M3-JEPA: Multimodal Alignment via Multi-gate MoE based on the Joint-Embedding Predictive Architecture

Hongyang Lei [*1] **Xiaolong Cheng** [*1] **Qi Qin** [*2] **Dan Wang** [1] **Huazhen Huang** [3] **Qingqing Gu** [1] **Yetao Wu** [1] **Luo Ji** [1]

Abstract

Current multimodal learning strategies primarily optimize in the original token space. Such a framework is easy to incorporate with the backbone of pretrained language model, but might result in modality collapse. To alleviate such issues, we leverage the Joint-Embedding Predictive Architecture (JEPA) on the multimodal tasks, which converts the input embedding into the output embedding space by a predictor and then conducts the cross-modal alignment on the latent space. We implement this predictor by a Multi-Gate Mixture of Experts (MMoE) and name the framework as M3-JEPA, accordingly. The gating function disentangles the modality-specific and shared information and derives information-theoretic optimality. The framework is implemented with both contrastive and regularization loss, and solved by alternative gradient descent (AGD) between different multimodal tasks. By thoroughly designed experiments, we show that M3-JEPA can obtain state-of-the-art performance on different modalities and tasks, generalize to unseen datasets and domains, and is computationally efficient in both training and inference. Our observation suggests that M3-JEPA might become a new basis to self-supervised learning in the open world.

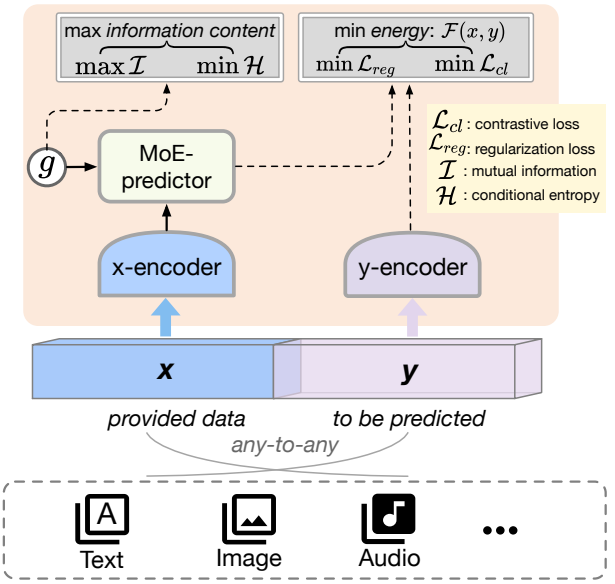

Figure 1. The paradigm of M3-JEPA on any-to-any multi-modality tasks. The self-supervised learning is conducted with two encoding branches of input and output signals, as well as an MoE predictor which projects the input embedding into the output latent space. M3-JEPA is an energy-based model that minimizes both contrastive and regularization losses. M3-JEPA is also conditioned on the inherent information content (g) which maximizes the mutual information and minimizes the conditional entropy.

1. Introduction

Human perception is inherently multi-modal, seamlessly integrating diverse sensory inputs from vision, hearing, touch, and other senses to comprehend the world. Inspired by this capability, modern modeling techniques also process and integrate information from multiple modalities like text, image, audio and video, becoming an important way to solve complex tasks involving heterogeneous data sources (Alayrac et al., 2022; Radford et al., 2021; Wang et al., 2022a; Liu et al.). Modern multi-modal studies are dominated by generative architecture, which can generally be classified into two main categories. The first category trains the model from scratch (Wang et al., 2022a; 2023a; Oquab et al., 2025), which asks for large-scale data therefore faces high training costs. Another category often employs a pretrained Large Language Model (LLM) as the backbone while finetuning a lightweight cross-modal connector (Alayrac et al., 2022; Li et al., 2023; Wu et al., 2024a; Liu et al., 2023; Jun Zhan, 2024). In this manner, the prior

[*]Equal contribution [1]Geely AI Lab, Zhejiang, China [2]Peking University, Beijing, China [3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. Correspondence to: Luo Ji <Luo.Ji1@geely.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

knowledge of LLM can be well preserved while the computational burden can also be alleviated (Zhang et al., 2024).

Although these studies achieve remarkable success in multimodal learning, however, they may be often subject to the issue of modality collapse during the cross-modal alignment, resulting from the conflicting gradients (Javaloy et al., 2022), missing modality or labels (Wu et al., 2024b), and mismatched data distribution or fusion (Ma et al., 2022). Especially, LLMs take advantage of self-supervised learning (SSL), which predicts the hidden parts of the data grounded by the visible parts, in the discrete token space. Nevertheless, this paradigm relatively struggles in continuous domains (*e.g.*, image or video), which causes the cross-modal alignment difficult to converge, and might fail to capture key information (Dawid & LeCun, 2024), especially in the existence of information uncertainty, redundancy, or ambiguity (*e.g.*, one picture can be described by two content different but semantically similar sentence).

To mitigate these issues, one possible solution is to switch from generative and probabilistic modeling to the energy-based model (EBM) paradigm, which minimizes a non-negative energy function of both the input and output embeddings. To better handle the uncertainty of information, the Joint-Embedding Predictive Architecture (JEPA) (Dawid & LeCun, 2024) is proposed, which employs a latent predictor that projects the necessary information from the input into the output embedding space, while filtering out irrelevant or misleading input features. JEPA then aligns the input and output signals in the latent level instead of the token or pixel level, with a similar idea proposed by the "platonic representation" (Huh et al., 2024). JEPA has been applied in vision, namely I-JEPA (Assran et al., 2023), in which an arbitrary part of an image is masked and predicted by the other image blocks; as well as the motion and content features of videos (Bardes et al., 2023). Nevertheless, there has been no generalized framework that applies the idea of JEPA to generalized multi-modality modeling, with any-to-any modality combinations either for the input or the output.

In this work, we propose a novel **M**ultimodal alignment paradigm via **M**ulti-gate **M**oE based on **JEPA**, in short, **M3-JEPA**. It studies the dependency of the unobserved part (y) on the observed part (x) in the embedding space of y, where x and y may belong to different modalities or any combination of them. Their embeddings are encoded by pretrained uni-modality encoders, with possibly several layers being finetuned during the multimodal learning. We implement the latent predictor by the Mixture-of-Experts (MoE) structure, as a lightweight cross-modal connector. To tackle the potential multimodal semantic discrepancy, we decouple the cross-modal information into modality-specific and shared components through the gating function of MoE. This MoE predictor is initialized randomly and trained by two terms of losses, the contrastive loss and the regularization loss. These two loss terms together form an energy function, and are adapted with the two-gate output of the MoE predictor. To avoid the representation collapse, the predictor is also driven by the optimal information content of an implicit latent variable, as discussed in (Dawid & LeCun, 2024). The paradigm of **M3-JEPA** is visualized by Figure 1, and the project code can be found at `https://github.com/HongyangLL/M3-JEPA`.

