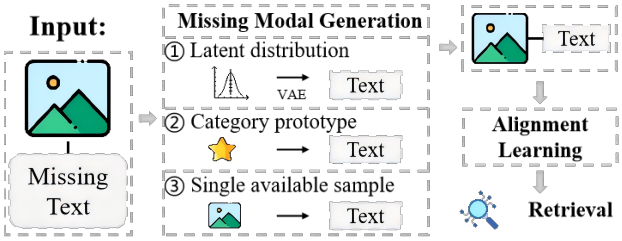
Causality-Aligned Semantic Recovery for Incomplete Cross-Modal Retrieval

Haipeng Chen[1, 2], Yu Liu[1*], Xun Yang[3], Yuheng Liang[1], Yingda Lyu[4]

[1]College of Computer Science and Technology, Jilin University, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3]University of Science and Technology of China, Hefei, China
[4]Public Computer Education and Research Center, Jilin University, China
{yul20, yhl24}@mails.jlu.edu.cn, xyang21@ustc.edu.cn, {chenhp, ydlv}@jlu.edu.cn

Abstract

Incomplete cross-modal retrieval (ICMR) requires models to recover missing modalities and robustly align heterogeneous ones for effective retrieval. Existing methods, however, fall short in both aspects. They often rely on limited semantic cues, such as single samples or coarse category prototypes, which compromises reconstruction quality. Moreover, these approaches are vulnerable to learning spurious cross-modal correlations, thereby impairing accurate alignment and hindering retrieval performance. To address these challenges, we propose **C**ausality-**A**ligned **S**emantic **R**ecovery (CASR), a novel method designed to both comprehensively restore missing modalities and mitigate spurious associations between vision and language. Our CASR involves two essential components: i) the Missing Modality Imagination (MMI) module, which combines category semantic priors with relevant contextual information to achieve high-quality semantic reconstruction; ii) the Explicit Causal Alignment (ECA) module, which explicitly learns environment-invariant attention, effectively eliminating the interference of spurious correlations and improving retrieval performance. Furthermore, we extend CASR to the challenging task of Partially Aligned Cross-Modal Retrieval, where we treat unlabeled unpaired data as a form of incomplete data. By leveraging MMI and ECA modules, we are able to learn robust representations in this setting. Extensive experiments on benchmark datasets under various missing rates demonstrate that CASR achieves superior robustness and retrieval performance.



(a) Modal recovery is insufficient with existing methods.

(b) Biased retrieval results caused by spurious correlations.

Figure 1: Limitations of existing ICMR methods.

1 Introduction

Cross-modal retrieval (CMR) is a fundamental task in multimodal learning, aiming to bridge heterogeneous data for applications like visual matching (Yang et al. 2024a) and audio-video understanding (Shvetsova et al. 2022). Existing CMR methods predominantly rely on fully paired data, an assumption that is often violated in real-world scenarios due to sensor failures, data collection costs, or privacy constraints. This data incompleteness severely degrades performance. Thus, Incomplete Cross-Modal Retrieval (ICMR) has become a critical and rapidly growing research area.

Within the ICMR paradigm, training data comprises partially paired visual-textual instances as well as samples containing only a single modality. Achieving robust ICMR hinges on two critical requirements: 1) the comprehensive semantic reconstruction of missing modalities, and 2) the accurate alignment of heterogeneous modalities. Recent studies (Jing et al. 2020; Zeng et al. 2021; Shi et al. 2024) have explored solutions from both perspectives. For semantic recovery of missing modalities, existing ICMR methods typically follow three strategies (see Fig.1(a)): latent distribution modeling, category prototypes, or direct imputation from a single available modality. For instance, DAVAE (Jing et al. 2020) uses a Variational Auto-Encoder (VAE) (Kingma and Welling 2013) to generate missing modalities by modeling latent distributions. Methods like PAN (Zeng et al. 2021), SPAL (Wang et al. 2024b), and OTPAL (Wang et al. 2024a) learn category prototypes that serve as a bridge the missing and available modalities, facilitating cross-modal alignment. DCT (Shi et al. 2024) directly fills in the missing modality from its available counterpart. Despite some progress, ex-

*Corresponding Author.

isting methods still struggle to adequately reconstruct the semantic information of missing modalities. This is because latent distribution models can introduce noise, while prototype-based and single-sample methods capture only partial semantic information, leading to impoverished representations. Beyond reconstruction, a major challenge lies in robust cross-modal alignment. Existing methods commonly adopt Empirical Risk Minimization (ERM) (Arjovsky et al. 2019) for training, which is susceptible to dataset biases and prone to learning spurious correlations between visual and textual modalities. As illustrated in Fig.1(b), due to the frequent co-occurrence of "person" and "bottle" in the training data, a model might erroneously retrieve images of a "person" when the query text only mentions "bottle". This demonstrates how spurious correlations can severely impair retrieval performance. These limitations highlight a fundamental research challenge: "*How can we fully recover the semantic content of missing modalities while simultaneously eliminating spurious cross-modal correlations to achieve robust ICMR?*"

To answer this question, we propose Causality-Aligned Semantic Recovery (CASR), a robust ICMR method comprising two key components: the Missing Modality Imagination (MMI) module and the Explicit Causal Alignment (ECA) module. The design of our MMI module is inspired by the human cognitive ability to infer missing information using category knowledge and contextual cues. Specifically, MMI fuses rich category semantic features generated by a large language model with fine-grained contextual information from similar samples to approximate the reconstruction of the missing modality representation. This approach allows for the reconstruction of missing modality representations that preserve global semantics while incorporating local details, thereby achieving high semantic completeness and representational fidelity. To mitigate spurious cross-modal correlations, the ECA module learns environment-invariant representations. It re-weights training samples to simulate diverse environments and extracts causal features that are consistent across them. This causal learning paradigm enables robust and reliable cross-modal alignment, free from dataset biases. Furthermore, we demonstrate the versatility of our CASR by extending it to the challenging task of Partially Aligned Cross-Modal Retrieval. In this setting, we treat each unpaired sample as two types of modality-missing scenarios: images without corresponding texts, and texts without corresponding images. We employ the MMI module for modality recovery and the ECA module for robust cross-modal matching. Our main contributions are summarized as follows:

