no instance-level correspondence at all. To tackle such a challenge, we enhance the OOM-IM connections by maximizing cross-modal interaction. To simplify such an unsupervised problem into a semi-supervised case, we select data pairs from cross-modal mappings and IM features, respectively. According to feature similarity, we assume that the data pairs closing to OOM mappings can be considered as correspondence. Under this assumption, we can leverage the emerging correspondence and solve the unsupervised case via the semi-supervised solution. To validate the proposed method, we carefully design experiments using various multi-modal datasets to validate the effectiveness of COX. Moreover, we provide extensive analyses in various scenarios to understand our method and inspire future research. To sum up, our contributions are three-fold:

- We discover a novel and practical problem named Out-of-Modal Generalization, which aims to explore a novel modality using the knowledge from known modalities.
- We consider two typical situations: Semi-Supervised OOM generalization and Unsupervised OOM generalization, and propose a connect & explore framework to tackle both problems from an information-theoretic perspective.
- We conduct extensive experiments to tackle the OOM generalization on various datasets and provide intuitive insights to help inspire future research.

2 Related Work

Modality Generalization generally focuses on leveraging the knowledge from some modalities and generalizing to another one. Existing studies are conducted in different settings and with various tasks. Cross-Modal Fine-Tuning mimics transfer learning by adapting the distribution of IM data to OOM data using the same model. Shen et al. (2023) proposed to conduct distribution alignment to achieve this goal which requires both pre-trained knowledge and labeled target modality data. Based on a similar problem setting, Cai et al. (2024) designed a gradual modality generation scheme that selects the top-k active feature patches from target modalities, and replaces them with source modalities patches. Such a progressive strategy can align target modal data to ensure generalization. Cross-Modal Generalization uses separate encoders and focus on generalizing to a different modality data from the same instance. Liang et al. (2021) used meta-learning to align OOM data to IM

Table 1: A comparison of different MML problems and their corresponding settings.

Problem	References	IM Knowledge	OOM Knowledge	Correspondence
Cross-Modal Fine-Tuning	Shen et al. (2023); Cai et al. (2024)	pre-trained & labeled	labeled	✗
Cross-Modal Generalization	Liang et al. (2021) Xia et al. (2024)	pre-trained & labeled pre-trained & labeled	pre-trained pre-trained & labeled	✓ ✓
MML w/o labeled Multi-Modal Data	Liang et al. (2023)	partially labeled	partially labels	✓
OOM Generalization	Semi-Supervised case (Section 3.3) Unsupervised case (Section 3.4)	pre-trained & labeled pre-trained & labeled	scarcely labeled ✗	A few ✗

space and generalize to OOM tasks dynamically. Xia et al. (2024) studied a different setting where IM and OOM data are both known during training. Then, a unified representation space is learned to help downstream generalization on OOM data. Some other studies considers generalization when all modalities are available, Ma et al. (2019) studied cross-modal generalization without paired data, Wang et al. (2023) applied the information bottleneck to CLIP training, Fang et al. (2024) conducted multi-modal fusion under limited clinical data, and Dong et al. (2023) considered domain generalization with fully-paired multi-modal data. A recent study MML without Labeled Multi-Modal Data (Liang et al., 2023) proposed a different setting where both IM and OOM data have labels, but they are not paired. Instead, additional unlabeled paired multi-modal data is given for learning the interaction between modalities. Moreover, Xue et al. (2022) understood the interactions and applied it to knowledge distillation. Except for cross-modal fine-tuning which follows transfer learning, existing MML works mostly require instance-level correspondence. This work proposes OOM Generalization, where there is no correspondence and the OOM knowledge is barely provided. The comparison of related works is shown in Table 1.

Modality Binding aims to learn a joint embedding space across different modalities. Contrastive Language-Image Pre-training CLIP (Radford et al., 2021) is the first work that aligns image with language data. Then, ImageBind (Girdhar et al., 2023) proposed to use vision modalities to bind various modalities into the same representation space. Further, LanguageBind (Zhu et al., 2023) proposed using language as an alternative solution, which binds various modalities similarly. Recently, FreeBind (Wang et al., 2024) extended the existing unified space into an additional expert space. Specifically, two types of binding were considered, namely space displacement bond and space combination bind. Since modality binding often requires a large amount of data with correspondence, the selected modalities are often quite common. Therefore, the OOM generalization problem can take advantage of the development of modality binding by leveraging the encoders as our IM perceptors to learn novel modalities.

3 OOM Generalization

In this section, we first formalize the OOM generalization setting. Then, we demonstrate the proposed method. Further, we consider a Semi-Supervised case where a few correspondences are available and an Unsupervised scenario where there is no correspondence, showing that the proposed method can successfully tackle both settings and effectively leverage unpaired OOM data.

3.1 Problem Setting

In OOM generalization, we are given a set of known modalities $\{\mathcal{M}^{\mathrm{I}}_1, \ldots, \mathcal{M}^{\mathrm{I}}_K\}$ where $\mathcal{M}^{\mathrm{I}}_{k\in\{1,\ldots,K\}} = \{(x^{\mathrm{I}}_{k,i}, y^{\mathrm{I}}_{k,i})^N_{i=1} \in \mathcal{X} \times \mathcal{Y}\}$ is composed of N number of labeled IM examples with its subscript i denoting the correspondence across different modalities. Moreover, we have an unknown modality $\mathcal{M}^{\mathrm{O}} = \{(x^{\mathrm{O}}_j)^M_{j=1}\}$ containing M unlabeled OOM examples. In some cases, it is possible to obtain few correspondences with IM data, then our OOM data could be $\mathcal{M}^{\mathrm{O}} = \{(x^{\mathrm{O}}_i, y^{\mathrm{O}}_i)\}^L_{i=1} \cup \{(x^{\mathrm{O}}_j)\}^M_{j=L+1}$, where $L \ll M$ and the subscript i traces the corresponding IM data instance and label.

To tackle OOM generalization, we propose a learning framework as shown in Figure 2. Particularly, we use a set of IM perceptors $\{g^{\mathrm{I}}_1, \ldots, g^{\mathrm{I}}_K\}$ to perceive IM data, which can be realized by many existing modality-binding models, such as ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2023).

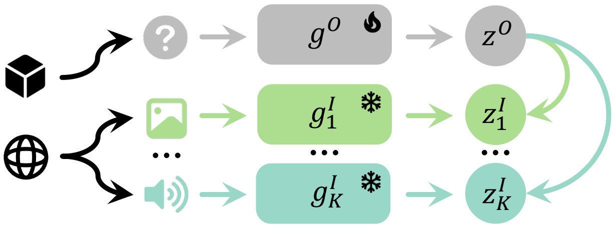

Figure 2: Learning framework of our OOM generalization.

Then, the features of IM data are obtained via $z_k^{\text{I}} = g_k^{\text{I}}(x_k^{\text{I}})$. Moreover, we use an OOM learner g^{O} to learn features z^{O} from OOM data through $z^{\text{O}} = g^{\text{O}}(x^{\text{O}})$. Our goal is to effectively generalize to OOM data by exploring the relationships between the OOM feature z^{O} and IM features $\{z_k^{\text{I}}\}_{k=1}^{K}$. Note that we only focus on the generalization performance of OOM data, the improvement of learning IM data is not the goal of this paper. Therefore, we freeze the parameters of all IM perceptors and only train the OOM learner during experiments. On top of the above models, we further define classifiers $h^{\text{O}}(x^{\text{O}}) := h^{\text{O}}(x^{\text{O}}; g^{\text{O}})$ and $h_k(x_k^{\text{I}}) := h_k(x_k^{\text{I}}; g_k^{\text{I}})$ that make predictions.