To better alleviate the gradient conflict issue across various modalities, M3-JEPA switches multiple directional multimodal tasks stepwisely, and solves them by Alternating Gradient Descent (AGD) (Akbari et al., 2023). To validate the reasonability of M3-JEPA, we conduct both theoretical study and empirical experiments. We provide an informative theoretic analysis, and discuss the convergence and optimal hyper-parameters. We also conduct comprehensive experiments across multiple multimodal tasks, including vision-language, audio-language, and vision classification tasks, while keeping the model architecture and training strategy the same. Experimental results demonstrate that our M3-JEPA achieves competitive results across different modalities and tasks, as well as remarkable generalization on unseen domains, compared to current state-of-the-art (SoTA) multimodal models. Furthermore, our approach is also computationally efficient from both training and inference aspects. We summarize our contributions as follows:

- We propose a novel any-to-any multi-modal alignment paradigm based on JEPA, to mitigate the potential modality collapse by aligning on the latent space.
- We leverage a computationally efficient multi-gate MoE architecture as the cross-modal predictor of JEPA, while freeze most parameters of modality encoders.
- We disentangle the gating function of MoE into the modality-specific and shared information, and also derive an information-theoretical analysis.
- We optimize M3-JEPA by alternating the gradient descent between different multi-directional multimodal tasks, and discuss its theoretical convergence.
- The experimental results demonstrate remarkable multimodal alignment accuracy and efficiency, encompassing text, image, audio and other modalities.

The rest of the paper is organized as follows. We first introduce the preliminary knowledge in Section 2. The methodology is stated in Section 3. The theoretical derivation is stated in Section 4. Experiment results and subsequent discussions are summarized in Section 5. The connection with previous works is stated in Section 6. Finally Section 7 concludes this paper.

2. Preliminary

2.1. Joint-Embedding Predictive Architecture

Given an input-output pair (x, y) which are encoded into (e_x, e_y), a joint-embedding predictive architecture (JEPA) (Dawid & LeCun, 2024) first converts the input encoding into the output's dimensional space by a predictor $\mathcal{P}(\cdot) : \mathbb{R}^x \to \mathbb{R}^y$, then minimizes its loss with e_y:

$$\min \mathcal{L}(e_{x\to y}, e_y) := \mathcal{F}(x, y), \quad e_{x\to y} = \mathcal{P}(e_x) \tag{1}$$

with $\mathcal{L}(\cdot, \cdot)$ denotes the loss; $\mathcal{F}$ is nonnegative and can be viewed as the energy between x and y. Subsequently, the JEPA framework belongs to energy-based model (EBM).

2.2. Mixture-of-Experts

The Mixture-of-Experts (MoE) architecture (Garmash & Monz, 2016; Shazeer et al., 2017) uses N feed-forward networks (FFN), namely "experts". The output of expert $(\mathbb{E}_{n=1}^N)$ is given by:

$$\mathbb{E}_n(e) = w_n^{out} \cdot \sigma(w_n^{in} \cdot e) \tag{2}$$

in which $\mathbb{E}_{n=1}^N$ denotes N experts, e is the input vector, w_n^{in} and w_n^{output} are learnable weights of the n-th expert, and σ is the activation function (*e.g.*, GeLU). An additional gating network $\mathbb{G}$ outputs an N-dimension normalized vector:

$$\mathbb{G}(e) = \text{softmax}(g \cdot e) \tag{3}$$

where g is a learnable matrix. The output of $\mathbb{G}$ routes each input via a few of the experts, and the output of MoE is:

$$\text{MoE}(e) = \sum_{n=1}^{N} \mathbb{G}(e)_n \mathbb{E}_n(e) \tag{4}$$

where $\mathbb{G}(e)_n$ denotes the probability of selecting expert $\mathbb{E}_n$.

Top-K MoE. The above MoE can be implemented with a top-K mechanism, which first ranks the gating scores, then keeps and summarizes top K experts instead of N experts

$$\text{MoE-K}(e) = \sum_{k=1}^{K} \text{top-K}(\mathbb{G}(e)_k)\mathbb{E}_{n=k}(e) \tag{5}$$

Multi-gate Mixture-of-Experts. The Multi-gate Mixture-of-Experts (MMoE) (Ma et al., 2018) expand the MoE architecture to the multi-task setting. Given L tasks, MMoE generates the outputs simultaneously by L parallel gates:

$$\text{MMoE}^l(z) = \sum_{n=1}^{N} (\mathbb{G}^l(z)_n)\mathbb{E}_n(z), \quad l = 1, \cdots, L \tag{6}$$

then use them to calculate the loss of each task:

$$\min \mathcal{L} := \sum_{l=1}^{L} \mathcal{L}^l(\text{MMoE}^l(z)) \tag{7}$$

where $\mathcal{L}$ and $\mathcal{L}^l(\cdot)$ are the total loss and task losses.

3. Methodology

This section illustrate our methodology, and Figure 2 provides a detailed introduction to the framework of M3-JEPA.

3.1. Problem Formulation

We develop M3-JEPA as an any-to-any modality framework. Assume we have total M modalities and T tasks. For each modality m, a uni-modal encoder can be employed to produce its latent embedding e_m. For the t-th task, we denote its input and output with (x^t, y^t), which contain the set of $\{m_x^t\}$ and $\{m_y^t\}$ modalities, respectively, with $1 \le m_x^t, m_y^t \le M$. Then the embedding of (x^t, y^t) can be the corresponding or the combination of modality latents:

$$e_x^t = \text{concat}(\{e_m\}), \quad m \in \{m_x^t\}$$
$$e_y^t = \text{concat}(\{e_m\}), \quad m \in \{m_y^t\}$$

Similar to Equation 1, we formulate an energy term $\mathcal{F}(x, y)$ which behaves both the training loss and also the alignment score during the inference

$$\mathcal{F}^t(x, y) = \mathcal{L}^t(e_{x\to y}^t, e_y^t) = \mathcal{L}^t(\mathcal{P}(e_x^t), e_y^t) \tag{8}$$

To make the formulation simple and clear, in the following derivations, we temporarily omit the superscript t until discussing the optimization method.

3.2. Losses

We implement the loss of JEPA ($\mathcal{L}$ in Equation 8) from two aspects, the regularization and contrastive losses.

Regularization loss. The regularization loss can be defined by the L-2 distance:

$$\mathcal{L}_{\text{reg}} = |e_{x\to y} - e_y|_2^2 \tag{9}$$

where $|\cdot|_2$ denotes the L-2 norm.

Contrastive loss. We conduct the contrastive learning by sampling the in-batch negatives:

$$\mathcal{L}_{\text{cl}} = \frac{1}{B} \sum_{i=1}^{B} \left[-\log \frac{\exp(\text{sim}(e_{x\to y}^i, e_y^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(e_{x\to y}^j, e_y^j)/\tau)} \right] \tag{10}$$

where B is the batch size, the superscripts i and j represents the i-th or the j-th sample within the batch. $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is a temperature parameter that controls the sharpness of the similarity distribution.

The total loss. The total loss is the linear combination between the regularization loss and contrastive loss:

$$\mathcal{L}(e_{x\to y}, e_y) = \alpha \mathcal{L}_{\text{reg}} + (1 - \alpha)\mathcal{L}_{\text{cl}} \tag{11}$$

in which α is the loss weight which determines the weighting balance between two loss terms.