- We propose a Causal-Aligned Semantic Recovery framework that enhances retrieval robustness under incomplete data by achieving fully semantic recovery of missing modalities and precise cross-modal alignment.
- To restore the missing semantics, the MMI module simulates human cognition by integrating the category semantic prior and the contextual prior of similar samples to guide the reasonable reconstruction.
- To enhance cross-modal semantic alignment, the ECA module suppresses spurious correlations by explicitly learning environment-invariant embedding based on causal inference.
- We demonstrate the effectiveness and robustness of CASR with different missing rates on five benchmark datasets through extensive experiments.

2 Related Work

Incomplete Cross-Modal Retrieval (ICMR) differs from traditional cross-modal retrieval (Liu et al. 2024, 2025a,b) in that it focuses on achieving robust retrieval when modalities are partially damaged or missing. Existing ICMR methods typically follow a two-step paradigm: recovering the missing modality semantics and performing cross-modal alignment. For instance, DAVAE (Jing et al. 2020) employs variational autoencoders to model latent distributions for generating missing modalities. PAN (Zeng et al. 2021) introduces semantic category prototypes to learn cross-modal invariant representations. DCT (Shi et al. 2024) leverages adjacent semantic correlations for retrieval, while SPAL (Wang et al. 2024b) and OTPAL (Wang et al. 2024a) construct shared semantic prototypes to associate data across modalities. Despite their progress, these approaches often rely on individual sample information or coarse prototypes that capture only partial semantics, leading to incomplete modality recovery. Moreover, they are vulnerable to biases in the training data and prone to learning spurious cross-modal correlations. In contrast, our CASR achieves more comprehensive semantic recovery of missing modalities while eliminating spurious correlations from a causal perspective, thereby significantly improving retrieval accuracy and robustness.

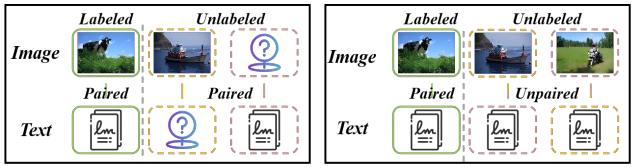
Causal Inference. Traditional multimodal methods often rely heavily on Empirical Risk Minimization (ERM), emphasizing statistical correlations among variables while overlooking the underlying causal relationships, which weakens model interpretability and robustness. To address this limitation, Causal Inference (Pearl, Glymour, and Jewell 2016) has recently been introduced into multimodal research as an effective debiasing mechanism that uncovers latent causal structures in data. It has shown promise in tasks such as self-supervised learning (Wang et al. 2021), video moment retrieval (Yang et al. 2021), medical image segmentation (Song et al. 2025a,b), and person re-identification (Liu et al. 2024). In contrast to existing ERM-based approaches for ICMR, our CASR explicitly learns cross-environment invariant embedding guided by causal inference. This strategy helps eliminate spurious correlations between visual and textual modalities, enabling the model to more accurately capture true cross-modal relationships and significantly enhancing its interpretability and robustness.

3 Method

Problem Formulation

Our CASR aims to solve the common incomplete cross-modal retrieval problem in real-world scenarios, where some data lack corresponding other modalities (as shown in Fig.2(a)). We define the training dataset as $\mathcal{D} = \{\mathcal{D}^f, \mathcal{D}^m\}$. Here, $\mathcal{D}^f = \{(\mathbf{x}_i^v, \mathbf{x}_i^t, y_i)\}_{i=1}^{n_f}$ denotes complete instances,

where $\mathbf{x}_i^v$, $\mathbf{x}_i^t$, y_i and n_f are visual embedding, text embedding, class label of the i-th instance and total number of complete instances, respectively. $\mathcal{D}^m = \{(\mathbf{x}_i^v, -, y_i) \vee (-, \mathbf{x}_i^t, y_i)\}_{i=1}^{n_m}$ is a modality-incomplete subset, where "–" indicates a missing modality and n_m is the number of incomplete instances. In addition, we also studied partially aligned cross-modal retrieval, as shown in Fig.2(b), in which there is a portion of unlabeled unpaired data. We treat these samples as modality-incomplete data and apply CASR to learn robust retrieval representations.



(a) Incomplete training data (b) Partially aligned training data

Figure 2: Two types of incomplete data addressed by CASR.

Our Proposed Method: CASR

Missing Modality Imagination

As noted in Sec. 1, existing methods often fail to fully recover missing modality semantics, impairing cross-modal alignment. Thus, we propose a Missing Modality Imagination (MMI) module that reconstructs the missing modality using semantic cues from related content. As shown in Fig.3, MMI contains prior retriever, prior fusion and noise filter.