3.2 METHODOLOGY: CONNECT & EXPLORE (COX)

Here we elucidate the proposed method based on the interactive relationship between modalities (Liang et al., 2023; Williams & Beer, 2010). Specifically, the total information of two modalities under a certain task is decomposed into 1) *commonality*[1] which indicates common attributes across modalities, 2) *uniqueness* that is only presented in each modality, and 3) *synergy* denoting the emerging information when modalities are presented together. Note that we do not consider 3) in this paper as our goal is generalizing to OOM data.

To generalize to an unknown modality based on common ones, we aim to extract the commonality that can help partially comprehend OOM data based on IM data. Then, we model the posterior distribution of OOM data by selecting anchor points with minimum uniqueness. To this end, the OOM generalization can be successfully established. The proposed COX method comprises two steps: 1) learning connections by mapping IM data to OOM data to extract commonality, and 2) exploring high uniqueness OOM data by matching their posterior to high-commonality OOM data.

Connection across Modalities that capture the shared knowledge is learnable through generative models (Lu, 2024). Here we follow the variational information bottleneck (VIB) framework (Alemi et al., 2016) to achieve this goal. We assume that given IM data X^{I} and OOM data X^{O}, the latent variable V extracted from X^{I}[2], and label Y, the joint distribution can be factorized as

$$p(X^{\text{I}}, X^{\text{O}}, V, Y) = p(V, Y|X^{\text{O}}, X^{\text{I}})p(X^{\text{O}}|X^{\text{I}})P(X^{\text{I}}), \tag{1}$$

where we assume $p(V, Y|X^{\text{O}}, X^{\text{I}}) = p(V|X^{\text{I}})p(Y|X^{\text{I}})$, corresponding to the Markov chains $V \leftrightarrow X^{\text{I}} \leftrightarrow X^{\text{O}}$ and $X^{\text{I}} \leftrightarrow Y \not\leftrightarrow X^{\text{O}}$. Such an assumption means that V is not related to X^{O} (Alemi et al., 2016) and the given label Y is not directly connected to X^{O} under our OOM setting. Intuitively, given an IM datum, *i.e.*, dog image, it is sufficient to infer the label "dog", and the same for inferring from an unknown OOM datum, *i.e.*, dog bark. Thus, in common multi-modal settings, the label prediction using IM information dog image is not further conditioned on OOM knowledge dog bark, because here the OOM knowledge is redundant when IM data is given.

Our goal is to extract valuable knowledge from IM data to leverage OOM data by maximizing the information commonality (Liang et al., 2023; Williams & Beer, 2010):

$$\max I(X^{\text{O}}; X^{\text{I}}; Y) = I(X^{\text{O}}; X^{\text{I}}) - I(X^{\text{O}}; X^{\text{I}}|Y), \tag{2}$$

where $I(X^{\text{O}}; X^{\text{I}}; Y)$ denotes the mutual information between X^{O} and X^{I} regarding the task Y, *i.e.*, the label; and $I(X^{\text{O}}; X^{\text{I}}|Y)$ indicates the conditional mutual information irrelevant to Y. We start with the first term:

$$I(X^{\text{O}}; X^{\text{I}}) = \int dx^{\text{O}}dx^{\text{I}}p(x^{\text{O}}, x^{\text{I}}) \log \frac{p(x^{\text{O}}, x^{\text{I}})}{p(x^{\text{O}})p(x^{\text{I}})} = \int dx^{\text{O}}dx^{\text{I}}p(x^{\text{O}}, x^{\text{I}}) \log \frac{p(x^{\text{O}}|x^{\text{I}})}{p(x^{\text{O}})}, \tag{3}$$

where $p(x^{\text{O}}|x^{\text{I}}) = \int dvp(x^{\text{O}}, v|x^{\text{I}}) = \int dvp(x^{\text{O}}|v)p(v|x^{\text{I}})$ can be approximated via a decoder $q(x^{\text{O}}|v)$. Since the Kullback Leibler (KL) divergence is always non-negative, we have $\text{KL}[p(X^{\text{O}}|V) \parallel q(X^{\text{O}}|V)] \geq 0 \Rightarrow \int dx^{\text{O}}p(x^{\text{O}}|v) \log p(x^{\text{O}}|v) \geq \int dx^{\text{O}}p(x^{\text{O}}|v) \log q(x^{\text{O}}|v)$, and thus we can have

$$\begin{aligned} I(X^{\text{O}}; X^{\text{I}}) &\geq \int dx^{\text{O}}dx^{\text{I}}p(x^{\text{O}}, x^{\text{I}}) \log \frac{\int dvq(x^{\text{O}}|v)p(v|x^{\text{I}})}{p(x^{\text{O}})} \qquad (4) \\ &= \int dx^{\text{O}}dx^{\text{I}}dvp(x^{\text{O}}, x^{\text{I}}) \log q(x^{\text{O}}|v)p(v|x^{\text{I}}) + H(X^{\text{O}}), \qquad (5) \end{aligned}$$

[1] It is originally termed "redundancy" which is negative. However, such property is quite positive for tackling our problem, and hence we rename it "commonality".

[2] Note that the latent variable V here is different from the feature representation z^{I} and z^{O}.

where the last term is independent of our optimization process. Further, we rewrite $p(x^{\mathrm{O}}, x^{\mathrm{I}}) = \int dv p(x^{\mathrm{O}}, x^{\mathrm{I}}, v) = \int dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}})$. Then, we have the following lower bound:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}) \geq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v) p(v|x^{\mathrm{I}}), \tag{6}$$

which is realized by sampling from the joint data distribution, the latent variable from our encoder $p(v|x^{\mathrm{I}})$, and the tractable variational approximation $q(x^{\mathrm{O}}|v)$.

Similarly, we can upper-bound the second term $I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y)$ (details shown in Appendix A.1):

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y) \leq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) - \log h^{\mathrm{O}}(y|x^{\mathrm{O}}), \tag{7}$$

where $h^{\mathrm{O}}(y|x^{\mathrm{O}})$ is our classifier model for predicting OOM data. To this end, we can lower-bound our objective by combining equation 6 and equation 7:

$$\begin{aligned} I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) \geq & \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v) p(v|x^{\mathrm{I}}) \\ & - \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) + \log h^{\mathrm{O}}(y|x^{\mathrm{O}}) = \mathcal{L}_{\text{con}}. \end{aligned} \tag{8}$$

