A Related work

A.1 Multimodal alignment

Multimodal learning addresses four key challenges (Liang et al., 2024c; Baltrušaitis et al., 2018; Liang et al., 2024d): managing interactions among redundant, unique, and synergistic features (Dumas et al., 2017; Liang et al., 2024a;b), aligning fine-grained and coarse-grained information (Wang et al., 2023; 2024a), reasoning across diverse features (Yang et al., 2023), and integrating external knowledge (Shen et al., 2022; Lyu et al., 2024). Among these challenges, multimodal alignment is one of the core challenges that many researchers aim to solve.

A common method in multimodal alignment is using cross-modal alignment by using attention mechanisms between pairwise modalities, such as vision-language (Tan & Bansal, 2019) and vision-language-audio (Tsai et al., 2019). Another effective approach is leveraging graph neural networks to align multimodal datasets (Yang et al., 2021; Wilf et al., 2023). For instance, Yang et al. (2021) transforms unaligned multimodal sequence data into nodes, with edges capturing interactions across modalities over time. Wilf et al. (2023) build graph structures for each modality—visual, textual, and acoustic—and create edges to represent their interactions.

To enhance the generalizability of cross-modal representations, Xia et al. (2024) employ a unified codebook approach, facilitating a joint embedding space for visual and audio modalities. Another prominent method (Radford et al., 2021) achieves cross-modal alignment by leveraging large collections of image-text pairs, making it a widely adopted strategy in multimodal learning (Zhang et al., 2022; Guzhov et al., 2022; Zhou et al., 2023).

A.2 Binding methods

Recent studies have focused on aligning multimodal datasets by leveraging binding properties in various modalities. ImageBind (Girdhar et al., 2023) aligns multimodal data by using image representation as the anchor and aligning each modality's embedding with the image embedding. Similarly, LanguageBind (Zhu et al., 2024) uses language representation as the anchor, aligning other modalities into the language space. PointBind (Guo et al., 2023) learns a joint embedding space across 3d point, language, image, and audio modalities by designating the point space as the central representation. Thanks to the efficacy of such a binding idea with a fixed anchor, several "-Bind" approaches have been studied in numerous domains (Teng et al., 2024; Xiao et al., 2024; Gao et al., 2024; Yang et al., 2024b; Balemans et al., 2024; Dhakal et al., 2024; Yang et al., 2024a) While these methods demonstrate strong performance in zero-shot cross-modality retrieval and classification tasks, they are constrained by their reliance on an existing single anchor modality.

Several approaches have integrated additional knowledge into multimodal representation spaces to address this limitation. Freebind (Wang et al., 2024a) introduces bi-modality spaces to enhance a pretrained image-paired unified space. It generates pseudo-embedding pairs across diverse modality pairs and aligns them with the pre-trained unified space using contrastive learning. Omnibind (Wang et al., 2024b) leverages multiple pretrained multimodal models to construct pseudo item-pair retrievals based on top-1 recall across various modality combinations using pairwise cross-modal alignment. Both methods show promising results in cross-modal retrieval by incorporating extra spaces into existing pairwise binding spaces. However, they still rely on fixed (pre-trained) representation spaces.

Unibind (Lyu et al., 2024) highlights the imbalanced representation when using image-centered representation spaces. To address this, Unibind employs large language models (LLMs) to create a unified and balanced representation space. It constructs a knowledge base with multimodal category descriptions, establishes LLM-augmented class-wise embedding centers, and aligns other modalities to these centers through contrastive learning. This approach attempts to balance representations across modalities but still depends heavily on large-scale pretrained LLMs and centers alignment around a single unified space.