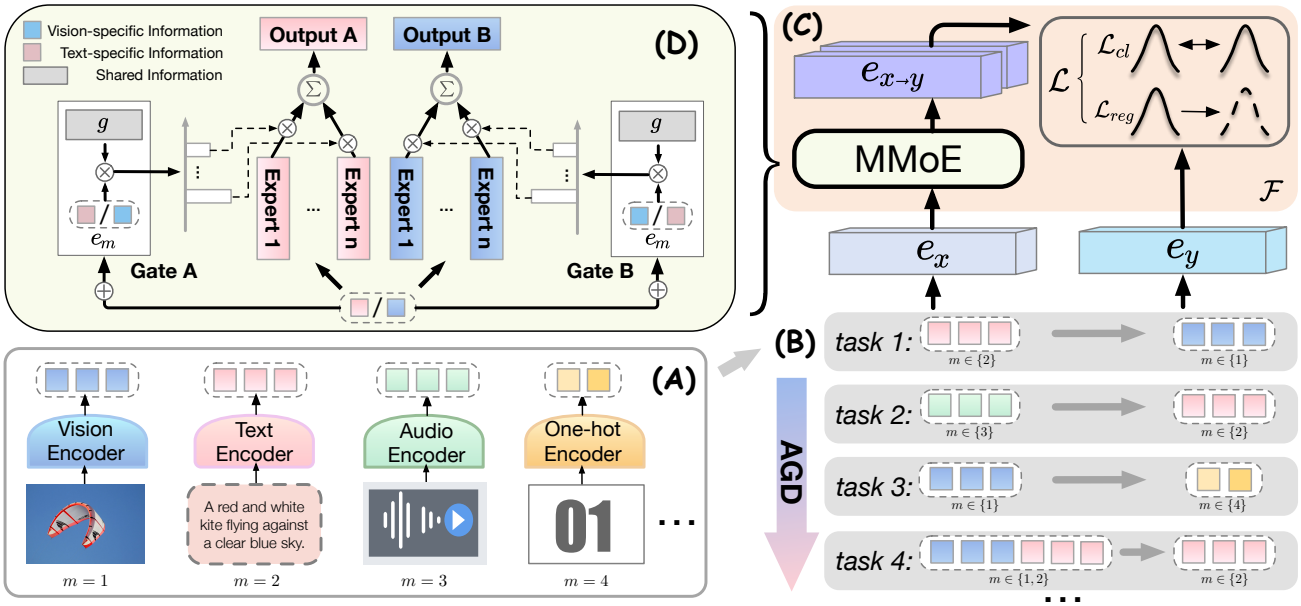

Figure 2. The entire framework of M3-JEPA with 4 modalities and 4 tasks indicated.
(A): Information is encoded by modality encoders ($m = 1, 2, 3, 4$), forming the input and output signals.
(B): Training is conducted by the alternative gradient descent (AGD), with different multimodal tasks switched stepwise. Tasks 1, 2, 3, 4 in the figure represent text-to-image, audio-to-text, image classification, and visual question-answering, respectively.
(C): Input and output are aligned on the latent space with an energy ($\mathcal{F}$), consisting of both contrastive (cl) and regularization (reg) losses.
(D): A Multi-gate MoE predictor projects the input latent vector to the output latent space, and generates parallel outputs (A and B) for different loss terms. Gating functions automatically separate modality-specific (e_m) and shared information (g).

3.3. Implementation of Predictor

In this paper, we implement the predictor of JEPA with a top-K MoE module:

$$e_{x\to y} = \mathcal{P}(e_x) := \text{MoE-K}(e_x) \tag{12}$$

For each modality, we implement N experts, which results in totally $M * N$ experts ($\mathbb{E}_N^M$). Furthermore, we implement the gating function with the following formulation

$$\mathbb{G} = \text{softmax}(g \cdot [e_x \oplus e_m]), m \in \{m_x^t\} \tag{13}$$

where $e_{1:M}$ is a learnable complementary matrices which convert the input of the gating function into a constant latent dimension, h. As a result, the gate matrix g is task-agnostic.

Disentangle modality-specific and shared information. Our MoE-predictor not only serves as a multimodal aligner, but also harmonize shared semantic information across modalities while preserving modality-specific details. This disentanglement is implemented with different terms in Equation 13:

- Modality-specific paths: the modality-specialized experts, as well as the gating embedding e_m.
- Shared representation: the projection matrix g creates a common subspace for the cross-modal gating.

Multi-gating on losses. We further expand our formation to the setting of MMoE architecture. In M3-JEPA, the multi-gate functions are employed to represent two loss terms, the regularization loss in Equation 9 and the contrastive loss in Equation 10. As a result, the number of gates $L = 2$.

3.4. Alternating Gradient Descent

At different training steps, we switch between T tasks and conduct the forward pass and back-propagation of the trainable parameter θ sequentially:

$$\theta(i+1) \leftarrow \theta(i) - \eta \nabla_\theta \mathcal{L}^t(\text{MMoE-K}(e_x^t(i)), e_y^t(i)) \quad \text{if } \text{mod}(i, T) = t, \quad t = 1, 2, \cdots, T \tag{14}$$

where i is the current step, η is the learning rate, and $\text{mod}(\cdot, \cdot)$ is the reminder after integer division. This optimization paradigm is called alternating gradient descent (AGD) with similar implementation on previous multimodal studies (Likhosherstov et al., 2022; Akbari et al., 2023).

The conventional joint optimization requires simultaneous updates to overlapping parameter subsets (e.g., a single con-

nector aligns both image-to-text and text-to-image), leading to potential gradient conflict (Javaloy et al., 2022). In contrast, AGD decouples these updates, mirroring the success of alternating training in multi-task learning.

4. The Theoretical Analysis

In this section, we show that M3-JEPA optimizes the information content by simultaneously maximizing the mutual information and minimizing the conditional entropy. We further discuss the optimality of loss weight, as well as the convergence condition.

4.1. The Information-Theoretic Analysis

We start this section with the information-theoretic analysis. To simplify the problem, we take the vision-language learning as an example[1], with $M = T = 2$. We then propose a new perspective of loss objective:

$$\theta^* \leftarrow \arg\min_{\theta} -\mathcal{I}(x;y) + \alpha\left(\mathcal{H}(y|x) + \mathcal{H}(x|y)\right) \quad (15)$$

in which $\mathcal{I}(x;y)$ is the mutual information between x and y, while $\mathcal{H}(y|x)$ and $\mathcal{H}(x|y)$ are their conditional entropies. COMPLETER (Lin et al., 2021) proves the connection between these terms and different loss types:

- Redundancy reduction: the contrastive loss $\mathcal{L}_{cl}$ maximizes $\mathcal{I}$ by separating negatives.
- Uncertainty reduction: the regularization loss $\mathcal{L}_{reg}$ minimizes $\mathcal{H}$ by regressing the positives.

Connection with the original loss. The above statements ensure that minimizing the averaged task-specific losses ($\mathcal{F}(x,y), \mathcal{F}(y,x)$) in Equation 11 is equivalent to the information-theoretic objective in Equation 15. Also, the loss weight α in both equations is consistent, balancing the compression ($\mathcal{L}_{cl}$ and $\mathcal{I}$) and predictiveness ($\mathcal{L}_{reg}$ and $\mathcal{H}$).

Information decomposition and coupling. Conditional entropy $\mathcal{H}(y|x)$ and $\mathcal{H}(x|y)$ represent the modality-specific information in text and image, respectively; while the mutual information $\mathcal{I}(x;y)$ quantifies the shared information between image and text. For a reasonable alignment, high mutual information (more shared content) and low conditional entropy (less modality-specific noises) are desirable, ensuring strong information coupling across modalities.

4.2. The Optimal Loss Weight

In this subsection, we derive the optimal loss weight from two different aspects, both of which suggest the optimal $\alpha = 0.5$. It is of significance to note that this theoretical conclusion is also validated by the subsequent empirical result (Figure 4).

Connection with the free energy. Loss in Equation 15 mirrors the free energy minimization principle

$$F = U - TS$$

in which $\mathcal{I}(x;y)$ corresponds to the internal energy U, and $\mathcal{H}(y|x) + \mathcal{H}(x|y)$ corresponds to the entropy TS. The critical temperature T_c where the energy/entropy balance occurs corresponds to $\alpha = 0.5$.