The above lower bound contains two part: 1) OOM data reconstruction where we reconstruct X^{O} using the latent V and 2) OOM data label prediction where we model the label distribution Y. In practice, we can approximate $p(x^{\mathrm{O}}, x^{\mathrm{I}}, y)$ using empirical samples from IM and OOM data. Moreover, we use encoder $p(v|x^{\mathrm{I}})$ without any prior assumptions because we can leverage the feature distribution from the pre-trained IM perceptors. Additionally, a classifier $h(y|x^{\mathrm{O}})$ is optimized to categorize OOM data based on given labels. Empirically, we can minimize

$$\mathcal{L}_{\text{con}} := \frac{1}{M} \sum_{i=1}^{M} \| x_i^{\mathrm{O}} - q(x_i^{\mathrm{O}}|v_i) p(v_i|x_i^{\mathrm{I}}) \|_2^2 - \log h^{\mathrm{O}}(y_i|x_i^{\mathrm{O}}), \tag{9}$$

where we use the reconstruction error $\| \cdot \|_2^2$[3] to realize the log-likelihood $q(x^{\mathrm{O}}|v)p(v|x^{\mathrm{I}})$, as similarly done by Kingma & Welling (2013). After building the connections, we can ensure the task-relevant information shared across modalities is learned, which helps partially understand OOM data regarding its commonality. However, note that the second term in Eq. 23 is not fully leveraged which contains $p(y|x^{\mathrm{I}})$ modeled by the IM perceptors. Take a step further, we can obtain $-\int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log \frac{p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}})}{h^{\mathrm{O}}(y|x^{\mathrm{O}})}$. Since $p(x^{\mathrm{O}}|x^{\mathrm{I}})p(x^{\mathrm{I}})$ is fixed in label prediction, we can derive $-\mathrm{KL}(p(y|x^{\mathrm{I}}) \parallel h^{\mathrm{O}}(y|x^{\mathrm{O}}))$ which implies that the label information related IM data can be harnessed to explore commonality. Next, we demonstrate how the commonality helps OOM generalization, and provide a solution to explore uniqueness.

Exploration of Uniqueness can be achieved via selecting and exploring the OOM data with high uniqueness. To identify these data, we can leverage the agreement and disagreement achieved by the optimal classifiers from various IM data. Our final goal is to optimize via

$$\min_{h^{\mathrm{O}}} \mathrm{KL}(h^{\mathrm{O}}(y|x_d^{\mathrm{O}}) \parallel h^{\mathrm{O}}(y|x_a^{\mathrm{O}})), \text{where } x_d^{\mathrm{O}} \in \mathcal{D}, x_a^{\mathrm{O}} \in \mathcal{A}, \tag{10}$$

in which h_1^* and h_2^* denote the optimal classifiers found in two IM data x_1^{I} and x_2^{I}, respectively, and x_d^{O} and x_a^{O} are selected from OOM data with modality disagreement $\mathcal{D} := \{x^{\mathrm{O}} : h_1^*(x^{\mathrm{O}}) \neq h_2^*(x^{\mathrm{O}})\}$ and agreement $\mathcal{A} := \{x^{\mathrm{O}} : h_1^*(x^{\mathrm{O}}) = h_2^*(x^{\mathrm{O}})\}$, respectively. Here we use two in-modalities for simplicity, but the conclusion can be extended to multiple modalities. Moreover, the data with agreement is considered anchor points that guide the exploration of those with disagreement. This objective aims to match the posterior of OOM data with uniqueness $h^{\mathrm{O}}(y|x_d^{\mathrm{O}})$ to the one of anchor points $h^{\mathrm{O}}(y|x_a^{\mathrm{O}})$. To justify this, we first define modality disagreement:

Definition 1 (Modality disagreement)**.** Given X_1, X_2 and target Y, as well as their corresponding optimal classifiers h_1^* and h_2^*, their modality disagreement is defined as $\alpha(h_1^*, h_2^*) = \mathbb{E}_{p(x_1,x_2)}[d(h_1^*, h_2^*)]$ where $d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$ is a distance function in the label space scoring the disagreement between h_1^* and h_2^*.

[3] Although training generative models in input space is computationally less efficient, we show in experiments that it is feasible to connect modalities in the feature space.

Theorem 1. Given two Bayes' optimal classifiers h_1^* and h_2^* from two in-modalities, under relaxed triangle inequality, inverse Lipschitz condition, and classifier optimality assumptions (Sridharan & Kakade, 2008), the modalities disagreement is upper-bounded by (see details in Appendix A.2)

$$\alpha(h_1^*, h_2^*) \leq I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + 2I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). \tag{11}$$

Finally, based on the decomposition of the task-related mutual information of X^{O}: $I(X^{\mathrm{O}}, Y) = I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y)$, as shown in Figure 3, we can achieve

$$\alpha(h_1^*, h_2^*) \leq I(X^{\mathrm{O}}, Y) - I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}), \tag{12}$$

where the first term denotes the overall information, the second term indicates the commonality shared between all modalities, and the third term stands for the uniqueness only preserved in OOM data. Intuitively, when we try to increase the modality disagreement, the commonality is decreased and OOM uniqueness is increased, which successfully justifies our learning objective: In order to explore the uniqueness of OOM data, we can explore the ones with high modality disagreement; conversely, the OOM data with high commonality and low uniqueness is found where agreement is achieved among h_1^* and h_2^*. Therefore, we select such data as anchor points that provide informative guidance to help explore uniqueness.

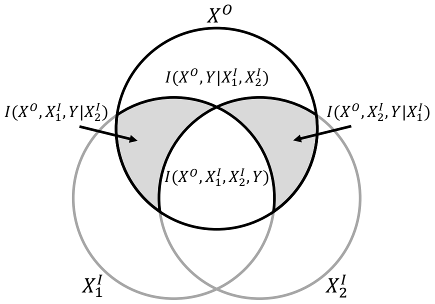

Figure 3: Decomposition of $I(X^{\mathrm{O}}, Y)$.

Next, we consider two realistic scenarios of OOM generalization and demonstrate how the proposed COX method can tackle them.

3.3 Semi-Supervised OOM Generalization

We start with a semi-supervised case where a few correspondences are available in OOM data, as shown in Figure 4 (a). Based on the VIB framework proposed in Section 3.2, we first leverage the OOM data $\{(x_i^{\mathrm{O}}, y_i^{\mathrm{O}})\}_{i=1}^L$ corresponding to IM data $\{(x_{k,i}^{\mathrm{I}}, y_{k,i}^{\mathrm{O}})\}_{i=1}^L, \forall k \in \{1, \ldots, K\}$ to build K connections using additional generative models that can be trained via a point-to-point mapping. As a result, the mappings on the OOM feature space can successfully match the OOM feature distribution, which allows us to directly apply IM data posteriors to select and explore the uniqueness of OOM data. Hence, we formulate our objective as

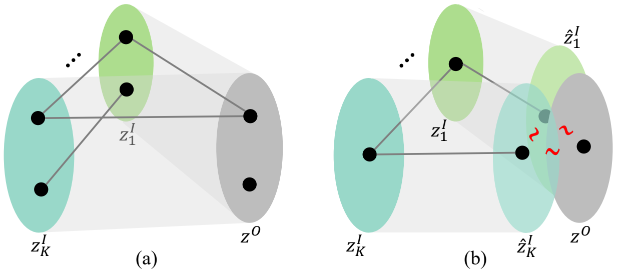

Figure 4: Two scenarios: (a) Semi-Supervised OOM Generalization and (b) Unsupervised OOM Generlaizaiton.

$$\min_{h^{\mathrm{O}}} \mathcal{L}_{\mathrm{ssl}} := \frac{1}{L}\sum_{i=1}^{L} \mathrm{CE}(h^{\mathrm{O}}(x_i^{\mathrm{O}}), y_i^{\mathrm{O}}) + \frac{1}{L+|\mathcal{D}|}\sum_{x_{d,j}\in\mathcal{D}}\sum_{x_{i=1}^{L}} \mathrm{KL}(h^{\mathrm{O}}(x_{d,j}^{\mathrm{O}}) \,\|\, h^{\mathrm{O}}(x_i^{\mathrm{O}}); h_1^*, h_2^*), \tag{13}$$

where the first term exploits labeled OOM data with correspondence and the second term explores OOM data $\mathcal{D}$ with modality disagreement by minimizing its KL divergence from the label posterior. Through the above objective, we can maximally exploit the uniqueness of OOM data to achieve effective OOM generalization.