(1) Prior Retriever. When humans attempt to infer missing modality information, they typically rely on two key cues: category-level knowledge and contextual similarity semantics. Inspired by this cognitive mechanism, we design our retrieval model to mimic such "reasoning" by constructing two prior memories: a category prior embedding memory $\mathbf{B}_p$ and a context prior embedding memory $\mathbf{B}_c$. Specifically, for each category, we first prompt GPT-4o (Hurst et al. 2024) with "Please generate an attribute-rich description for [class name]", obtaining a detailed textual description. We then feed these descriptions into a frozen large language model (LLM) to extract the category prior embedding $\mathbf{B}_p$. Leveraging the LLM's strengths in semantic understanding and contextual modeling, $\mathbf{B}_p$ effectively captures the underlying semantic structure and fine-grained distinctions between categories, offering high-quality semantic prior support for subsequent modality recovery and cross-modal alignment. Additionally, we leverage the complete instances from the current batch to build $\mathbf{B}_c$. We exploit available modality to retrieve valuable information from $\mathbf{B}_p$ and $\mathbf{B}_c$. Take text missing as an example, we first use the image representation $\mathbf{x}_i^v$ as a query to retrieve the most similar category prototype $\mathbf{p}_i \in \mathbb{R}^{1\times d}$ from $\mathbf{B}_p$, where $d = 1024$ is the feature dimension. Then, we use $\mathbf{x}_i^v$ to compute the cosine similarity with the elements in $\mathbf{B}_c$ to identify the top-K similar visual representations $S_i^v = \{\mathbf{s}_i^v\}_{i=1}^K$ and their corresponding textual representations $S_i^t = \{\mathbf{s}_i^t\}_{i=1}^K$. The top-$K$ retrieved instances offer auxiliary contextual cues to facilitate the recovery of the missing modality.

(2) Prior Fusion and Noise Filter. To ensure the integrity of semantic restoration, MMI module optimizes content restoration by integrating multi-source auxiliary information. Specifically, we employ a cross-attention block to establish interactions among $\mathbf{p}_i$, S_i^v, and S_i^t, constructing the text-level representation $\tilde{\mathbf{t}}_i \in \mathbb{R}^{1\times d}$ as:

$$\tilde{\mathbf{t}}_i = \text{Softmax}\left(\frac{f_t^Q(\mathbf{p}_i) f_t^K(S_i^v)^\top}{\sqrt{d}}\right) f_t^V(S_i^t), \quad (1)$$

where $f_t^Q(\cdot)$, $f_t^K(\cdot)$, $f_t^V(\cdot)$ denote the query, key, and value functions for the text modality. In addition, considering the potential noise in the text representation $\tilde{\mathbf{t}}_i$, we introduce Fast Fourier Transform to filter the noise in the frequency domain, which can be defined as:

$$\mathbf{t}_i = \mathcal{F}^{-1}(\mathbf{f} \odot \mathcal{F}(\tilde{\mathbf{t}}_i)), \quad (2)$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the Fourier transform and its inverse, respectively, and $\odot$ represents the element-wise (Hadamard) product. $\mathbf{f}$ acts as a learnable filter. In addition, when the image modality is missing, the above process can be used to obtain the corresponding visual embedding $\mathbf{v}_i$.

Explicit Causal Alignment

After the MMI module recovers the missing modality, ICMR task learns the correspondence between visual modality V and textual modality L and infers retrieval result R accordingly. This process can be formulated as:

$$P(R|V, L). \quad (3)$$

To maximize probability $P(R|V, L)$, we encourage cross-modal representations of the same instance to be as similar as possible. To achieve this, we introduce a consistency loss:

$$\mathcal{L}_{con}(\mathbf{v}_i, \mathbf{t}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\mathbf{v}_i \mathbf{t}_i^\top}{\|\mathbf{v}_i\| \cdot \|\mathbf{t}_i\|}\right), \quad (4)$$

where $\mathbf{v}_i$ and $\mathbf{t}_i$ are the common representations of the i_{th} sample. However, existing methods directly learning $P(R \mid V, T)$ inevitably capture spurious correlations (*c.f.*, Sec.1). As Fig.3 illustrates, we employ a Structural Causal Model (SCM) (Pearl, Glymour, and Jewell 2016) to analyze dataset bias. This reveals a confounder Z obscuring the true causal pathways $V \rightarrow R$ and $L \rightarrow R$, inducing spurious correlations $V \leftarrow Z \rightarrow L$. To mitigate this confounding influence, we propose an Explicit Causal Alignment (ECA) module.

While collecting data uniformly across diverse environments would eliminate the confounding effect of Z, it is often costly and impractical (Arjovsky et al. 2019). Fortunately, by implementing causal intervention through the do calculus $do(\cdot)$, we can learn $P^* = P(R \mid do(V), do(L))$ by blocking the confounding paths $Z \rightarrow V$ and $Z \rightarrow L$, thereby eliminating Z as a confounder (see Fig.3). Thus, we stratify Z into $Z = \{z_1, z_2, ..., z_k\}$, where each z_i represents a distinct contextual environment. Given an instance $(\mathbf{v}_i, \mathbf{t}_i)$, the de-confounded model is formulated as follows:

$$P^* = \sum_{z_j \in Z}[P(z_j)P(R|f^v(\mathbf{v}_i), f^t(\mathbf{t}_i))], \quad (5)$$