Derivation from the convergence assumption. For simplicity, we show this derivation with the case of two tasks, $x \to y$ and $y \to x$. Given a large enough step i, it is reasonable to assume that the losses of two consecutive steps converge to each other. Then the stable total loss can be approximated by

$$\begin{aligned}\mathcal{L} &\to \frac{1}{2}(\mathcal{L}(x \to y) + \mathcal{L}(y \to x)) \\ &\to \frac{1}{2}\big(-\mathcal{I}(x;y) + \mathcal{H}(y|x) - \mathcal{I}(y;x) + \mathcal{H}(x|y)\big) \\ &= -\mathcal{I}(x;y) + \frac{1}{2}\big(\mathcal{H}(y|x) + \mathcal{H}(x|y)\big) \end{aligned} \quad (16)$$

based on the fact $\mathcal{I}(x;y) = \mathcal{I}(y;x)$. Comparison between Equation 11, 15 and 16 indicates the optimal $\alpha = 0.5$.

4.3. Convergence of AGD on M3-JEPA

We assume for each task t, minimizing the task-specific loss in Equation 11 is independent and convex. From this assumption, from previous derivation (Jain & Kar, 2017; Pascal et al., 2022; Wibisono et al., 2022), AGD on multiple tasks (Equation 14) is guaranteed the convergence to a local optimum if each subtask is convex and optimally solved.

5. Experiment

To empirically verify our theoretical claims of M3-JEPA, we design the experiments to address the following research questions:

RQ1: Can M3-JEPA align well on the cross-modal representation on typical multimodal tasks?
RQ2: Can the same architecture be arbitrarily expanded to more forms of information (other than well-studied text, image and audio)), or generalized to unseen data and domains?
RQ3: Can M3-JEPA take multiple-modality as input (or output), instead of single-modality?
RQ4: Are MoE, finetuning and AGD all reasonable components of M3-JEPA?
RQ5: Are both contrastive and regularization losses necessary to achieve the optimality, and how about their weights?

[1] In this case, both x and y consist of only one modality. Two related tasks (image $\to$ text and text $\to$ image) can be simply represented by $x \to y$ and $y \to x$.

Table 1. Finetuned results on Vision-Language Retrieval tasks.

Method	# Trainable Params	Flickr30K						COCO					
		Image → Text			Text → Image			Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Lightweight models													
TinyCLIP (Wu et al., 2023)	63M+31M	84.9	-	-	66.0	-	-	56.9	-	-	38.5	-	-
MobileCLIP (Vasu et al., 2024)	<30.7M	85.9	-	-	67.7	-	-	58.7	-	-	40.4	-	-
Dual-encoder models													
CLIP (Radford et al., 2021)	428M	88.0	98.7	99.4	68.7	90.6	95.2	-	-	-	-	-	-
ALIGN (Cohen, 1997)	820M	88.6	98.7	99.7	75.7	93.8	96.8	77.0	93.5	96.9	59.9	83.3	89.8
FILIP (Yao et al., 2022)	417M	89.8	99.2	99.8	75.0	93.4	96.3	78.9	94.4	97.4	61.2	84.3	90.6
Florence (Yuan et al., 2021)	893M	90.9	99.1	-	76.7	93.6	-	81.8	95.2	-	63.2	85.7	-
BEIT-3 (Wang et al., 2023b)	1.9B	94.9	99.9	100.0	81.5	95.6	97.8	84.8	96.5	98.3	67.2	87.7	92.8
Fusion-encoder models													
UNITER (Chen et al., 2020)	303M	83.6	95.7	97.7	68.7	89.2	93.9	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	70.0	91.1	95.5	54.0	80.8	88.5
VinVL (Zhang et al., 2021)	345M	-	-	-	-	-	-	75.4	92.9	96.2	58.8	83.5	90.3
Dual encoder + Fusion encoder													
ALBEF (Li et al., 2021)	233M	94.1	99.5	99.7	82.8	96.3	98.1	77.6	94.3	97.2	60.7	84.3	90.5
BLIP (Li et al., 2022)	446M	97.1	100.0	100.0	86.7	97.3	98.7	82.4	95.4	97.9	65.1	86.3	91.8
BLIP-2 w/ ViT-L (Li et al., 2023)	474M	96.9	100.0	100.0	88.6	97.6	98.9	83.5	96.0	98.0	66.3	86.5	91.8
BLIP-2 w/ ViT-g (Li et al., 2023)	1.2B	97.6	100.0	100.0	89.7	98.1	98.9	85.4	97.0	98.5	68.3	87.7	92.6
Ours													
M3-JEPA	140M	**97.8**	**100.0**	**100.0**	**97.8**	**100.0**	**100.0**	**87.7**	**99.6**	**99.9**	**89.7**	**99.7**	**99.9**

RQ6: Is M3-JEPA a computational efficient framework for both training and inference?

In the following subsections, we discuss different experimental results and answer the above questions.

5.1. Setting

M3-JEPA employs pretrained uni-modal encoders and connects their latent spaces with an MoE. We use LLama3-8B (Dubey et al., 2024), Dinov2-Large (Oquab et al., 2025) and LanguageBind (Zhu et al., 2024) to encode text, image and audio modalities, respectively. We select 3 layers of the modality encoders to be finetuned by LoRA (Hu et al., 2022) (with rank=64), while keep the rest of parameters frozen. We implement an MMoE predictor with $N = 12$, $K = 4$ and $L = 2$. The inner hidden size (h) is 2048 and a dropout rate is set to 0.1. The MoE predictor is initialized randomly with full parameters updated during the training.

All tasks have the batch size of 128, solved by the Adam optimizer with the lr schedule of cosine, warmup of 0.1 and weight decay of 0.005. For retrieval tasks, we evaluate recall-based metrics including R@1, R@5 and R@10. For classification tasks, we provide metrics such as Accuracy, Precision, Recall and F1 score. Further details of implementation, datasets and benchmarks are summarized in Appendix A.

5.2. Vision-Language Retrieval

To answer **RQ1**, we first validate M3-JEPA on image-text retrieval tasks, including COCO (Lin et al., 2014b) and Flickr30K (Plummer et al., 2015) and evaluate it on the test set. Table 1 shows the experimental results, compared to previous state-of-the-art baselines across different architectures. M3-JEPA obtained superior performance on both image-to-text and text-to-image tasks. This observation indicates that our framework achieves stronger cross-modal alignment on the latent space, capturing different levels of abstraction on the cross-modal relationship.

M3-JEPA also demonstrates superior computational efficiency by observing the size of training parameters. To partially address **RQ6**, notice that M3-JEPA has only 140M trainable parameters, which is significantly smaller than the 1.2B BLIP-2, the second-best method in Table 1.

In Figure 3, we also visualize the similarity matrix for 10 text-image pairs from COCO, as a snapshot of the image-text retrieval performance. It shows that M3-JEPA can differentiate positive pairs (the diagonal grids) and negative pairs (the non-diagonal grids) well, which ensures retrieval performance.

5.3. Audio-Language Retrieval

In address **RQ2**, we attempt to adapt M3-JEPA to a new modality and simultaneously examine its generalization ability. In more detail, we experiment on audio-text retrieval tasks by replacing the image encoder with the audio encoder. For a fair comparison, we inherit the same experimental setting from LanguageBind (Zhu et al., 2024), *i.e.*, training on a held-out audio-language dataset (including wavtext5k (Deshmukh et al., 2023) and freesound (Fonseca et al., 2022)) while evaluate on Clotho (Drossos et al., 2020)

Table 2. Audio-text retrieval results. Results of AVFIC, ImageBind and VALOR are obtained from Zhu et al. (2024) directly. We download the original model of LanguageBind and evaluate it by ourselves to collect the results of all metrics.