3.4 Unsupervised OOM Generalization

As for the unsupervised case, we propose two-phase training: 1) we first conduct a warm-up training to initialize the OOM feature space and the connection, and 2) then, we enhance the connection by creating emergent correspondence and further exploring OOM data.

Specifically, we select anchor points from OOM data by directly applying modality agreement among all Bayes' optimal classifiers from IM data via

$$\mathcal{A}_{\text{sorted}} = \text{SORT}_T(\mathcal{A}, \frac{1}{K}\sum_{k=1}^{K} \max h_k^*(x^{\text{O}})), \text{ where } \mathcal{A} = \{\forall x^{\text{O}} \in \mathcal{M}^{\text{O}}\colon h_1^*(x^{\text{O}}) = \cdots = h_K^*(x^{\text{O}})\}, \quad (14)$$

where the $\text{SORT}_T(\cdot,\cdot)$ is a sort function, which ranks each element x^{O} in $\mathcal{A}$ based on the value of $\frac{1}{K}\sum_{k=1}^{K} \max h_k^*(x^{\text{O}})$ from large to small. Here, we select anchor points with the top-T largest likelihood averaged over all K IM classifiers. Then, we warm up the OOM learner via minimizing cross-entropy loss $\min \frac{1}{T}\sum_{x^{\text{O}} \in \mathcal{A}_{\text{sorted}}} \text{CE}(h^{\text{O}}(x^{\text{O}}), \arg\max h_k^*(x^{\text{O}}))$. Additionally, we also warm up the connection by leveraging class-wise information. Specifically, we compute the cluster centroids for each modality via $\frac{1}{|\mathcal{C}_y|}\sum_{x_i^{\text{O}} \in \mathcal{C}_y := \{x^{\text{O}} : h^{\text{O}}(x^{\text{O}}) = y, y \in \mathcal{Y}\}} z_i^{\text{O}}$ and pair them to each IM centroid correspondingly. To this end, we can build up initial connections by following the VIB framework.

After the warm-up, we aim to further enhance both our connection and OOM exploration by creating emergent correspondence, as shown in Figure 4 (b). To tackle this, we map all IM data into the OOM feature space. If an OOM feature is close to all mappings $v_{k,i}, \forall k = \{1, \ldots, K\}$, then they can form a strong correspondence. Further, we select such OOM data as anchor points, which is further labeled the same as the corresponding IM data. Formally, we optimize OOM learners via

$$\min_{h^{\text{O}}} \mathcal{L}_{\text{uns}} := \frac{1}{|\mathcal{A}|}\sum_{(x_a^{\text{O}}, y) \in \mathcal{A}} \text{CE}(h^{\text{O}}(x_a^{\text{O}}), y) + \frac{1}{|\mathcal{A}| + |\mathcal{D}|}\sum_{x_d^{\text{O}} \in \mathcal{D}}\sum_{x_a^{\text{O}} \in \mathcal{A}} \text{KL}(h^{\text{O}}(x_d^{\text{O}}) \,\|\, h^{\text{O}}(x_a^{\text{O}}); h_1^*, h_2^*), \quad (15)$$

where $\mathcal{A}$ denotes the updated anchor points which are realized by sorting the Euclidean distance: $\mathcal{A} := \text{SORT}_S(\{(x_j^{\text{O}}, y_i^{\text{I}})\}_{j=1}^{M}, -\min_{i \in \{1,\ldots,N\}} \frac{1}{K}\sum_{k=1}^{K} \|z_j^{\text{O}} - v_{k,i}\|)$, where the first term computes the cross-entropy loss from the anchor points, and the second term calculates the KL divergence between the OOM data with modality disagreement and the OOM anchor points.

After these two steps, we can effectively tackle the unsupervised OOM generalization. In practice, we connect modalities and select anchor points in the feature space, and hence our application to both two scenarios can be efficient. In the next section, we carefully conduct extensive experiments to justify the effectiveness and extendibility of the proposed COX method under various settings.

4 EXPERIMENTS

In our experiments, we first elucidate the experimental details. Then, we provide performance comparisons to various baseline methods on different datasets. Finally, we conduct empirical analyses to provide an intuitive understanding of the proposed method.

4.1 IMPLEMENTATION DETAILS

Datasets. We consider datasets with at least three modalities: 1) TVL dataset (Fu et al., 2024) contains tactile sensing, RGB image, and class name which can be transformed into language; 2) LLVIP (Jia et al., 2021) dataset has infrared thermal data, RGB image, and annotations for pedestrian detection. We follow Zhu et al. (2023) to crop the pedestrian and background which stand for two classes. Further, we use the OpenAI template (Radford et al., 2021) to create language description; 3) NYU-D dataset (Silberman et al., 2012) contains RGB image, depth data, and class name that can be transformed into language description as well; 4) VGGS dataset (Chen et al., 2020a) includes video data, corresponding sound, and the language description; 5) MSR-VTT (Xu et al., 2016) includes videos and text description, we break down the videos into video frames and the audio data; 6) MOSEI dataset (Zadeh et al., 2018) contains videos from 7 classes of emotions, we extract audio data from the videos and use the emotion type to create language descriptions.

Models. We employ two types of IM perceptors, namely ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2023) which correspondingly contain 6 and 5 encoders to process different modalities. We select one modality for each experiment as OOM and then choose the rest as IM. For IM data, we use the existing encoders to extract their features. As for OOM data, we conduct preprocessing to ensure its compatibility. Then, we initialize an OOM learner from scratch using ViT-T/16 to learn from the OOM data using the guidance from IM perceptors. Note that for the TVL dataset, there are no existing encoders to process tactile modality. Therefore, when the tactile modality is chosen as IM data, we fine-tune the encoder using contrastive learning on the training

Table 2: Classification performance comparison of different methods across multiple datasets with different OOM modalities.

Setting	IM Perceptor	Method	TVL			LLVIP			NYU-D			VGGS		
			RGB	Lan	Tac	RGB	Lan	The	RGB	Dep	Lan	Aud	Vid	Lan
Semi-Supervised	ImageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		ERM	23.1	19.5	22.7	54.6	53.1	54.1	45.2	44.5	38.1	9.3	10.2	8.4
		EntMin	24.0	21.8	23.6	56.7	57.0	55.4	48.0	46.3	39.3	10.5	13.3	8.9
		COX	**31.2**	**25.3**	**26.5**	**59.2**	**58.3**	**58.3**	**52.3**	**50.7**	**44.2**	**16.8**	**18.4**	**11.7**
		aligned	79.5	29.8	35.8	65.4	61.8	63.4	61.8	54.0	52.7	27.8	29.3	19.1
	LanguageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		ERM	23.6	20.1	22.6	56.5	54.9	58.3	44.8	44.5	39.9	9.8	13.7	9.9
		EntMin	25.7	23.1	25.1	59.8	60.0	62.2	49.4	47.3	42.7	11.9	14.5	12.8
		COX	**33.5**	**26.3**	**27.3**	**61.2**	**62.3**	**66.4**	**58.8**	**53.5**	**48.4**	**18.3**	**22.1**	**13.4**
		aligned	81.6	31.2	38.3	74.1	73.2	87.2	68.6	65.1	57.7	38.6	32.5	20.9
Unsupervised	ImageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		SSL	6.3	4.3	5.1	52.3	56.1	52.4	14.6	13.6	18.9	2.5	6.9	3.8
		COX	**18.9**	**15.4**	**17.1**	**54.8**	**57.2**	**53.8**	**21.7**	**22.0**	**19.5**	**9.3**	**10.2**	**10.5**
		aligned	79.5	29.8	35.8	65.4	61.8	63.4	61.8	54.0	52.7	27.8	29.3	19.1
	LanguageBind	Random	0.4	0.3	0.2	48.2	47.3	51.0	10.2	11.3	10.2	0.3	0.3	0.3
		SSL	6.8	6.5	5.1	54.6	57.8	53.8	16.9	18.1	16.3	7.2	5.6	4.8
		COX	**19.3**	**19.2**	**18.6**	**55.0**	**56.4**	**55.7**	**24.5**	**23.1**	**20.4**	**10.0**	**11.6**	**10.4**
		aligned	81.6	31.2	38.3	74.1	73.2	87.2	68.6	65.1	57.7	38.6	32.5	20.9