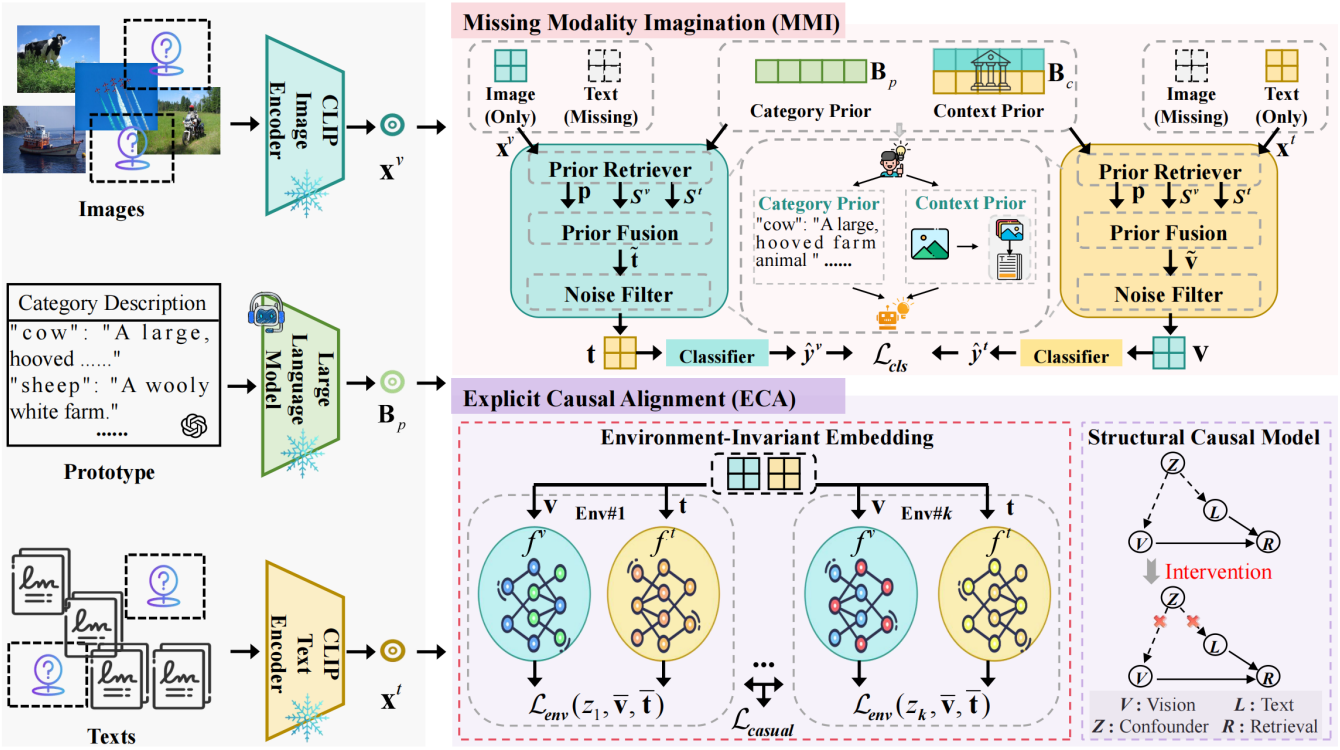


Figure 3: Overview of the proposed CASR framework, which integrates two key modules: MMI for recovering missing modality semantics and ECA for reinforcing causal cross-modal alignment, jointly enhancing retrieval robustness.

where $f^v(\cdot)$ and $f^t(\cdot)$ represent the functions that produce the final visual and text embedding under dataset contextual environment z_j, respectively. By adopting the cross-modal consistency loss $\mathcal{L}_{con}$, maximizing the probability of $P^* = P(R|do(V), do(L))$ is equivalent to minimizing $\mathcal{L}_{con}$:

$$\mathcal{L}_{bd} = \sum_{z_j \in Z} \sum_{(\mathbf{v}_i, \mathbf{t}_i) \in z_j} \mathcal{L}_{con}(f^v(\mathbf{v}_i), f^t(\mathbf{t}_i)). \quad (6)$$

Eq.(6) poses two challenges: 1) how to obtain environments z_j with different biases; and 2) how to learn visual embedding function $f^v(\cdot)$ and text embedding function $f^t(\cdot)$.

To address challenge 1), we re-weight training samples to simulate environments with different distribution biases. As each sample carries unique contextual cues, this re-weighting effectively shifts the dataset's contextual distribution (Deng and Zhang 2022). The loss in environment z_j in Eq.(6) can thus be expressed as:

$$\mathcal{L}_{env}(z_j) = \sum_{(\mathbf{v}_i, \mathbf{t}_i, w_{ij}) \in z_j} w_{ij} * \mathcal{L}_{con}(f^v(\mathbf{v}_i), f^t(\mathbf{t}_i)), \quad (7)$$

where w_{ij} is the weight of instance $(\mathbf{v}_i, \mathbf{t}_i)$ in environment z_j. We set the sum of weights for a instance across all k environments to 1: $\sum_{j=1}^{k} w_{ij} = 1$. Causal alignment should avoid spurious correlations and remain invariant across different environments. To this end, inspired by Invariant Risk Minimization (Arjovsky et al. 2019), we extend Eq.(6) by adding a constraint that there exists an optimal classifier that is invariant and effective across environments:

$$\mathcal{L}_{inv}((\mathbf{v}_i, \mathbf{t}_i)) = \sum_{z_j \in Z} \big[\mathcal{L}_{env}(z_j, f^v(\mathbf{v}_i), f^t(\mathbf{t}_i)) + \alpha * \|\nabla_w|_{w=1.0} \mathcal{L}_{env}(z_j, f^v(\mathbf{v}_i), f^t(\mathbf{t}_i))\|^2\big], \quad (8)$$

where α is the trade-off hyper-parameter and $w = 1.0$ is a "dummy" classifier to calculate the gradient penalty term.

Next we need to address challenge 2), which is to design $f^v(\cdot)$ and $f^t(\cdot)$ to learn the final embedding of vision and text, respectively. Since causal embedding lacks explicit annotations, we introduce a Graph Attention Network (GAT) (Veličković et al. 2017) for each modality to mine latent causal embedding and effectively filter out irrelevant information that can easily lead to spurious correlations. In this process, samples from each modality are treated as graph nodes, with adjacency relations used to identify relative neighbors and construct local graphs. Thus, the feature learning process is formulated as:

$$\overline{\mathbf{v}}_i = f^v(\mathbf{v}_i; \theta^v), \quad \overline{\mathbf{t}}_i = f^t(\mathbf{t}_i; \theta^t), \quad (9)$$

where θ^v and θ^t represent the learnable parameters of $f^v(\cdot)$ and $f^t(\cdot)$, respectively. To effectively learn causal embedding, attention should be environment invariant, as its goal is to learn causal embedding that are not disturbed by environmental changes. This goal can be achieved by penalizing the differences in attention parameters under different environments. Integrating the visual and textual attention modules