B PROOFS

B.1 PROOF OF PROPOSITION 1

Using the chain rule of the mutual information, we observe that

$$\begin{aligned} I(\mathbf{X}_1, f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i) &= I(\mathbf{X}_1; \mathbf{X}_i) + I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i | \mathbf{X}_1) \\ &= I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i) + I(\mathbf{X}_1; \mathbf{X}_i | f_1^{\text{suf}}(\mathbf{X}_1)), \end{aligned} \tag{12}$$

Since $f_1^{\text{suf}}(\mathbf{X}_1)$ is a deterministic function of $\mathbf{X}_1$, we have

$$I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i | \mathbf{X}_1) = 0. \tag{13}$$

Moreover, f_1^{suf} obtained in Definition 1 with proper choice of $\mathcal{Z}$ achieves the maximum mutual information, implying together with $I(\mathbf{X}; \mathbf{Y}) \leq \min\{H(\mathbf{X}), H(\mathbf{Y})\}$ that $I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_1) = H(\mathbf{X}_1)$, where $H(\mathbf{X}_1)$ is the entropy of $\mathbf{X}_1$ (Polyanskiy & Wu, 2024). In other words, we have $H(\mathbf{X}_1 | f_1^{\text{suf}}(\mathbf{X}_1)) = H(\mathbf{X}_1) - I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_1) = 0$. This gives

$$\begin{aligned} I(\mathbf{X}_1; \mathbf{X}_i | f_1^{\text{suf}}(\mathbf{X}_1)) &= H(\mathbf{X}_1 | f_1^{\text{suf}}(\mathbf{X}_1)) - H(\mathbf{X}_1 | f_1^{\text{suf}}(\mathbf{X}_1), \mathbf{X}_i) \\ &= 0 \end{aligned} \tag{14}$$

Substituting (13) and (14) into (12) yields

$$I(\mathbf{X}_1; \mathbf{X}_i) = I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i). \tag{15}$$

We conclude the proof of Proposition 1 by noting that the optimality of FABIND (i.e., $I(f_1^{\text{suf}}(\mathbf{X}_1); \mathbf{X}_i) = I(f_1^{\text{suf}}(\mathbf{X}_1); f_i^{\text{FB}}(\mathbf{X}_i))$, $\forall i \in \{2, \cdots, M\}$) yields

$$I(\mathbf{X}_1; \mathbf{X}_i) = I(f_1^{\text{suf}}(\mathbf{X}_1); f_i^{\text{FB}}(\mathbf{X}_i)). \tag{16}$$

B.2 PROOF OF PROPOSITION 2

Using the chain rule of mutual information, we have

$$\begin{aligned} I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1, \mathbf{X}_i) &= I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1) + I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i | \mathbf{X}_1) \\ &= I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i) + I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1 | \mathbf{X}_i). \end{aligned} \tag{17}$$

Moreover, since $f_1^{\text{ins}}(\mathbf{X}_1)$ is a deterministic function of $\mathbf{X}_1$, we have $I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i | \mathbf{X}_1) = 0$, leading to $I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1) = I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i) + I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1 | \mathbf{X}_i)$. Then, using the assumption $I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1) < \epsilon$, it follows that

$$\begin{aligned} \epsilon &> I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i) + I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_1 | \mathbf{X}_i) \\ &\overset{\text{(a)}}{\geq} I(f_1^{\text{ins}}(\mathbf{X}_1); \mathbf{X}_i) \\ &\overset{\text{(b)}}{\geq} I(f_1^{\text{ins}}(\mathbf{X}_1); f_i^{\text{FB}}(\mathbf{X}_i)), \end{aligned} \tag{18}$$

where the labeled inequalities follow from: (a) the non-negativity of mutual information; (b) the data processing inequality. This concludes the proof of Proposition 2.

B.3 PROOF OF THEOREM 1

To prove Theorem 1, we leverage the reverse inequality of M-variable Hölder inequality (Seo, 2013, eq. (2.8)). For the sake of completeness, we state the inequality in Lemma 1.