set. For ImageBind, the tactile encoder is aligned with the image encoder, and for LanguageBind, it is aligned with the language encoder, which is the same as the original training process. For training the connection between modalities, we employ multi-layer perceptrons to realize the variational information bottleneck framework. Moreover, to obtain the optimal classifier from each in-modality, we utilize the extracted features and train a linear layer as classification heads.

Setup. We consider two scenarios of OOM generalization: For the semi-supervised case, we sample 10% of the training data as labeled data with each class having a balanced number of labels. For the unsupervised case, we have no labels at all. For selecting the number of anchor points, we choose the same number of examples for the warm-up and training phases, which is 10% of the total training set. To train the OOM learner, we use the Adam optimizer with an initial learning rate of $1e-3$ with weight decay $1e-5$, and train the model for 50 epochs.

Baseline methods. Since there is no existing baseline method to compare with under our setting, we implement four methods for comparison, namely: Random where the model is randomly initialized, ERM where only labeled data is used to minimize the empirical risk, EntMin (Grandvalet & Bengio, 2004) which minimize the entropy of unlabeled data meanwhile conduct ERM, SSL which conducts self-supervised learning using Gaussian noise perturbation on the input, and MoCo He et al. (2020) which updates model parameters with ensembling and meanwhile conducts contrastive learning. Note that we use MoCo for comparison for retrieval task in Table 3 because it is not for classification, and it is combined with EntMin in the semi-supervised case. Moreover, we use a pre-trained encoder as an upper-limit baseline "aligned". Next, we carefully compare the performance of our COX to these baseline methods.

4.2 Performance Comparison

For performance comparisons, we conduct classification and cross-modal retrieval to validate the proposed COX. There are seven modalities are considered, namely RGB image, language, tactile, thermal, depth, audio, and video which are simplified as RGB, Lan, Tac, The, Dep, Aud, and Vid, respectively. For each column, we choose one modality as OOM data, the rest modalities are selected IM data. For the retrieval task, we report the recall rate in both top 1 (R@1) and top 5 (R@5). The results are shown in Tables 2 and 3. We can see that the proposed COX clearly shows the best performance in both scenarios. Specifically, COX can achieve more than 5% performance improvement for most of the OOM setting, which justifies that leveraging the knowledge from IM perceptors can indeed help OOM generalization compared to using OOM data alone. Moreover, even though the performance is relatively limited compared to the fully pre-trained baseline under the unsupervised case, considering it is an extremely challenging setting, we can still largely improve the performance for over 10% compared to the Random baseline, which demonstrates that the unsupervised OOM generalization is indeed learnable further leads to a novel research direction for improving the gen-

Table 3: Cross-modal retrieval performance comparison of different methods across multiple datasets with different OOM modalities.

Setting	IM Perceptor	Method	MSR-VTT						MOSEI					
			Aud		Lan		Vid		Aud		Lan		Vid	
			R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Semi-Supervised	ImageBind	Random	5.4	25.1	5.0	25.4	5.4	24.2	14.3	42.5	14.4	42.8	14.1	42.1
		ERM	15.6	30.3	16.1	35.2	18.5	38.2	28.0	45.3	29.3	47.1	33.4	48.2
		EntMin	18.5	32.4	19.2	38.5	21.0	39.4	29.6	46.7	32.0	48.7	35.4	50.5
		MoCo	20.5	33.9	21.1	38.9	23.4	43.5	30.1	47.3	32.7	50.1	36.2	51.0
		COX	**23.3**	**35.8**	**23.4**	**39.1**	**26.5**	**48.8**	**32.4**	**48.0**	**33.8**	**50.4**	**38.8**	**53.7**
		Aligned	35.5	51.5	32.3	52.4	36.8	61.8	42.9	66.4	48.2	69.4	50.5	71.6
	LanguageBind	Random	5.2	24.3	5.4	25.1	5.0	25.6	13.5	43.1	14.2	42.7	14.6	41.9
		ERM	16.3	31.1	16.5	36.2	18.7	37.9	27.3	45.5	28.4	47.6	33.4	49.3
		EntMin	19.6	33.4	19.8	38.6	22.4	37.9	30.2	45.5	33.5	49.0	36.0	49.7
		MoCo	21.1	34.8	20.9	39.2	24.5	38.6	31.1	46.7	34.5	**50.5**	37.0	51.7
		COX	**25.2**	**36.0**	**24.1**	**40.0**	**28.7**	**49.5**	**34.6**	**49.8**	**34.6**	50.2	**39.2**	**55.4**
		Aligned	42.0	53.6	38.8	58.6	44.8	70.0	44.6	68.9	49.5	67.4	51.1	68.3
Unsupervised	ImageBind	Random	5.4	25.1	5.0	25.4	5.4	24.2	14.3	42.5	14.4	42.8	14.1	42.1
		SSL	8.9	28.4	9.3	28.1	10.1	29.5	17.4	48.8	16.2	45.2	16.0	45.0
		MoCo	9.2	28.9	9.5	28.4	10.6	30.0	17.8	50.3	16.6	45.8	17.1	44.4
		COX	**13.5**	**30.4**	**16.5**	**32.4**	**15.2**	**34.8**	**20.8**	**53.7**	**18.7**	**46.7**	**18.2**	**48.9**
		Aligned	35.5	51.5	32.3	52.4	36.8	61.8	42.9	66.4	48.2	69.4	50.5	71.6
	LanguageBind	Random	5.2	24.3	5.4	25.1	5.0	25.6	13.5	43.1	14.2	42.7	14.6	41.9
		SSL	9.2	28.9	11.0	28.8	10.3	28.7	18.0	48.9	18.4	45.0	17.8	45.6
		MoCo	9.6	29.4	11.1	28.5	11.0	29.3	18.8	50.7	18.5	45.2	18.0	45.5
		COX	**14.8**	**31.1**	**18.4**	**34.4**	**15.4**	**35.0**	**23.1**	**52.8**	**19.4**	**47.2**	**20.4**	**49.9**
		Aligned	42.0	53.6	38.8	58.6	44.8	70.0	44.6	68.9	49.5	67.4	51.1	68.3

eralization performance. Additionally, note that the performance of COX is affected by the quality of IM perceptors, as using LanguageBind shows relatively higher performance compared to using ImageBind. Thus, it would be potentially helpful to leverage sophisticated IM perceptors to benefit the generalization performance.