Method	Clotho						Audiocaps					
	Audio → Text			Text → Audio			Audio → Text			Text → Audio		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
AVFIC (Nagrani et al., 2022)	-	-	-	3.0	-	17.5	-	-	-	8.7	-	37.7
ImageBind (Girdhar et al., 2023)	-	-	-	6.0	-	28.4	-	-	-	9.3	-	42.3
VALOR (Liu et al., 2025)	-	-	-	8.4	-	-	-	-	-	-	-	-
LanguageBind (Zhu et al., 2024)	16.1	39.9	**53.2**	15.5	38.6	51.7	17.8	47.3	64.0	16.5	48.7	64.6
M3-JEPA (ours)	**17.0**	**40.8**	53.0	**20.1**	**45.2**	**58.7**	**20.4**	**50.8**	**66.6**	**19.8**	**51.4**	**66.8**

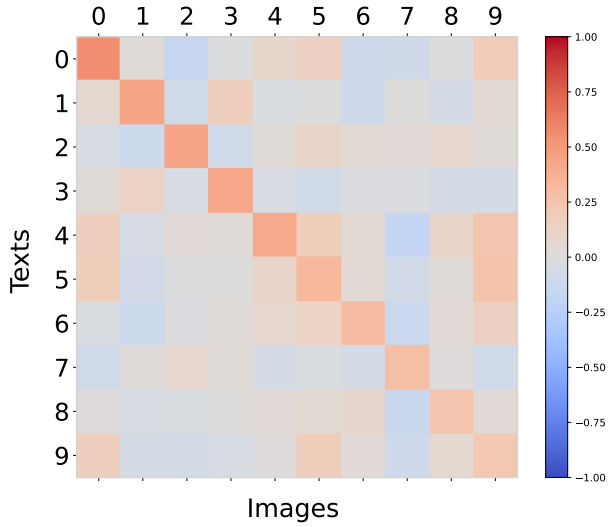

Figure 3. The similarity matrix of the text-image pairs. The colors indicate different levels of retrieval scores. The up-triangle part indicates text-to-image and the down-triangle part indicates image-to-text. The diagonal grids correspond to the ground truth pairs.

and Audiocaps (Kim et al., 2019) in a zero-shot manner[2]. Since LanguageBind (Zhu et al., 2024) provides limited results of metrics (only R@1 and R@10 on text-to-audio), we re-run the evaluation test of LanguageBind on all text-to-audio and audio-to-text metrics[3]. As shown in Table 2, M3-JEPA still demonstrates superior performance over all baselines, showcasing its alignment capability in the audio modality and an exceptional level of generalization. In this zero-shot setting, M3-JEPA also becomes extremely sensitive to the dataset bias (*i.e.*, the ratio of freesound to wavtext5k) and their sampling intervals and lengths, due to the sparsity of audio samples.

[2]One can refer to Zhu et al. (2024) for more details of training datasets and other settings.

[3]For overlapping metrics, our results are close to the originally reported values in (Zhu et al., 2024) and do not change the conclusion.

Table 3. Image classification results on ImageNet-1K. All results are in percentage.

Method	Accuracy	Precision	Recall	F1 score
CLIP-ViT (Radford et al., 2021)	82.1	82.4	82.0	82.0
DinoV2 (Oquab et al., 2025)	83.2	83.5	83.3	83.1
M3-JEPA (ours)	**86.6**	**86.9**	**86.6**	**86.5**

5.4. Vision Classification

This subsection further addresses **RQ2** by expanding to the image classification task. In this situation, we assume M3-JEPA can still learn well, by treating the classified labels as another modality. To validate this hypothesis, we train and evaluate M3-JEPA on ImageNet-1K (Deng et al., 2009) and compare it to DinoV2 (Oquab et al., 2025) and CLIP-ViT (Radford et al., 2021). Specifically, DinoV2 is vision-only, while CLIP-ViT is a text-guided vision model; they capture distinct types of information and offer complementary perspectives of performance comparison, but may need explicit classification heads. Unlike these methods, we encode the classification labels by one-hot, such that they can be involved in a self-supervised manner as a separate modality. As shown in Table 3, M3-JEPA outperforms DinoV2 and CLIP-ViT on all classification metrics, indicating the potential of M3-JEPA to represent the inherent knowledge of the natural world, not limited to the traditionally studied modalities (text, image, audio, etc).

5.5. Vision-Language Understanding

To address **RQ3**, we examine M3-JEPA in a more challenging scenario, where the input or output consists of multiple modalities, instead of uni-modality. To test M3-JEPA's adaptation ability to this situation, we conduct the Visual Question Answering (VQA) task, in which the model is concurrently prompted with an image and textual question, and is expected to provide a reasonable textual answer. In this situation, we integrate the image encoding and text encoding by simple concatenation into the input, then feed it into the MMoE predictor. The rest of the algorithm pipeline

is kept the same.

Training and evaluation are performed on VQAv2 (Goyal et al., 2017) and NLVR-2 (Suhr et al., 2019). Results are exhibited in Table 4. M3-JEPA adapts well to this multimodal input scenario, obtaining the second-best result obtained for each split of test sets. Although results of M3-JEPA are lower than BEiT-3(Wang et al., 2023b), it might be due to the large pretrained corpus of BEiT-3, including MSCOCO (Lin et al., 2014a) and Visual Genome (Krishna et al., 2017)). Furthermore, M3-JEPA could be further improved by implementing smarter fusion of multimodal information (*e.g.*, cross-attention between input modalities), instead of simple concatenation. We will leave this further attempt to future work. To better discuss this observation, we put a detailed bad case analysis in Appendix B.3.

Table 4. VQA scores on VQAv2 and NLVR-2. For each test set, the bold number indicates the best result and the underlined number indicates the second best.

Method	VQAv2		NLVR-2	
	test-dev	**test-std**	**dev**	**test-P**
ALBEF (Li et al., 2021)	75.8	76.0	82.6	83.14
BLIP (Li et al., 2022)	78.3	78.3	82.2	82.2
X-VLM (Zeng et al., 2022)	78.2	78.4	84.4	84.8
SimVLM (Wang et al., 2022b)	80.0	80.3	84.5	85.2
OFA (Wang et al., 2022a)	82.0	82.0	-	-
Flamingo (Alayrac et al., 2022)	82.0	82.1	-	-
CoCa (Yu et al., 2022)	82.3	82.3	86.1	87.0
BLIP-2 (Li et al., 2023)	82.2	82.3	-	-
BEiT-3 (Wang et al., 2023b)	**84.2**	**84.0**	**91.5**	**92.6**
M3-JEPA (ours)	<u>82.3</u>	<u>82.5</u>	<u>86.8</u>	<u>87.6</u>

5.6. Analysis and Discussions

We conduct further analysis, including ablation and sensitivity studies, as well as an efficiency analysis, to address **RQ4**, **RQ5** and **RQ6**, and obtain a deeper insight on M3-JEPA.

Ablation on the M3-JEPA approach. To validate the design of the MoE predictor design, we replace it with a comparable-size MLP, with results shown in the first row of Table 5. Compared to the last row (the formal M3-JEPA), a significant drop in performance is observed, demonstrating the effectiveness of MoE.

Table 5. Ablation of the M3-JEPA approach on COCO.