Lemma 1 (Reverse inequality of the M-variable Hölder inequality (Seo, 2013)). *Consider M sequences $(x_{i,j})_{j\in[n]}$, $i \in [M]$ of n positive scalars such that for some $0 < c_m \leq c_M < \infty$,*

$$0 < c_m \leq x_{i,j} \leq c_M < \infty, \ \forall i, j. \tag{19}$$

Then,

$$\prod_{i=1}^{M} \left(\sum_{j=1}^{n} x_{i,j} \right)^{\frac{1}{n}} \leq \frac{(c_m + c_M)^2}{4 c_m c_M} \sum_{j=1}^{n} \left(\prod_{i=1}^{M} x_{i,j} \right)^{\frac{1}{n}}. \tag{20}$$

Now we start by writing the summation of InfoNCE losses for each $f_l^{(t)}(\boldsymbol{x}'_{l,k}), l \in [M]$ to $f_i(\mathbf{X}_i)$ as

$$\sum_{l=1}^{M} I_{\mathrm{NCE}}(f_l(\mathbf{X}'_l); f_i(\mathbf{X}_i)|\tau) = -\frac{1}{|\mathcal{I}_B|}\sum_{k=1}^{|\mathcal{I}_B|}\sum_{l=1}^{M}\log\frac{\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,k})}{\tau}\right)}{\sum_{j\in\mathcal{I}_B}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau}\right)}. \tag{21}$$

Then, the inner summation in (21) is bounded as

$$\begin{aligned}
&\sum_{l=1}^{M}\log\frac{\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,k})}{\tau}\right)}{\sum_{j\in\mathcal{I}_B}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau}\right)} \\
&= \frac{1}{\tau}\sum_{l=1}^{M} f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,k}) - \log\prod_{l=1}^{M}\sum_{j\in\mathcal{I}_B}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau}\right) \\
&\overset{(a)}{\geq} \frac{1}{\tau}\sum_{l=1}^{M} f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,k}) - \log\left(C_{\mathcal{F},k,i}\sum_{j\in\mathcal{I}_B}\prod_{l=1}^{M}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau|\mathcal{I}_B|}\right)\right)^{|\mathcal{I}_B|} \\
&\overset{(b)}{=} \frac{M}{\tau}\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,k}) - |\mathcal{I}_B|\log\sum_{j\in\mathcal{I}_B}\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,j})}{\tau|\mathcal{I}_B|}\right) - |\mathcal{I}_B|\log C_{\mathcal{F},k,i} \\
&= |\mathcal{I}_B|\log\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,k})}{\tau|\mathcal{I}_B|}\right) - |\mathcal{I}_B|\log\sum_{j\in\mathcal{I}_B}\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,j})}{\tau|\mathcal{I}_B|}\right) - |\mathcal{I}_B|\log C_{\mathcal{F},k,i} \\
&= |\mathcal{I}_B|\log\frac{\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,k})}{\tau|\mathcal{I}_B|}\right)}{\sum_{j\in\mathcal{I}_B}\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,j})}{\tau|\mathcal{I}_B|}\right)} - |\mathcal{I}_B|\log C_{\mathcal{F},k,i},
\end{aligned} \tag{22}$$

where the labeled (in)equalities follow from: (a) Lemma 1 and $C_{\mathcal{F},k,i} = \frac{(c^{\min}_{\mathcal{F},k,i}+c^{\max}_{\mathcal{F},k,i})^2}{4c^{\min}_{\mathcal{F},k,i}c^{\max}_{\mathcal{F},k,i}}$ with

$$\begin{aligned}
c^{\min}_{\mathcal{F},k,i} &= \min_{\ell\in[M],j\in\mathcal{I}_B}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau}\right), \text{ and} \\
c^{\max}_{\mathcal{F},k,i} &= \max_{\ell\in[M],j\in\mathcal{I}_B}\exp\left(\frac{f_l^\top(\boldsymbol{x}'_{l,k})f_i(\boldsymbol{x}_{i,j})}{\tau}\right);
\end{aligned} \tag{23}$$

and (b) the definition of anchor embedding (7). Substituting (22) into (21) gives