4.3 Empirical Analysis

To provide an intuitive justification for the proposed method, here we conduct empirical analyses using the MSR-VTT dataset on various OOM scenarios and modalities.

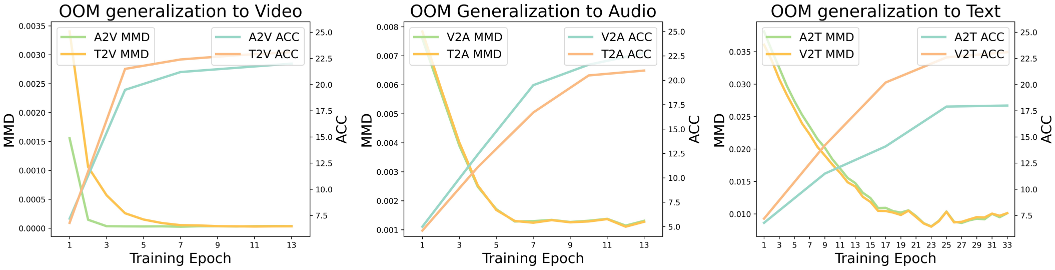

Figure 5: Connection effect on maximum mean discrepancy and accuracy across modalities.

Connection mitigates modality gap. To understand the performance of our VIB-based connection learning, here we show its effect on generalization out-of-modal. Specifically, during connection training, we compute the maximum mean discrepancy (MMD) between the mapping of each IM data and the OOM data. Meanwhile, we evenly select 6 points during the training and extract the IM mappings which are used to learn a classification head as the optimal classifier. Based on our theoretical result, we apply the classifiers to OOM data and compute their accuracies, as shown in Figure 5. We can see that as training goes on, the MMD between each IM mapping and OOM data is decreasing and the corresponding accuracy is increasing, which shows that: 1) our connection can indeed close the modality gap between their features and 2) as the mappings of IM data getting close to OOM data, the optimal classifier shows better classification results on OOM data, which benefits the knowledge transfer from known modalities to unknown ones.

Modality disagreement identifies uncertainty. To understand the effect of modality disagreement, we conduct an analysis under the semi-supervised scenario by training the OOM learner to use only labeled data for 10 epochs. Then, we leverage the modality disagreement criteria to separate OOM data into those with disagreement and agreement and show their prediction accuracies in

Figure 6 (a). We can see that the accuracy for OOM data with disagreement is significantly lower than those with agreement, meaning that the prediction uncertainty, *i.e.*, data with low accuracy, is effectively identified by the proposed modality disagreement.

Modality agreement alleviates uncertainty. Further, we conduct training by following the procedure proposed in Section 3.3 and again show the accuracies of OOM data with disagreement and agreement in Figure 6 (b). We can see that the performance gap between the two types of data is largely mitigated, which justifies the methodology of exploring OOM data using the guidance of modality agreement. As a result, we can achieve almost comparable performance on both types of data, benefiting the overall generalization performance.

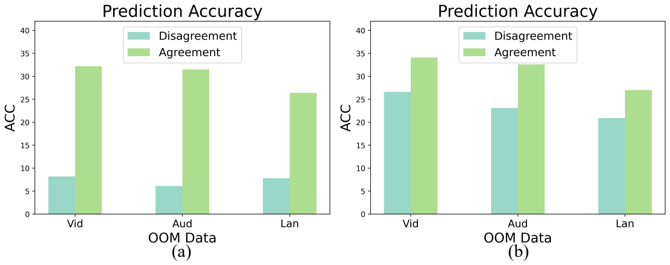

Figure 6: Prediction accuracy of OOM data with modality disagreement and modalities agreement, respectively. (a) Before exploration. (b) After exploration.

Ablation study. Additionally, we conduct an ablation study to justify the effect of our methodology. Specifically, we consider three ablations: 1) "w/o connection" where we remove the connection and directly apply the modality disagreement criteria on the original features of IM data and OOM data, 2) "w/o exploration" where we only leverage the OOM data with agreement for training, 3) For unsupervised scenario, we consider "w/o warm-up" where we do not conduct the warm-up phase and directly training the model. The results in Table 4 show that all modules are essential for achieving effective OOM generalization. Specifically, the connection is vital for the knowledge transduction of IM data to OOM data, without which the generalization performance is largely degraded. The conclusion is consistent with the connection analysis where directly applying optimal classifiers across modalities leads to poor accuracy. Moreover, removing exploration also hinders the performance because the uniqueness of OOM data is largely ignored. Additionally, we find that the warm-up phase is essential for the unsupervised case. As initialized models have no classification capability, we need pre-training to form basic feature clusters that are consistent with IM data, further enabling effective OOM generalization.

Table 4: Ablation study on various settings.

Setting	Ablation	MSR-VTT R@1		
		Aud	Lan	Vid
Semi	w/o connection	8.7	7.9	10.3
	w/o exploration	16.4	16.5	18.8
	COX	**25.2**	**24.1**	**40.0**
Unsup.	w/o warm-up	7.4	11.5	10.5
	COX	**14.8**	**18.4**	**15.4**

Discussion on computational efficiency. Note that we conduct the feature connection mostly on the feature space, the computational cost of training VIB framework work is quite acceptable. The main cost is training the OOM learner which is the basic training with cross-entropy loss optimization and can be implemented on a single NVIDIA 3090/4090 GPU.

5 Conclusion and Limitation

In this paper, we study a novel and promising research direction dubbed Out-of-Modal (OOM) Generalization which aims to leverage knowledge from existing modalities to generalize to an unknown modality without instance-level correspondence. We consider two scenarios where there are a few correspondences and there is no correspondence, *i.e.*, semi-supervised and unsupervised cases, respectively. To tackle these problems, we propose a connect & explore (COX) method which first learns connections across modalities to extract common knowledge and then explores the unique knowledge of OOM data based on modality disagreement. Extensive experiments are conducted to justify the proposed method and intuitive insights are provided to inspire future studies. However, our research is limited to several aspects which we hope to address in the future. First, although challenging as it is, the performance is relatively limited compared to fully-aligned models, which requires more investigations to enhance generalization. Second, our OOM generalization is mostly conducted within the modalities from the same dataset. In the future, we hope to discover scenarios where the OOM data is from a different dataset with a large modality gap.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Lincan Cai, Shuang Li, Wenxuan Ma, Jingxuan Kang, Binhui Xie, Zixun Sun, and Chengwei Zhu. Enhancing cross-modal fine-tuning with gradually intermediate modality generation. *arXiv preprint arXiv:2406.09003*, 2024.

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)*, pp. 721–725. IEEE, 2020a.

Jiaao Chen, Zichao Yang, and Diyi Yang. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*, 2020b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36:78674–78695, 2023.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. Dynamic multimodal information bottleneck for multimodality classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7696–7706, 2024.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=tFEOOH9eH0.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Heidi E Harley, Erika A Putman, and Herbert L Roitblat. Bottlenose dolphins perceive object features through echolocation. *Nature*, 424(6949):667–669, 2003.

Masukichi Hashimoto. Origin of the compass. *Memoirs of the Research Department of the Toyo Bunko (The Oriental Library)*, 1:69–92, 1926.

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*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956, 2021.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3496–3504, 2021.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: Robot manipulation with multimodal prompts. In *International Conference on Machine Learning*, pp. 14975–15022. PMLR, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov. Cross-modal generalization: Learning in low resource modalities via meta-alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2680–2689, 2021.