and the environment invariance constraint into Eq.(8) yields:

$$\mathcal{L}_{causal}(\overline{\mathbf{v}}_i, \overline{\mathbf{t}}_i) = \sum_{z_j \in Z} \Big[\mathcal{L}_{env}(z_j, \overline{\mathbf{v}}_i, \overline{\mathbf{t}}_i) + \alpha * \|\nabla_w|_{w=1.0} \mathcal{L}_{env}(z_j, \overline{\mathbf{v}}_i, \overline{\mathbf{t}}_i)\|^2 + \beta * \Big(\Big\|\theta_j^v - \hat{\theta}^v\Big\|^2 + \Big\|\theta_j^t - \hat{\theta}^t\Big\|^2\Big)\Big], \quad (10)$$

where β is the balance coefficient. $\hat{\theta}^v$ and $\hat{\theta}^t$ denote the average attention parameters learned across different environments. During inference, each modality uses a single attention net with parameters fixed to $\hat{\theta}^v$ and $\hat{\theta}^t$, respectively. Finally, we address the problem of generating diverse environments (*i.e.*, sampling weights w_{ij}). ECA module learns environments that the current embedding is not optimal by:

$$\arg\max_{w} \sum_{z_j \in Z} [\|\nabla_w|_{w=1.0} \mathcal{L}_{env}(z_j, \overline{\mathbf{v}}_i, \overline{\mathbf{t}}_i)\|^2]. \quad (11)$$

We update the environments every e epochs.

Partially Aligned Cross-modal Retrieval

CASR is extended to address partially aligned cross-modal retrieval, which involves training data with semantically unpaired image-text pairs (see Fig.2(b)). Diverging from methods that leverage category prototypes (*e.g.*, SPAL (Wang et al. 2024b) and OTPAL (Wang et al. 2024a)), CASR reinterprets unpaired data as a modality-missing problem. For example, when the instance $(\mathbf{x}_i^v, \mathbf{x}_i^t)$ exhibits a semantic mismatch, we regard $\mathbf{x}_i^v$ as missing its corresponding text modality, and $\mathbf{x}_i^t$ as missing its corresponding image modality. The missing modality is recovered using the MMI module, while the ECA module further refines cross-modal alignment. This strategy mitigates the reliance on category prototypes that capture only partial semantics, thereby enhancing the model's robustness to noisy data.

Learning Strategies

To achieve robust ICMR, we design a unified learning objective that combines two losses: 1) We use the classification loss $\mathcal{L}_{cls}$ to preserve category-level semantics:

$$\mathcal{L}_{cls} = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \cdot \log(\hat{y}_i^v) + y_i \cdot \log(\hat{y}_i^t)\right), \quad (12)$$

where $\hat{y}_i^v$ and $\hat{y}_i^t$ are predicted values by the two MLP classifier layers. Note that for unlabeled data in partially aligned cross-modal retrieval, we follow SPAL (Wang et al. 2024b) to generate pseudo-labels $\breve{y}_i$. 2) We also adopt the proposed $\mathcal{L}_{causal}$ to promote cross-modal alignment as described in Sec. 3. The total loss is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{causal} + \lambda * \mathcal{L}_{cls}, \quad (13)$$

where λ denotes the balance coefficient. The CASR process is summarized in Algorithm 1.

Algorithm 1: CASR

Input: Data $\mathcal{D}$, Visual Attention f^v, Text Attention f^t

for $i = 1$ to N epochs **do**
  Missing modality recovery by MMI
  if $i\%e == 0$ **then**
    for $j = 1$ to M **do**
      Update sample weights (environments)
    end for
  end if
  Update the model parameters by minimizing $\mathcal{L}_{total}$
end for

4 Experiments and Results

In this section, we conduct extensive experiments to answer the following research questions: RQ1: To what extent does the proposed CASR framework alleviate the issue of incomplete data? RQ2: For the different components in CASR, what are their roles and impacts on performance? RQ3: What are the learning patterns and insights of CASR?

Experimental Settings

Datasets and Data Splitting. We conduct experiments on four image-text retrieval datasets: Wikipedia (Rasiwasia et al. 2010), Pascal-Sentence (Rashtchian et al. 2010), NUS-WIDE-10K (Chua et al. 2009), and XmediaNet (Peng, Qi, and Yuan 2018), along with one video-text dataset, MSR-VTT (Xu et al. 2016). For incomplete cross-modal retrieval (ICMR) setting, we randomly remove a portion of vision or text modalities to simulate incomplete scenarios. The configuration (50%P, 0%I, 50%T) indicates that 50% of the training samples are complete, while the remaining 50% contain only text features. For partially aligned cross-modal retrieval (PACMR) setting, we randomly designate 20%, 40%, or 60% of the training samples as labeled paired data, with the rest treated as unlabeled unpaired samples.

Implementation Details. Our method is implemented in PyTorch and all experiments are conducted on an NVIDIA 4090 GPU. Following prior work (Shi et al. 2024), we use the pre-trained CLIP model (Radford et al. 2021) to extract 1024-dimensional features for both images and texts. For the MMI module, we adopt frozen Deepseek-LLM (Bi et al. 2024) as the category prior encoder. The model is optimized using the Adam optimizer with a fixed learning rate of 0.001 and a batch size of 256. We train for 100, 100, 200, and 100 epochs on the Wikipedia, Pascal Sentence, XmediaNet, and NUS-WIDE-10K datasets, respectively. Mean Average Precision (mAP) is used as the evaluation metric, following (Shi et al. 2024; Wang et al. 2024b). For video-text retrieval, we use CLIP4CLIP (Luo et al. 2022) as the feature extractor for both modalities. The model is trained for 10 epochs with a batch size of 64, and performance is evaluated by the sum of Recall@K (K = 1, 5, 10). The hyper-parameters α, β and λ are set to 1.0, 0.4 and 0.6, respectively.