$$\begin{aligned}
\sum_{l=1}^{M} I_{\mathrm{NCE}}(f_l(\mathbf{X}'_l); f_i(\mathbf{X}_i)|\tau) &\leq -\frac{1}{|\mathcal{I}_B|}\sum_{k=1}^{|\mathcal{I}_B|}\left[|\mathcal{I}_B|\log\frac{\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,k})}{\tau|\mathcal{I}_B|}\right)}{\sum_{j\in\mathcal{I}_B}\exp\left(\frac{M\boldsymbol{a}_k^\top f_i(\boldsymbol{x}_{i,j})}{\tau|\mathcal{I}_B|}\right)} - |\mathcal{I}_B|\log C_{\mathcal{F},k,i}\right] \\
&= |\mathcal{I}_B| I_{\mathrm{NCE}}\left(\mathbf{A}; f_i(\mathbf{X}_i)\,\middle|\,\frac{\tau|\mathcal{I}_B|}{M}\right) + \sum_{k=1}^{|\mathcal{I}_B|}\log C_{\mathcal{F},k,i}.
\end{aligned} \tag{24}$$

Rearranging (24) and setting $\tilde{\tau} = \frac{\tau|\mathcal{I}_B|}{M}$ in (23) and (24) yield

$$I_{\mathrm{NCE}}\left(\mathbf{A}; f_i(\mathbf{X}_i) \mid \tilde{\tau}\right) \geq \frac{1}{|\mathcal{I}_B|}\sum_{l=1}^{M} I_{\mathrm{NCE}}\left(f_l(\mathbf{X}'_l); f_i(\mathbf{X}_i)\,\middle|\,\frac{\tilde{\tau}M}{|\mathcal{I}_B|}\right) - \frac{1}{|\mathcal{I}_B|}\sum_{k=1}^{|\mathcal{I}_B|}\log C_{\mathcal{F},k,i}, \tag{25}$$

which concludes the proof of Theorem 1.

C Experiment Details

C.1 Experiments with synthetic datasets

Synthetic datasets. We employ a latent variable model (Bishop & Nasrabadi, 2006) for generating synthetic multimodal datasets. A latent variable model is a statistical model for data $\mathbf{X} \in \mathbb{R}^{d_x}$,

under which $\mathbf{X}$ is generated according to a conditional probability distribution $P_{\mathbf{X}|\mathbf{Z}}$, where $\mathbf{Z} \in \mathbb{R}^{d_z}$ is the latent variable. In terms of the representation learning framework, $\mathbf{Z}$ can be seen as a true representation of $\mathbf{X}$. Moreover, we assume that the class label $\mathbf{Y} \in [K]$ and the latent variable $\mathbf{Z}$ are jointly distributed according to $P_{\mathbf{Z},\mathbf{Y}}$.

For the marginal distribution of $\mathbf{Z}$, we make use of a Gaussian mixture model (GMM) (Bishop & Nasrabadi, 2006), and hence the probability density function (PDF) of $\mathbf{Z}$ is a weighted sum of Gaussian densities. In particular, the PDF of $\mathbf{Z}$ is defined as follows:

$$p_{\mathbf{Z}}(\boldsymbol{z}) = \prod_{y=1}^{K} \pi_y \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \tag{26}$$

where K is the number of mixture components, $\pi_y = \Pr(\mathbf{Y} = y)$ is the component prior probability, and $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ denotes Gaussian PDF with mean $\boldsymbol{\mu}_y \in \mathbb{R}^{d_z}$ and covariance matrix $\boldsymbol{\Sigma}_y \in \mathbb{R}^{d_z \times d_z}$. This leads to the conditional PDF of $\mathbf{Z}$ as $p_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{z}|y) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$.

Once a latent variable $\boldsymbol{z}$ is generated from GMM in (26), we generate data samples $(\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,N})$ for i-th modality using the conditional PDFs of $\mathbf{X}_i$ given $\boldsymbol{z}$, denoted by $p_{\mathbf{X}_i|\mathbf{Z}}(\boldsymbol{x}_i|\boldsymbol{z})$. Specifically, we use the model $\mathbf{X}_i = g_i(\mathbf{Z}_i) + \mathbf{N}$, where $g_i : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ is a non-linear projection from latent space to observation space, and $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, I_{d_x})$ is Gaussian noise with zero-mean and identity covariance matrix. To make the inherent correlation between $\mathbf{X}_i$ and $\mathbf{Z}_i$ different among modalities, we choose g_i such that

$$g_i(\mathbf{Z}) = \Theta_i^{(2)} \text{sigmoid}\left(\Theta_i^{(1)} \mathbf{Z}\right), \tag{27}$$