Paul Pu Liang, Chun Kai Ling, Yun Cheng, Alex Obolenskiy, Yudong Liu, Rohan Pandey, Alex Wilf, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal learning without labeled multimodal data: Guarantees and applications. *arXiv preprint arXiv:2306.04539*, 2023.

Zhou Lu. A theory of multimodal learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7598–7607, 2019.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning*, pp. 689–696, 2011.

Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Muhammad Arifur Rahman, David J Brown, Nicholas Shopland, Andrew Burton, and Mufti Mahmud. Explainable multimodal machine learning for engagement analysis by continuous performance test. In *International Conference on Human-Computer Interaction*, pp. 386–399. Springer, 2022.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339–9347, 2019.

Junhong Shen, Liam Li, Lucio M Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: Align then refine. In *International Conference on Machine Learning*, pp. 31030–31056. PMLR, 2023.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *COLT*, number 114, pp. 403–414, 2008.

Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 171–188. Springer, 2020.

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

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

Ying Wang, Tim GJ Rudner, and Andrew G Wilson. Visual explanations of image-text representations via multi-modal information bottleneck attribution. *Advances in Neural Information Processing Systems*, 36:16009–16027, 2023.

Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng Huang, Yang Zhao, Tao Jin, Peng Gao, et al. Freebind: Free lunch in unified multimodal space via knowledge fusion. In *Forty-first International Conference on Machine Learning*, 2024.

LA Weinstein. Electromagnetic waves. *Radio i svyaz', Moscow*, 1988.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. *arXiv preprint arXiv:2206.06487*, 2022.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *ICML*, 2023.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Rowan Zellers, Prithviraj Ammanabrolu, Ronan Le Bras, Gunhee Kim, et al. Fusing pre-trained language models with multimodal prompts through reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10845–10856, 2023.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10394–10403, 2019.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

A Appendix

A.1 Lower Bound of Our VIB framework

Recall that we have the following factorization:

$$p(X^{\mathrm{I}}, X^{\mathrm{O}}, V, Y) = p(V, Y|X^{\mathrm{O}}, X^{\mathrm{I}})p(X^{\mathrm{O}}|X^{\mathrm{I}})P(X^{\mathrm{I}}), \tag{16}$$

with Markov chains $V \leftrightarrow X^{\mathrm{I}} \leftrightarrow X^{\mathrm{O}}$ and $X^{\mathrm{I}} \leftrightarrow Y \not\leftrightarrow X^{\mathrm{O}}$. Our goal is to maximize the information redundancy (Liang et al., 2023; Williams & Beer, 2010):

$$\max I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) = I(X^{\mathrm{O}}; X^{\mathrm{I}}) - I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y), \tag{17}$$

where the first term is lower-bounded by:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}) \geq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v) p(v|x^{\mathrm{I}}), \tag{18}$$

Then, we consider the second term $I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y)$:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y) = \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log \frac{p(x^{\mathrm{O}}, x^{\mathrm{I}}|y)}{p(x^{\mathrm{O}}|y) p(x^{\mathrm{I}}|y)} \tag{19}$$

$$= \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) \log \frac{p(x^{\mathrm{O}}, x^{\mathrm{I}}, y)}{p(y|x^{\mathrm{O}})} - H(Y) + H(Y|X^{\mathrm{I}}) + H(X^{\mathrm{O}}) + H(X^{\mathrm{I}}). \tag{20}$$

Note that we use the factorization $p(x^{\mathrm{O}}, x^{\mathrm{I}}, y) = p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}})$, and further ignore the entropy terms[4], then we have:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}|Y) \leq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) \log p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) - \log h(y|x^{\mathrm{O}}), \tag{21}$$

which is based on the positivity of KL divergence between our classifier $h(y|x^{\mathrm{O}})$ and $p(y|x^{\mathrm{O}})$.

To this end, we can lower-bound our objective by combining Eqs. 18 and 21:

$$I(X^{\mathrm{O}}; X^{\mathrm{I}}; Y) \geq \int dx^{\mathrm{O}} dx^{\mathrm{I}} dv p(x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(v|x^{\mathrm{I}}) \log q(x^{\mathrm{O}}|v) p(v|x^{\mathrm{I}}) \tag{22}$$

$$- \int dx^{\mathrm{O}} dx^{\mathrm{I}} dy p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) \log p(y|x^{\mathrm{I}}) p(x^{\mathrm{O}}|x^{\mathrm{I}}) p(x^{\mathrm{I}}) + \log h(y|x^{\mathrm{O}}) = \mathcal{L}_{con}. \tag{23}$$

A.2 Proof of Theorem 1

Proof.

Assumption 1 (Relaxed triangle inequality). For the distance function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, there exists $c_d \geq 1$ such that $\forall \hat{y}_1, \hat{y}_2, \hat{y}_3 \in \hat{\mathcal{Y}} d(\hat{y}_1, \hat{y}_2) \leq c_d(d(\hat{y}_1, \hat{y}_3) + d(\hat{y}_2, \hat{y}_3))$.

Assumption 2 (Inverse Lipschitz condition). For the function d, it holds that $\forall h$, $\mathbb{E}[d(h(x_1, x_2), h^*(x_1, x_2))] \leq |\mathcal{L}(h) - \mathcal{L}(h^*)|$, where h^* is the Bayes optimal classifier on both x_1 and x_2; and $\mathbb{E}[d(h(x), h^*(x))] \leq |\mathcal{L}(h) - \mathcal{L}(h^*)|$, where h^* is the Bayes optimal classifier on x.

Assumption 3 (Classifier optimality). For any classifiers h in comparison to the Bayes' optimal classifier h^*, there exists constants $\epsilon > 0$ such that $|\mathcal{L}(h) - \mathcal{L}(h^*)|^2 \leq \epsilon$.

To bridge h_1^* and h_2^*, we use $h_{1,2}^*$ and h^* to denote the Bayes' optimal classifier on both IM data and all data, respectively. Then, we capture the relationship between the uniqueness of OOM data given both IM data and the difference in their Bayes' optimal prediction errors:

$$|\mathcal{L}(h_{1,2}^*) - \mathcal{L}(h^*)|^2 = |\mathbb{E}_X \mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{X_1^{\mathrm{I}}, X_2^{\mathrm{I}}} \mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}, X_2^{\mathrm{I}}), y)|^2 \tag{24}$$

$$\leq |\mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}, X_2^{\mathrm{I}}), y)|^2 \tag{25}$$

$$\leq \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}})) \tag{26}$$

$$\leq \mathbb{E}_X \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}})) \tag{27}$$

$$= I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). \tag{28}$$

[4] We focus on the optimization of $p(Y|X^{\mathrm{O}})$, and $p(Y|X^{\mathrm{I}})$ is given and frozen in our setting.