Comparisons with SOTA Methods (RQ1)

We comprehensively evaluate CASR across different training paradigms, comparing it with: 1) Supervised: DAVAE (Jing et al. 2020), PAN (Zeng et al. 2021), C^3CMR (Wang et al. 2022), and DCT (Shi et al. 2024). 2) Semi-supervised: SCL_{ss} (Liu et al. 2022), SPAL (Wang et al. 2024b), and OTPAL (Wang et al. 2024a).

Imbalance Setting	Wikipedia				Pascal Sentence				NUS-WIDE			
	DAVAE	DCT	OTPAL	Ours	DAVAE	DCT	OTPAL	Ours	DAVAE	DCT	OTPAL	Ours
(100%P, 0%I, 0%T)	0.578	0.602	0.622	**0.658**	0.710	0.760	0.782	**0.806**	0.594	0.625	0.638	**0.693**
(50%P, 50%I, 0%T)	0.559	0.578	0.604	**0.649**	0.714	0.747	0.762	**0.780**	0.591	0.615	0.628	**0.673**
(50%P, 0%I, 50%T)	0.585	0.591	0.612	**0.651**	0.651	0.725	0.744	**0.767**	0.594	0.615	0.630	**0.665**
(50%P, 25%I, 25%T)	0.579	0.587	0.599	**0.637**	0.735	0.734	0.749	**0.778**	0.594	0.618	0.634	**0.668**
(30%P, 70%I, 0%T)	0.544	0.574	0.607	**0.633**	0.687	0.729	0.751	**0.775**	0.581	0.601	0.617	**0.669**
(30%P, 0%I, 70%T)	0.568	0.584	0.608	**0.643**	0.651	0.707	0.731	**0.766**	0.586	0.614	0.631	**0.675**
(30%P, 35%I, 35%T)	0.581	0.586	0.612	**0.649**	0.715	0.721	0.737	**0.765**	0.602	0.614	0.619	**0.659**
(10%P, 90%I, 0%T)	0.482	0.541	0.603	**0.647**	0.578	0.719	0.730	**0.744**	0.538	0.593	0.621	**0.676**
(10%P, 0%I, 90%T)	0.506	0.573	0.628	**0.648**	0.545	0.719	0.723	**0.749**	0.555	0.597	0.618	**0.684**
(10%P, 45%I, 45%T)	0.549	0.589	0.601	**0.653**	0.633	0.707	0.726	**0.755**	0.582	0.603	0.624	**0.679**

Table 1: Average mAP scores with different modality missing rates on three datasets.

Datasets	Methods	Venue	20% Paired data			40% Paired data			60% Paired data		
			I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
NUS-WIDE	PAN	SIGIR21	0.580	0.590	0.585	0.590	0.603	0.596	0.596	0.608	0.602
	C^3CMR	MM22	0.574	0.580	0.577	0.584	0.597	0.590	0.596	0.604	0.600
	DCT	TOMM23	0.551	0.511	0.531	0.559	0.553	0.556	0.572	0.571	0.571
	SCL_{ss}	TMM22	0.607	0.613	0.610	0.611	0.619	0.615	0.612	0.622	0.617
	OTPAL	MM24	0.631	0.639	0.635	0.636	0.638	0.637	0.635	0.645	0.640
	Ours	AAAI26	**0.673**	**0.665**	**0.669**	**0.675**	**0.677**	**0.676**	**0.675**	**0.684**	**0.680**
Wikipedia	PAN	SIGIR21	0.541	0.525	0.533	0.562	0.548	0.555	0.577	0.561	0.569
	C^3CMR	MM22	0.499	0.481	0.490	0.530	0.518	0.524	0.553	0.535	0.544
	DCT	TOMM23	0.452	0.399	0.425	0.548	0.520	0.534	0.560	0.539	0.549
	SCL_{ss}	TMM22	0.562	0.538	0.550	0.599	0.577	0.588	0.613	0.579	0.596
	OTPAL	MM24	0.570	0.556	0.563	0.609	0.589	0.599	0.614	0.600	0.607
	Ours	AAAI26	**0.621**	**0.603**	**0.612**	**0.641**	**0.635**	**0.638**	**0.655**	**0.644**	**0.650**
XmediaNet	PAN	SIGIR21	0.451	0.485	0.468	0.456	0.486	0.471	0.457	0.491	0.474
	C^3CMR	MM22	0.542	0.548	0.545	0.576	0.586	0.581	0.595	0.601	0.598
	DCT	TOMM23	0.573	0.544	0.558	0.610	0.567	0.588	0.586	0.595	0.590
	SCL_{ss}	TMM22	0.636	0.648	0.642	0.672	0.682	0.677	0.675	0.685	0.680
	OTPAL	MM24	0.726	0.706	0.716	0.731	0.744	0.737	0.738	0.752	0.745
	Ours	AAAI26	**0.760**	**0.748**	**0.754**	**0.768**	**0.763**	**0.766**	**0.784**	**0.783**	**0.784**

Table 2: Retrieval performance for mAP scores compared to existing methods on three datasets with partially aligned data.

Comparison on ICMR setting. To evaluate the robustness of CASR under incomplete data, we conducted experiments with varying proportions of missing modalities, as shown in Tab.1. CASR consistently achieved the best retrieval performance across all imbalance settings, with particularly outstanding results under severe modality-missing conditions. For example, in setting (10%P, 90%I, 0%T), CASR outperformed OTPAL by 7.3%, 1.9%, and 8.8% on the three datasets, respectively. Furthermore, we evaluated CASR in the incomplete video-text retrieval scenario (see Tab.3). Using CLIP4CLIP (denoted as "*C*") as the baseline, we combined it with DCT, OTPAL and CASR, respectively. The results demonstrate that CASR significantly enhances the robustness of CLIP4CLIP. These findings indicate that CASR possesses stronger capabilities in both modality recovery and cross-modal alignment.