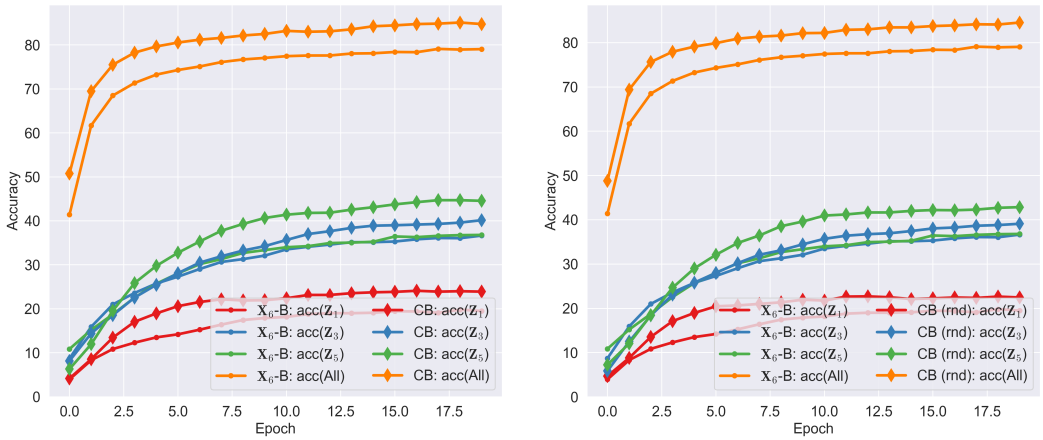
where $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ is applied element-wise, and $\Theta_i^{(1)} \in \mathbb{R}^{d_x \times d_z}$ and $\Theta_i^{(2)} \in \mathbb{R}^{d_x \times d_x}$ are matrices randomly generated from Gaussian distribution. Moreover, after $\Theta_i^{(1)}, i \in [M]$ are generated, we set arbitrary columns of them all zero, so that the number of all zero columns decreases in i. For example, 60% of columns of $\Theta_1^{(1)}$ are all-zero, while only 10% of columns of $\Theta_M^{(1)}$ are all-zero. This enables approximate control the correlation between $\mathbf{X}_i$ and $\mathbf{Z}$, providing estimates of best modality ($\mathbf{X}_M$) or worst modality ($\mathbf{X}_1$). To have meaningful labels for this latent model, which requires for downstream tasks, we set the labels $\mathbf{Y}$ being the component index in GMM. In particular, since there are K components in GMM (26), there exists K categories in $\mathbf{Y}$. We conduct experiments with three different synthetic datasets by setting $M = 4, 6, 8$. For all synthetic datasets, we fix $d_x = 16$, $d_z = 8$, and $K = 50$.

Experiment details. We initialize two different versions of backbones for all modalities, where the first is a random backbone (highlighted by (rnd) in figures), and the second is a backbone pretrained with InfoNCE loss. For each backbone, we use a simple multilayer perceptron (MLP). Comparing the results with these two versions of backbone provides how much both FABIND and CENTROBIND are robust to backbone quality. Given the backbones for M modalities, we align the corresponding embedding spaces using either FABIND with anchor $\mathbf{X}_i$ (denoted by $\mathbf{X}_i$-B in figures) or CENTROBIND (denoted by CB in figures). Finally, with the encoders aligned by either FABIND or CENTROBIND, we evaluate classification accuracy as a measure of representation quality. We use a simple MLP for the classifier. To distinguish between accuracy with embeddings from a single modality and the one with concatenated embeddings from all modalities, we denote by acc($\mathbf{Z}_i$) the accuracy with embeddings from i-th modality and by acc(All) the accuracy with embeddings from all modalities.