Then, we first capture the redundancy between one IM data and OOM data given another IM data:

$$
\begin{aligned}
|\mathcal{L}(h_1^*) - \mathcal{L}(h^*)|^2 &= |\mathbb{E}_X \mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{X_1^{\mathrm{I}}} \mathbb{E}_{Y|X_1^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}), y)|^2 && (29)\\
&\leq |\mathbb{E}_{Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, X^{\mathrm{O}}} \ell(h^*(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}), y) - \mathbb{E}_{Y|X_1^{\mathrm{I}}} \ell(h_1^*(x_1^{\mathrm{I}}), y)|^2 && (30)\\
&\leq \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}})) && (31)\\
&\leq \mathbb{E}_X \mathrm{KL}(p(y|x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}}) \parallel p(y|x_1^{\mathrm{I}})) && (32)\\
&= I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}). && (33)
\end{aligned}
$$

Further leveraging triangle inequality through the Bayes' optimal classifier h^* and the inverse Lipschitz condition, we have:

$$
\begin{aligned}
\mathbb{E}_{p(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}})}[d(h_1^*, h_{1,2}^*)] &\leq \mathbb{E}_{p(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}})}[d(h_1^*, h^*)] + \mathbb{E}_{p(x_1^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}})}[d(h^*, h_{1,2}^*)] && (34)\\
&\leq |\mathcal{L}(h_1^*) - \mathcal{L}(h^*)|^2 + |\mathcal{L}(h^*) - \mathcal{L}(h_{1,2}^*)|^2 && (35)\\
&\leq I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}). && (36)
\end{aligned}
$$

Symmetrically, we can have $|\mathcal{L}(h_2^*) - \mathcal{L}(h^*)|^2 \leq I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}})$ and further obtain $\mathbb{E}_{p(x_2^{\mathrm{I}}, x_2^{\mathrm{I}}, x^{\mathrm{O}})}[d(h_2^*, h_{1,2}^*)] \leq I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}})$. Then combining with Eq. 36:

$$
\mathbb{E}_{p(x_1^{\mathrm{I}}, x_2^{\mathrm{I}})}[d(h_1^*, h_2^*)] \leq I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + 2I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}) \qquad (37)
$$

Finally, based on the decomposition of the task-related mutual information of X^{O}: $I(X^{\mathrm{O}}, Y) = I(X^{\mathrm{O}}, X_2^{\mathrm{I}}, Y|X_1^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, Y|X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}) + I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y)$, as shown in Figure 3, we can achieve:

$$
\alpha(h_1^*, h_2^*) := \mathbb{E}_{p(x_1^{\mathrm{I}}, x_2^{\mathrm{I}})}[d(h_1^*, h_2^*)] \leq I(X^{\mathrm{O}}, Y) - I(X^{\mathrm{O}}, X_1^{\mathrm{I}}, X_2^{\mathrm{I}}, Y) + I(X^{\mathrm{O}}, Y|X_1^{\mathrm{I}}, X_2^{\mathrm{I}}), \qquad (38)
$$

□

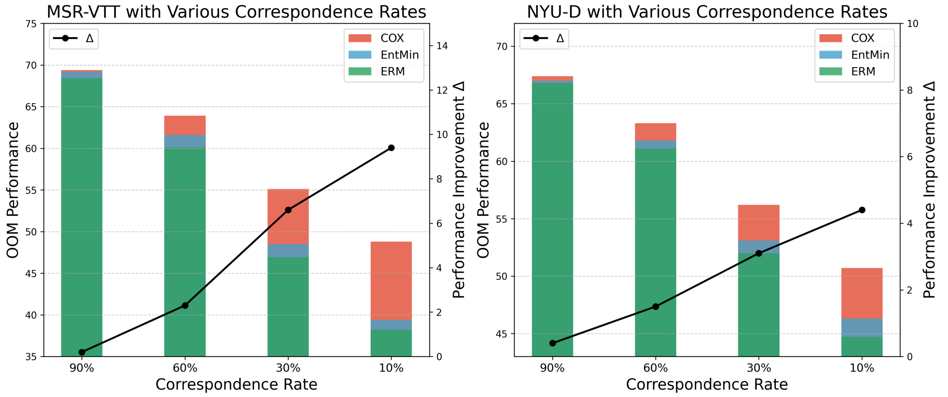

Figure 7: Performance benefits brought by COX under various correspondence rate in OOM data.

A.3 Additional Experiments

We conduct additional experiments to further justify the proposed COX. First, we study the performance benefits brought by COX under various correspondence rates in OOM data. Specifically, we choose MSR-VTT and NYU-D datasets and use Vid and Dep as OOM modalities, respectively, and show the result in Figure 7. First of all, we observe that COX brings more benefits when correspondence is more scarce. This is because sufficient correspondence can maximally uncover the knowledge of OOM data. As correspondence gets less, the knowledge that can be explored from correspondence decreases. However, COX leverages the knowledge from IM data which brings more benefits even with less correspondence. Thus, the increased benefits of COX under

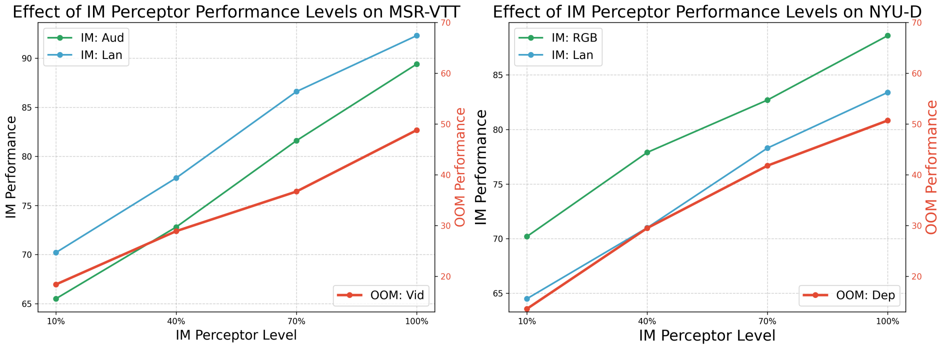

Figure 8: Effect of Varying IM Perceptor Performance Level.

low-correspondence scenarios demonstrate its effectiveness in tackling OOM generalization without correspondence.

Moreover, we testify how varied performance levels of IM perceptors could affect the OOM performance. To achieve this, we change the number of IM data in each dataset as 10%, 40%, 70%, and 100%, and test the OOM performance of COX, as shown in Figure 8. We can see that the OOM performance is significantly affected by the accuracy level of IM perceptors. When the performance of IM perceptors improves, the OOM performance of COX is also enhanced. Therefore, improving the performance of IM perceptors is vital for OOM generalization using COX.

MSR-VTT	Vision	NYU-D	Language
FreeMatch	45.2	MixText	21.2
COX	**52.3**	COX	**23.4**

Table 5: Comparison with competitive uni-modal methods from Vision and Language.

Setting	Method	MSR-VTT	NYU-D
Unsup	MoCo	30.0	15.7
	MoCo+COX	**35.4**	**23.8**

Table 6: Combining COX with MoCo for knowledge extraction from OOM to IM data.

Further, to understand the contribution of COX on uni-modal study, we conduct comparison and combination with uni-modal methods. First, we consider two uni-modalities vision and language from MSR-VTT and NYU-D datasets, respectively. By comparing to FreeMatch Wang et al. (2022) and MixText Chen et al. (2020b), two competitive semi-supervised learning methods that correspondingly deal with vision and language data, we show the performance of COX in Table 5. Even though the two baselines were effective under their original setting, their performance is still limited when applied to challenging multi-modal datasets with scarce knowledge. As we can see, COX still shows very effective performance compared to them, again justifying the benefits from COX by leveraging IM data.

Then, we consider combining COX with the well-known unsupervised method MoCo He et al. (2020) and show the performance benefits brought by COX for enhancing unknown modality. We show the result in Table 6. Delightfully, we observe significant performance improvement on both Vid from MSR-VTT and RGB from the NYU-D dataset. This is because COX unleashes the potential label information from IM data to enhance label prediction of OOM data, i.e., Vid and RGB here. Such a finding implies that COX can extract knowledge from other modalities to enhance new ones, which is the main goal of this study.