Comparison on PACMR setting. We further evaluated the robustness of CASR in the PACMR setting, as shown in Table 2. Unlike supervised methods (e.g., C^3CMR and DCT) or unsupervised methods (*e.g.*, OTPAL), our approach treats the unlabeled unpaired samples as a form of incomplete data. Experimental results under various noise rates demonstrate that CASR consistently achieves state-of-the-art performance. Even with only 20% labeled data across the three datasets, CASR surpasses the second-best method, OTPAL, by 5.3%, 8.7%, and 5.3% in terms of mAP scores. These results clearly demonstrate that CASR is not only effective in handling missing modalities, but also applicable to complex real-world scenarios involving unlabeled unpaired samples.

Imbalance Setting	MSR-VTT			
	C	*C*-DCT	*C*-OTPAL	Ours
(100%P, 0%I, 0%T)	387.9	394.3	391.6	**407.5**
(50%P, 25%I, 25%T)	304.8	327.9	335.4	**349.2**
(30%P, 35%I, 35%T)	309.7	330.4	340.1	**347.2**
(10%P, 90%I, 0%T)	301.5	323.4	329.1	**342.5**
(10%P, 45%I, 45%T)	303.4	324.0	332.7	**343.6**

Table 3: Performance on MSR-VTT dataset under different modality missing rates."*C*" stands for CLIP4CLIP.

In-depth Studies of CASR (RQ2)

Contributions of the CASR's components. To fully understand CASR, we evaluate the effectiveness of the MMI and ECA modules on Pascal Sentence and NUS-WIDE under the (50%P, 25%I, 25%T) setting. The corresponding re-

Methods	Pascal Sentence			NUS-WIDE		
	I2T	T2I	Avg	I2T	T2I	Avg
Baseline	0.726	0.692	0.709	0.631	0.633	0.632
w/o MMI	0.759	0.762	0.761	0.645	0.661	0.653
w/o ECA	0.764	0.761	0.763	0.657	0.653	0.655
CASR	**0.789**	**0.767**	**0.778**	**0.674**	**0.664**	**0.668**

Table 4: Effects of CASR's components on Pascal Sentence and NUS-WIDE under the (50%P, 25%I, 25%T) setting.

sults are reported in Tab.4. 1) *Effectiveness of MMI.* Disabling MMI ("w/o MMI") results in a 2.2% drop in Average mAP on Pascal Sentence (0.761 vs. 0.778), highlighting MMI's essential role in recovering complete semantics. 2) *Effectiveness of ECA.* Excluding ECA ("w/o ECA") causes a 1.9% decrease in Average mAP on NUS-WIDE (0.655 vs. 0.668), demonstrating ECA's effectiveness in enhancing cross-modal causal alignment.

Parametric Sensitivity Analysis. We present the sensitivity analysis of the hyper-parameters α, β, and λ in Fig.4. Experiments are conducted on Wikipedia and NUS-WIDE datasets under the (10%P, 45%I, 45%T) setting. α and β balance the gradient penalty and the attention discrepancy penalty, respectively. λ represents the intensity of the classification loss. Our model exhibits minimal fluctuations across a wide range of hyper-parameters, indicating that performance remains stable and robust to parameter variations.

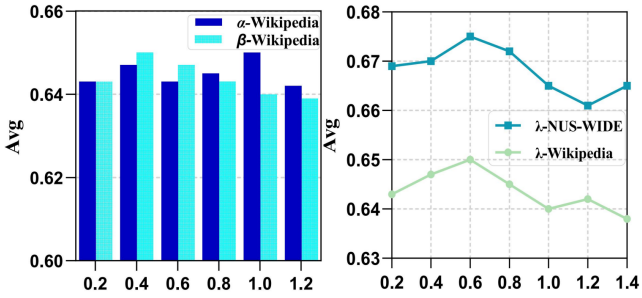


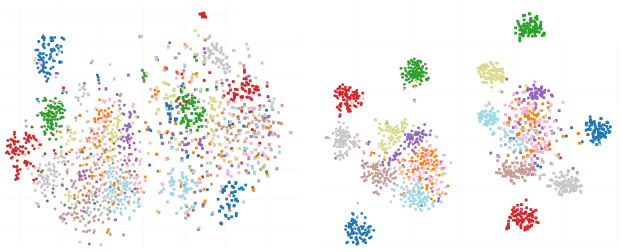
Figure 4: Ablation studies of Hyper-parameters.

Qualitative Analysis (RQ3)

By nature, CASR is equipped with intrinsic visual explainability. To capture the learning insights of CASR, we perform a series of detailed visual analyses.

Feature Visualization. In Fig.5, we visualize the visual and textual representations learned by the models on the NUS-WIDE-10K test set using t-SNE. Specifically, we compare the features learned by DCT (Fig.5(a)) with those learned by CASR (Fig.5(b)). It is evident that, compared to DCT, the image and text features learned by CASR exhibit more compact intra-class clustering and clearer inter-class separation. This semantically discriminative distribution verifies the effectiveness of CASR in enhancing both semantic discrimination and cross-modal alignment invariance.

Retrieval Visualization. Looking closely at Fig.6, compared to DCL and OTPAL, CASR is more effective in mitigating spurious cross-modal associations in retrieval,



(a) DCT (b) CASR

Figure 5: T-SNE visualization of NUS-WIDE test data.

thereby focusing more accurately on true causal alignments. As discussed in Sec.1, the frequent co-occurrence of "person" and "bottle" in the training data can mislead models like DCL and OTPAL. When retrieving using a textual query that only mentions "bottle", these models tend to be influenced by the spurious co-occurrence, causing the correct image to appear lower in the ranking. In contrast, CASR, with the aid of the ECA module, uncovers more authentic causal alignments, allowing it to rank semantically matching images at the top of the retrieval results.