Additional experimental results on synthetic datasets with $M = 6, 8$ number of modalities are shown in Figure 3 and Figure 4.

C.2 EXPERIMENTS WITH REAL-WORLD DATASETS

Training details. We utilize Low-Rank Adaptation (Hu et al., 2022) for training CENTROBIND and FABIND , enhancing training efficiency and achieving impressive results with fewer iterations. For parameter settings, we set a learning rate of 0.001, the AdamW optimizer (Loshchilov & Hutter, 2019) with a batch size of 16, and a temperature of 0.3 for InfoNCE. Training CENTROBIND requires augmentation. We augment video frames with various transformations, including random perspective



(a) FABIND at $\mathbf{X}_6$ vs. CENTROBIND

(b) When CENTROBIND uses random backbones.

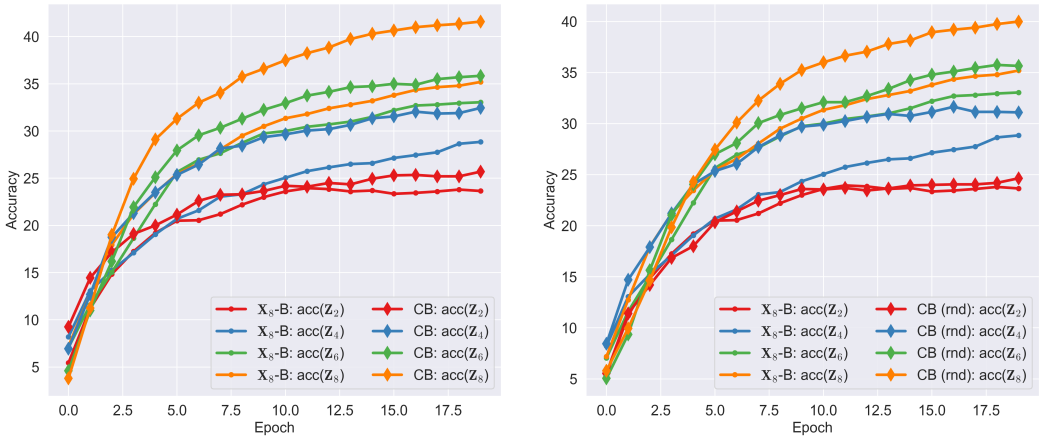
(c) When FABIND uses random backbones.

(d) When all backbones are random backbones.

Figure 3: Experiment results with synthetic dataset of $M = 6$ modalities. Abbreviation: $\mathbf{X}_i$-B or CB: applying FABIND method to backbones with anchor $\mathbf{X}_i$ or applying CENTROBIND; acc($\mathbf{Z}_i$) or acc(All): accuracy of $\mathbf{Z}_i$ or of concatenated embeddings $(\mathbf{Z}_1, \cdots, \mathbf{Z}_M)$; (rnd): if random backbones are used for $\mathbf{X}_i$-B or CB.

shifts, random flips and rotation, color jitter, Gaussian blur, and auto-contrast adjustment. For the audio modality, we apply a low-pass filter, speed changes, echo effect, room impulse response convolution, and background noise. For the text modality, we generate paraphrased sentences using the Phi-3 language model served using Ollama [4].

[4] https://ollama.com/library/phi3



(a) FABIND at $\mathbf{X}_6$ vs. CENTROBIND

(b) When CENTROBIND uses random backbones.

(c) When FABIND uses random backbones.

(d) When all backbones are random backbones.

Figure 4: Experiment results with synthetic dataset of $M = 8$ modalities. Abbreviation: $\mathbf{X}_i$-B or CB: applying FABIND method to backbones with anchor $\mathbf{X}_i$ or applying CENTROBIND; acc($\mathbf{Z}_i$) or acc(All): accuracy of $\mathbf{Z}_i$ or of concatenated embeddings $(\mathbf{Z}_1, \cdots, \mathbf{Z}_M)$; (rnd): if random backbones are used for $\mathbf{X}_i$-B or CB.