MoE	AGD	Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
×	✓	74.4	86.0	92.2	82.3	89.5	92.6
✓	×	68.2	68.7	81.1	74.2	88.7	92.4
✓	✓	**87.7**	**99.6**	**99.9**	**89.7**	**99.7**	**99.9**

Additionally, we examine the impact of the AGD approach. Ablating it to a non-alternating optimization also leads to significant performance degradation, as indicated by the second row of Table 5.

Ablation on finetuning of modality encoder. During training, one can choose the uni-modality encoder to be completely frozen, finetuned on its full layers or part of layers (we test for 3 layers). We show this ablation result in Table 6. Results indicate that finetuning has a positive impact on both performances of image-to-text and text-to-image, while full-layer LoRA generally performs better than 3-layer LoRA. Nevertheless, full-layer LoRA is also subject to a higher time cost. As a result, we apply 3-layer LoRA in all the aforementioned formal experiments, and we achieve state-of-the-art performance on vision-language and audio-language tasks with only 3 layers finetuned.

Table 6. Ablation of modality encoder finetuning on COCO.

Approach	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
freeze	75.4	88.6	94.5	84.3	90.1	97.8
3-layer LoRA	87.7	**99.6**	**99.9**	89.7	99.7	**99.9**
full-layer LoRA	**92.1**	99.4	**99.9**	**91.1**	**99.8**	**99.9**

Sensitivity analysis on the loss weight. We then conduct the sensitivity study on an important hyper-parameter, the loss weight α between $\mathcal{L}_{cl}$ and $\mathcal{L}_{reg}$. The experiment is conducted on the text-to-image task of COCO, with the R@1 results shown in Figure 4. One can observe that the best performance is achieved when $\alpha = 0.5$, indicating that both contrastive loss and regularization loss are necessary. We then select 0.5 as the formal choice of α in all formal experiments, which also empirically verifies the theoretical conclusion in Section 4.2.

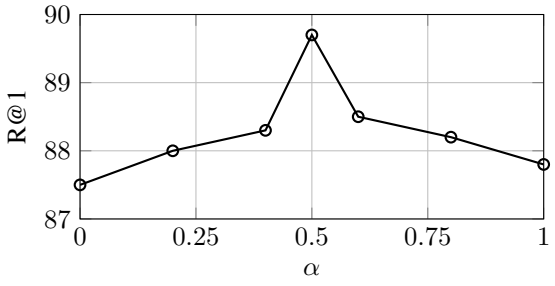

Figure 4. Sensitivity plot of the loss weight α.

Analysis of computational efficiency. In order to answer **RQ6**, here we discuss the computational efficiency of M3-JEPA from the aspects of training and inference costs, respectively. Although M3-JEPA has relatively heavy modality encoders, its training cost is relatively low, since only the lightweight MMoE predictor and several layers of encoders (3 layers in the formal experiment) need to be trained, while

the rest parts of encoder parameters are freeze. To elaborate the comparison of training costs to baselines, we also include the # of trainable parameters in Table 1.

For the inference cost, one should notice that M3-JEPA supports the modality precomputing and online caching, while the cross-modality alignment only happens in the latent space of the MMoE predictor (which is relatively lightweight compared to the uni-modality encoders), significantly reducing the memory overhead. To illustrate this difference, we calculate the averaged retrieval time (RT) of M3-JEPA on COCO, a typical image-text retrieval task. We observe the RT of M3-JEPA is only 0.02 seconds, comparing to CLIP, a classical dual-encoder approach, with an RT of 0.16s, and BLIP-2, which relies on a lightweight Q-former, with an RT of 0.05s. These results indicate M3-JEPA with modality pre-computing is much more computationally efficient, and can be potentially applied as an efficient multimodal retriever on massive documents.

For inference with dynamic inputs (*i.e.*, user-provided images/text) which are intractable to be precomputed, M3-JEPA's retrieval latency is dominated by the modality encoder inference (approximately 0.1s/image for DINOv2, 0.3s/text for LLaMA-3-8B). In such a situation, it is recommended to cache frequent queries (e.g., common prompts in retrieval systems), for latency-sensitive applications.

6. Related Works

6.1. Multimodal Learning

Modern multimodal Learning has achieved outstanding performance, which can be classified into two main categories: training end-to-end, or training a connector based on pretrained unimodal models. The end-to-end multimodal models can have various architectures, such as dual-encoder (Radford et al., 2021; Jia et al., 2021) or fusion encoder (Li et al., 2020; Chen et al., 2020; Jia et al., 2021; Wang et al., 2022a). Such end-to-end pre-training consumes large-scale multi-modal datasets. As the model scale continues to increase, it may incur prohibitively high computational costs and often struggle to adapt to novel modalities or tasks without extensive re-training.

Another category of multimodal learning focused on integrating unimodal models to achieve high-quality multimodal alignment (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Zhang et al., 2024). For example, Flamingo (Alayrac et al., 2022) integrates visual information into each layer of a frozen Large LLM through the use of cross-attention. BLIP-2 (Li et al., 2023) introduces a Q-former before feeding into the LLM, and propose also a two-stage training process. LLaVA (Liu et al., 2023) aligns vision and language representation by employing a two-layer multilayer perceptron (MLP). M3-JEPA further advances this concept by interconnecting vision, text and audio encoders through a multi-gate MoE. We design the gating function to separate the model-specific and shared information to optimize the cross-modal performance.

6.2. Multimodal Models with MoE Structures

There are also pioneer works that integrate multimodal models with MoE-like structures. For instance, LIMoE (Mustafa et al., 2022) feeds both images and text signals into a single MoE encoder, trained by a contrastive loss. VL-MoE (Shen et al., 2023) uses modality-specific experts for image-text modeling. MoE-LLaVA (Lin et al., 2024) proposes MoE-Tuning, a training strategy for large vision-language models, and construct a sparse model with a constant computational cost. In contrast, M3-JEPA adopts the multi-gate MoE architecture, which simultaneously feeds parallel output terms to both contrastive and regularization losses.

6.3. Multimodal Models Trained by AGD

Although most multimodal learning is trained jointly and concurrently, there have also been attempts that train the model by alternating gradient descent (AGD), *i.e.*, alternating back-propagation across different tasks. Such efforts include Polyvit (Likhosherstov et al., 2022) and IMP (Akbari et al., 2023), revealing that the alternation of diverse modalities, tasks, and resolutions may enhance the model's cross-domain performance and generalization capabilities. In our study, we apply AGD on the joint-embedding-predictive architecture, in which the gradient descent happens on the latent space, which further mitigates the gradient conflict issue from cross-modality samples.

7. Conclusion

This paper introduces M3-JEPA, a novel any-to-any multimodal training framework, which is lightweight, scalable, and computationally efficient. M3-JEPA effectively integrates pretrained unimodality encoders, projects the input embedding into the output embedding space through a multi-gate MoE predictor, and finally aligns the cross-modal information on the latent space. The MoE predictor includes two gates, corresponding to the contrastive and regularization losses, while each gating function is designed with separated model-specific and shared learnable encodings. The M3-JEPA is optimized by alternating the multi-directional multimodal tasks, facilitating the cross-modal alignment in multiple directions. By theoretical derivation, we demonstrate that M3-JEPA is information-theoretic optimal, while its convergence can be ensured with reasonable assumptions on the subtasks. Extensive experiments finally demonstrate that M3-JEPA can achieve state-of-the-art performance on different types of tasks and modalities, and generalize well on unseen datasets and domains.