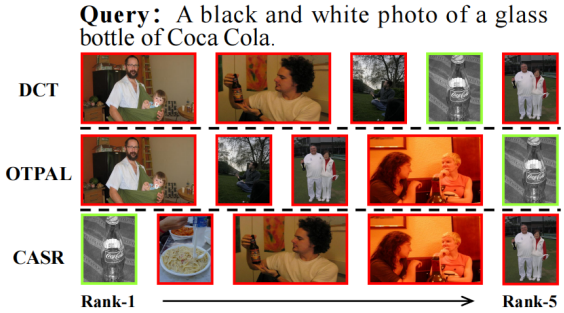


Figure 6: Text-to-image retrieval results on Pascal Sentence.

5 Conclusion

In this paper, we propose Causality-Aligned Semantic Recovery (CASR), a novel method to enhance the robustness of ICMR. We identify two critical limitations in existing approaches: insufficient semantic reconstruction and vulnerability to spurious cross-modal correlations. To address these issues, we design two core modules: the Missing Modality Imagination (MMI) module and the Explicit Causal Alignment (ECA) module. Specifically, the MMI module leverages category and contextual priors to comprehensively reconstruct the semantics of missing modalities, while the ECA module learns environment-invariant embedding to suppress spurious correlations and reinforce true causal alignments. Extensive experiments demonstrate the effectiveness of CASR in improving both retrieval accuracy and robustness. Looking ahead, we will investigate the applicability of CASR in real-world scenarios where incomplete modalities frequently occur, such as video question answering (Yang et al. 2024b) and text-to-pointcloud localization(Xu et al. 2025).

Acknowledgments

This research is supported by the National Natural Science Foundation of China (62276112, 62272435, and U22A2094) and Jilin Province Science and Technology Development Plan Key RD Project(20230201088GX).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 1–9.

Deng, X.; and Zhang, Z. 2022. Deep causal metric learning. In *International Conference on Machine Learning*, 4993–5006. PMLR.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jing, M.; Li, J.; Zhu, L.; Lu, K.; Yang, Y.; and Huang, Z. 2020. Incomplete cross-modal retrieval with dual-aligned variational autoencoders. In *Proceedings of the 28th ACM international conference on multimedia*, 3283–3291.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Liu, Y.; Chen, H.; Liang, Y.; Yang, Y.; Yang, X.; and Lyu, Y. 2025a. Enhancing Semantic Clarity: Discriminative and Fine-grained Information Mining for Remote Sensing Image-Text Retrieval. In Kwok, J., ed., *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 5815–5823. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Liu, Y.; Chen, H.; Qin, G.; Song, J.; and Yang, X. 2025b. Bias Mitigation and Representation Optimization for Noise-Robust Cross-Modal Retrieyang2024robustval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21(11).

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

Liu, Y.; Wu, J.; Qu, L.; Gan, T.; Yin, J.; and Nie, L. 2022. Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 25: 2851–2863.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Peng, Y.; Qi, J.; and Yuan, J., Yuxinnd Hockenmaier. 2018. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11): 5585–5599.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rashtchian, C.; Young, P.; Hodosh, M.; and Hockenmaier, J. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 139–147.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, 251–260.

Shi, D.; Zhu, L.; Li, J.; Dong, G.; and Zhang, H. 2024. Incomplete cross-modal retrieval with deep correlation transfer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5): 1–21.

Shvetsova, N.; Chen, B.; Rouditchenko, A.; Thomas, S.; Kingsbury, B.; Feris, R. S.; Harwath, D.; Glass, J.; and Kuehne, H. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 20020–20029.

Song, J.; Chen, H.; Lyu, Y.; Nie, W.; and Liu, A.-A. 2025a. Causality-inspired Unsupervised Domain Adaptation with Target Style Imitation for Medical Image Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Song, J.; Chen, H.; Qin, J.; and Zhao, N. 2025b. Dual-supervised Asymmetric Co-training for Semi-supervised Medical Domain Generalization. *IEEE Transactions on Multimedia*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, J.; Gong, T.; Yan, Y.; Qin, G.; and Chen, H. 2024a. Partially Aligned Cross-modal Retrieval via Optimal Transport-based Prototype Alignment Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 701–709.

Wang, J.; Gong, T.; Yan, Y.; Qin, G.; and Chen, H. 2024b. Semi-supervised Prototype Semantic Association Learning for Robust Cross-modal Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 872–881.

Wang, J.; Gong, T.; Zeng, Z.; Sun, C.; and Yan, Y. 2022. C3cmr: Cross-modality cross-instance contrastive learning for cross-media retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4300–4308.

Wang, T.; Yue, Z.; Huang, J.; Sun, Q.; and Zhang, H. 2021. Self-supervised learning disentangled group representation

as feature. *Advances in Neural Information Processing Systems*, 34: 18225–18240.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, Y.; Qu, H.; Liu, J.; Zhang, W.; and Yang, X. 2025. CMMLoc: Advancing Text-to-PointCloud Localization with Cauchy-Mixture-Model Based Framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 6637–6647.

Yang, X.; Chang, T.; Zhang, T.; Wang, S.; Hong, R.; and Wang, M. 2024a. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *International Journal of Computer Vision*, 132(11): 4823–4849.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 1–10.

Yang, X.; Zeng, J.; Guo, D.; Wang, S.; Dong, J.; and Wang, M. 2024b. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67(10): 202104.

Zeng, Z.; Wang, S.; Xu, N.; and Mao, W. 2021. PAN: Prototype-based adaptive network for robust cross-modal retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 1125–1134.