Impact Statement

In this paper, we apply JEPA to multimodal tasks, including vision-language retrieval, audio-language retrieval, image classification, and vision-language understanding. We implement the predictor of JEPA with a multi-gate MoE architecture, which attempts to separate the model-specific and model-shared information. Two gates of MoE generate parallel outputs which are used to calculate the contrastive loss and the regularization loss, respectively. We name this framework as M3-JEPA, and optimize it by AGD, an alternative optimization strategy between different multimodal tasks. An information-theoretic analysis is provided, which connects our loss with maximization of the mutual information and minimization of the conditional entropy. We also discuss the convergence of AGD, and the optimal loss weight in this formulation. M3-JEPA achieves state-of-the-art performance across various tasks and modalities, and also exhibits strong generalization and high computational efficiency. This work opens new directions for self-supervised multimodal learning and open-world understanding.

References

Akbari, H., Kondratyuk, D., Cui, Y., Hornung, R., Wang, H., and Adam, H. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79142–79154, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: unified vision-language pre-training with mixture-of-modality-experts. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Bardes, A., Ponce, J., and LeCun, Y. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features, 2023. URL https://arxiv.org/abs/2307.12698.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Cohen, G. H. Align: a program to superimpose protein coordinates, accounting for insertions and deletions. *Journal of applied crystallography*, 30(6):1160–1161, 1997.

Dawid, A. and LeCun, Y. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10), October 2024. ISSN 1742-5468. doi: 10.1088/1742-5468/ad292b. Publisher Copyright: © 2024 The Author(s).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Deshmukh, S., Elizalde, B., and Wang, H. Audio retrieval with wavtext5k and clap training. In *Interspeech 2023*, pp. 2948–2952, 2023. doi: 10.21437/Interspeech.2023-1136.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022.

Font, F., Roma, G., and Serra, X. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 411–412, 2013.

Garmash, E. and Monz, C. Ensemble learning for multi-source neural machine translation. In Matsumoto, Y. and Prasad, R. (eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1409–1418, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://aclanthology.org/C16-1133/.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Jain, P. and Kar, P. *Non-convex Optimization for Machine Learning*. 2017. doi: 10.1561/2200000058.

Javaloy, A., Meghdadi, M., and Valera, I. Mitigating modality collapse in multimodal VAEs via impartial optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9938–9964. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/javaloy22a.html.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jin, P., Zhu, B., Yuan, L., and Yan, S. Moh: Multi-head attention as mixture-of-head attention. *arXiv preprint arXiv:2410.11842*, 2024.

Jun Zhan, Junqi Dai, J. Y. Y. Z. D. Z. Z. L. X. Z. R. Y. G. Z. L. L. H. Y. J. F. T. G. T. S. Y. J. X. Q. Anygpt: Unified multimodal llm with discrete sequence modeling, 2024. URL https://arxiv.org/abs/2402.12226.

Kim, C. D., Kim, B., Lee, H., and Kim, G. AudioCaps: Generating captions for audios in the wild. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123 (1):32–73, May 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Likhosherstov, V., Arnab, A., Choromanski, K., Lucic, M., Tay, Y., Weller, A., and Dehghani, M. Polyvit: Co-training vision transformers on images, videos and audio. *Transactions on Machine Learning Research*, 2022.

Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Zhang, J., Ning, M., and Yuan, L. MoE-LLaVA: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In Fleet, D., Pajdla,

T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014a. Springer International Publishing. ISBN 978-3-319-10602-1.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014b.

Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., and Peng, X. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Liu, D., Zhang, R., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. In *Forty-first International Conference on Machine Learning*.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Liu, J., Chen, S., He, X., Guo, L., Zhu, X., Wang, W., and Tang, J. VALOR: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (2):708–724, 2025. doi: 10.1109/TPAMI.2024.3479776.

Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., and Chi, E. H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 1930–1939, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220007. URL https://doi.org/10.1145/3219819.3220007.

Ma, M., Ren, J., Zhao, L., Testuggine, D., and Peng, X. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18156–18165, 2022. doi: 10.1109/CVPR52688.2022.01764.

Mustafa, B., Riquelme, C., Puigcerver, J., Jenatton, R., and Houlsby, N. Multimodal contrastive learning with limoe: the language-image mixture of experts. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 9564–9576. Curran Associates, Inc., 2022.

Nagrani, A., Seo, P. H., Seybold, B., Hauth, A., Manen, S., Sun, C., and Schmid, C. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pp. 407–426. Springer, 2022.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. In *International Conference on Learning Representations*, 2025.

Pascal, L., Michiardi, P., Bost, X., Huet, B., and Zuluaga, M. A. Improved optimization strategies for deep multitask networks, 2022. URL https://arxiv.org/abs/2109.11678.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Shen, S., Yao, Z., Li, C., Darrell, T., Keutzer, K., and He, Y. Scaling vision-language models with sparse mixture of experts. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11329–11344, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.758. URL https://aclanthology.org/2023.findings-emnlp.758/.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Vasu, P. K. A., Pouransari, H., Faghri, F., Vemulapalli, R., and Tuzel, O. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15963–15974, 2024.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pp. 23318–23340. PMLR, 2022a.

Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., and Zhou, C. One-peace: Exploring one general representation model toward unlimited modalities, 2023a. URL https://arxiv.org/abs/2305.11172.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., and Wei, F. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19175–19186, 2023b. doi: 10.1109/CVPR52729.2023.01838.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=GUrhfTuf_3.

Wibisono, A., Tao, M., and Piliouras, G. Alternating mirror descent for constrained min-max games. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=w4X7GLThiuJ.

Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X. S., Wang, X., et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. NExT-GPT: Any-to-any multimodal LLM. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53366–53397. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/wu24e.html.

Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., and Faghri, F. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=Je5SHCKpPa.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cpDhcsEDC2.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Zeng, Y., Zhang, X., and Li, H. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25994–26009. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zeng22c.html.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., and Yu, D. MM-LLMs: Recent advances in MultiModal large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.738. URL https://aclanthology.org/2024.findings-acl.738/.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5579–5588, 2021.

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QmZKc7UZCy.

A. Extra Implementation Details

A.1. Image-Text Retrieval

For image-text retrieval, we examine M3-JEPA on two famous datasets, COCO and Flickr30K, with the detailed introductions below:

- COCO (Lin et al., 2014b): 330,000 images with object annotations and captions, providing a rich resource for multi-label image classification and visual understanding. It encompasses 80 object categories with 5,000 training and 1,000 testing images per category. The test set contains 5,000 samples.
- Flickr30K (Plummer et al., 2015): comprises over 30,000 images, each paired with five descriptive sentences. It is sourced from Flickr and reflects the diversity and complexity of real-world data, making it suitable for tasks such as image annotation, visual question answering, and image retrieval. The test set contains 1,000 samples.

For each dataset, we finetune M3-JEPA on its training set and evaluate it on its test set, respectively. The initial learning rate is 0.001 and the final learning rate is 5.5e-6.

A.2. Audio-Text Retrieval

Details of our audio-language datasets are as follows:

- Clotho (Drossos et al., 2020): 4,981 audio samples with 24,905 descriptions, sourced from the Freesound platform and crowdsourced from English-speaking contributors.
- Audiocaps (Kim et al., 2019): A curated subset of AudioSet, focusing on audio captions and enabling the study of audio-text relationships.
- Wavtext5k (Deshmukh et al., 2023): The WavText5K data was sourced from two main websites: BigSoundBank and SoundBible 3 (details can be found in (Deshmukh et al., 2023)). WavText5K contains 4505 audios, 4348 descriptions, 4525 audio titles, and 2058 tags.
- Freesound(Font et al., 2013): The Freesound dataset contains 363,618 samples, totaling 2,162.10 hours of audio.

To zero-shot evaluate the performance of M3-JEPA on audio-text retrieval, similar to Similar to LanguageBind (Zhu et al., 2024), we zero-shot evaluate the performance of M3-JEPA on Clotho and Audiocaps based on the knowledge of other datasets. In more details, we train M3-JEPA on the mixture of AudioCaps, WavText5K, and Freesound then test it on the Clotho dataset; then we train M3-JEPA on the mixture of Clotho, WavText5K, and Freesound then test it on AudioCaps dataset. The initial learning rate is 5.0e-4 and the final learning rate is 2.5e-6.

A.3. Image Classification

We train and test M3-JEPA on the popular benchmark, ImageNet-1K (Deng et al., 2009), which contain 1,281,167 training images, 50,000 validation images, and 100,000 test images across totally 1,000 classes. For image classification, the initial learning rate is 1.0e-3 and the final learning rate is 5.5e-6.

A.4. Visual Question Answer

We consider the following two datasets for the VQA task:

- VQAv2 (Antol et al., 2015): 265,016 images with multiple questions per image, assessing the ability of models to understand and answer questions about visual content.
- NLVR-2 (Suhr et al., 2019): 107,292 pairs of images with corresponding sentences, testing the visual reasoning capabilities of models.

Specifically, the VQA task requires the model to answer textual questions about input images. We starts from the pretrained version of M3-JEPA by the COCO datasets, then finetune it following previous works (Wang et al., 2023b;a; Bao et al., 2022), which formulate the VQA task as a textual answer retrieval problem. The initial learning rate is 2.0e-4 and the final learning rate is 2.4e-5.

B. More Experimental Results

B.1. Parameter Comparisons

To provide a more comprehensive discussion, here we summarize the parameter statistics of M3-JEPA, compared to vision-language model baselines, as listed in Table 7. Specifically, M3-JEPA can also be viewed as a lightweight knowledge connector, similar to BLIP-2. Therefore, although our M3-JEPA has a large total parameter size (around 8.5B), its trainable parameter is only 140M, even smaller than BLIP-2, which ensures its training efficiency.

B.2. Ablation of MoE Hyperparameters

Here we conduct the ablation studies on the structural parameters of MoE, including the number of experts n, and the top-k ranking mechanism. For n, we compare different values on VQAvq2, with results shown in Table 8. One can observe that a larger n can benefit the VQA performance, where $n = 12$ outperforms 2 and 8. However, increasing n

Table 7. Parameter statistics of vision-language methodologies.

Method	# total parameter	# trainable parameter
CLIP	428M	*the same*
ALIGN	820M	*the same*
FLIP	417M	*the same*
BEiT-3	1.9B	*the same*
UNITER	303M	*the same*
OSCAR	345M	*the same*
BLIP-2	4.1B	474M
M3-JEPA	8.5B	140M

also amplifies the computational cost, therefore we do not attempt a larger n. For k, we conduct the ablation study on COCO, with image-text retrieval results shown in Table 9. The results validate the optimality of our formal settings, which is $k = 4$.

Table 8. Ablation of n on the validation set of VQAvq2. The reported score is the accuracy of VQA answers.

n	2	8	12
score	55.15	59.84	68.03

Table 9. Ablation of k on COCO. The reported metric is R@1.

k	Flickr30K		COCO	
	Image → Text	Text → Image	Image → Text	Text → Image
2	96.0	95.5	85.0	82.0
4	**89.7**	**87.9**	**97.8**	**97.8**
6	88.0	86.5	97.5	97.0

B.3. Typical Failure Case

Currently, we found that M3-JEPA may suffer performance degradation in the existence of multimodality input or output, *e.g.* the VQA task. As discussed before, part of reasons may be the insufficient pretraining, as well as its simple concatenation strategy of modality input embeddings. To better demonstrate this phenomenon, here we exhibit a typical bad case of M3-JEPA on VQAv2. The input image is in Figure 5, while the textual questions, answers, and corresponding scores are shown in Table 10.

This example is challenging to M3-JEPA since it contains multiple visual objects as well as corresponding questions. As indicated in Table 10, M3-JEPA successfully predicts the number and color of the horse, but fails to identify the number of steps. This may be due to the simple concatenation of visual and question embeddings before passing to the MoE predictor. The MoE tends to focus on dominant objects (e.g., the horse), while failing to capture minor details like the stairs. A specifically designed mechanism, such as the cross-modal attention, or a unified positional embedding, may help capture finer-grained objects and alleviate such issues.

Figure 5. A typical failure case on the VQA task. The input image contains multiple visual objects (a horse and a staircase).

C. Limitation and Future Direction

In this paper, we propose M3-JEPA and demonstrate its effectiveness across different modalities and tasks. However, the following limitations still persist and hinder M3-JEPA from a more generalized foundation model:

- Generative capability: the joint-embedding-predictive architecture (JEPA) reformulate the next-token prediction into the alignment on the latent space, to filter the modality noise and capture the core cross-modal information. As a cost, JEPA is not a generative framework which prevents its wider applications. Nevertheless, we believe there might be potential solutions that are able to incorporate generative learning with JEPA.
- Modality expansion: different modalities have different information intensity. Therefore, popular modality encoders generally encode different modalities into various embedding dimensions. Built upon these pretrained modality encoders, M3-JEPA needs to adapt with these different input embedding dimensions. Although we have disentangled the modality-specific and shared components inside the architecture, the entire M3-JEPA framework can not be claimed as modality-agnostic. Introducing a new modality need to manually select its modality encoder, add the corresponding modality-specific expert, and redetermined the subsequent training pipeline. Incorporating the idea of meta or hyper networks into the MoE predicter may increase its adaptability to new information or modality, making it a true modality-agnostic framework.
- Generality as a world model: in the future, one may mask out an arbitrary part of world, and ask a generalized world model to predict the masked part, given the information of unmasked content. Without losing the generality, the masked and unmasked parts

Table 10. A typical failure case on the VQA task, which corresponds to the image in Figure 5. The bold answer indicated the model's predicted answer. M3-JEPA correctly answers questions about the house, but fails to identify the number of steps.

Question	Answer	Score
What kind of horse is this?	**brown and white**	**0.6**
	clydesdale	1.0
	others	0.0
How many horses are in the picture?	**1**	**1.0**
	others	0.0
How many steps to the building?	5	0.9
	10	0.3
	4	0.3
	6	**0.3**
	20	0.9
	others	0.0

should be assumed as a combination of different modalities, instead of a single modality. In this paper, we explore such a case, the VQA task. We show that simply concatenating the visual and question embeddings as the input can have reasonable performance, but is worse than the best baseline, BEiT-3. Better cross-modality encoding techniques may be needed to further enhance M3-JEPA's performance in such cases, such as Mixture-of-Heads (MoH) (Jin et al., 2024). Furthermore, the spatial-temporal representation and corresponding positional embedding may also be crucial to expand M3-JEPA in related scenarios, *e.g.*, and video comprehension.

- Possibility of other optimization: in this paper, we mainly apply AGD, an alternative optimization strategy between tasks, to train M3-JEPA. We conduct the ablation study of AGD and compare its result to a joint optimization, and suggests AGD's superiority. Nevertheless, we admit there are many details (*e.g.*, various tasks, modalities, losses) and different possible joint optimization implementations, which has not been explored by us. The ultimate optimization strategies can still be experimented.