M3-Jepa: Multimodal Alignment via Multi-directional MoE based on the JEPA framework

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

[1]Geely AI Lab, Geely Automobile Research Institute
[2]Peking University
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{hongyang.lei, xiaolong.cheng2, dan.wang35,yetao.wu, kun.fan1
qingqing.gu3, luo.ji1}@geely.com
qiqin@pku.edu.cn
hz.huang@siat.ac.cn

Abstract

Current multimodal alignment strategies primarily use single or unified modality encoders, while optimizing the alignment on the original token space. Such a framework is easy to implement and incorporate with the pretrained knowledge, but might result in information bias. To deal with such issues, the joint encoding predictive architecture (JEPA) learns the alignment loss on the latent space, with a predictor to convert the input encoding to the output latent space. However, the application of JEPA in multimodal scenarios is limited so far. In this paper, we introduce M3-Jepa, a scalable multimodal alignment framework, with the predictor implemented by a multi-directional mixture of experts (MoE). We demonstrate the framework can maximize the mutual information with information theory derivations, by alternating the optimization between different uni-directional 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 training and inference. Our study indicates that M3-Jepa might provide a new paradigm to self-supervised learning and open-world modeling.

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; 2023; Liu et al.). Modern multi-modal studies are mostly built on self-supervised learning (SSL) and generative architecture, which can be classified into two main categories. The first category aim to learn the multi-modal information by training from scratch (Wang et al., 2022a; 2023; Oquab et al., 2023). Such methodology asks for large-scale data therefore faces high training costs. It may also limit the acquisition of intra-modal information and introduce inter-modality information bias. For example, vision-language pre-training studies might be predominantly text-guided (Radford et al., 2021), which hinders the retainment of visual information due to inherent limitations of text descriptions (Oquab et al., 2023). Another category either employs the pretrained Large Language Model (LLM) as the backbone (Wu et al., 2023; Liu et al., 2023; Jun Zhan, 2024) or a lightweight cross-modal predictor (Alayrac et al., 2022; Li et al., 2023a; Bai et al., 2023) in order to preserve fine-grained uni-modality information and alleviate computational burdens (Zhang et al., 2024). Success of such methodologies indicates the "platonic representation" (Huh et al., 2024), which indicates representations across various domains

---

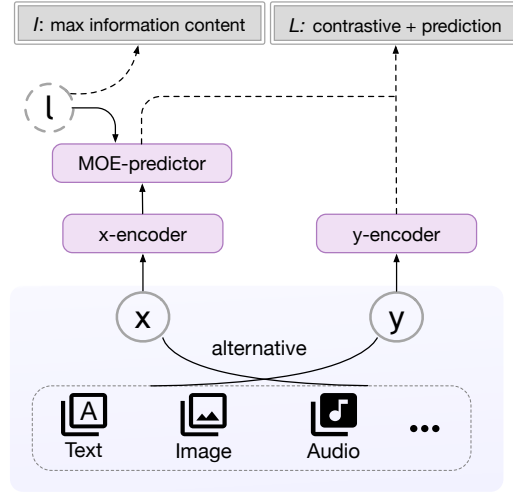*Equal contribution.
†Corresponding author.

Figure 1: Paradigm of our M3-Jepa. The self-supervised learning is conducted with two encoding branches of input and output signals, and a multi-directional MoE predictor to match the input latent embedding into the target latent embedding by minimizing contrastive and prediction losses. The MoE predictor is conditioned on the multi-modal routed information which is learned with information entropy minimization. Training is performed with alternative any-to-any multi-modality tasks.

are increasingly converging to the same latent space. Such representation similarity spans different model architectures, training objectives, and even data modalities.

Typical generative architecture tries to reconstruct the target signal conditioned by the input signal directly, usually with a decoder structure (Assran et al., 2023). However, such approach is not good at eliminating irrelevant details, and may face difficulty when input details are not easily predictable, or representing complex uncertainties in continuous spaces (Dawid & LeCun, 2023). It will also create bias given ambiguous semantic alignment of multi-modal samples. To take a step ahead on the representation convergence on multi-modalities, one may consider to reconstruct multi-modal signals in the latent-space, however, which might induce the representation collapse, especially when the latent variables contain excess information (Dawid & LeCun, 2023). Such consideration calls for a different framework with better alignment in the latent space, and reduced information content of latent variables.

In this work, we propose a novel **M**ultimodal alignment paradigm via **M**ulti-directional **M**oE based on **Jepa**, in short, **M3-Jepa**. It studies the dependency of the unobserved part (y) on the observed part (x) in their embedding space, in which x and y may belong to different modalities or any combination of them. We employ the joint-embedding predictive architecture (JEPA) (Dawid & LeCun, 2023; Assran et al., 2023) to align the representation of different modalities into the same space. JEPA achieves the alignment by first encoding the observed and unobserved parts, then use a predictor to convert one embedding into another embedding space, driven by the contrastive and predictive losses simultaneously. We implement the predictor by recapping the lightweight cross-modal predictor, which is designed as the Mixture-of-Experts (MoE) structure. We solve the framework by Alternating Gradient Descent (AGD) of different multi-modal tasks, similar to IMP (Akbari et al., 2023). To tackle the challenge of the multi-modal semantic discrepancy (paired cross-modal data may exhibit inconsistent information despite their intended correspondence), we decouple the cross-modal information into modality-independent and shared components through distinct experts. This multi-directional MoE is used to conduct multiple unidirectional alignment tasks and is solved by Alternating Gradient Descent (AGD) of different multi-modal tasks, similar with IMP (Akbari et al., 2023). The router information of MoE is the latent factor l which helps better alignment. To avoid the representation collapse, we solve the problem by minimizing the information content of predictor (Dawid & LeCun, 2023). Our **M3-Jepa** paradigm is depicted by Figure 1, and project codes can be found at `https://github.com/HongyangLL/M3-JEPA`.

To empirically validate these properties and explicitly evaluate the alignment effectiveness, we conducted comprehensive experiments across multiple multi-modal tasks, while keeping the model architecture and training strategy consistent. Experimental results demonstrate that our M3-Jepa achieves competitive results across different modalities and tasks, good generalization on unseen task and modality, compared to current state-of-the-art (SoTA) multi-modal studies. Furthermore, our approach is able to conduct large-scale online multi-modal retrieval tasks due to its unique architecture. We summarize our contributions as follows:

- We propose a novel modality-agnostic multi-modal alignment paradigm, with the alignment conducted on the latent space, which is computationally efficient especially when employed as a retriever.
- We leverage multi-directional MoE as the cross-modal predictor, optimizing by alternating the gradient descent between different unidirectional alignment tasks.
- We derive an information-theoretical explanation analysis, demonstrating the optimality of M3-Jepa.
- Our experimental results demonstrate remarkable multi-modal alignment accuracy and efficiency, encompassing text, image and audio modalities.
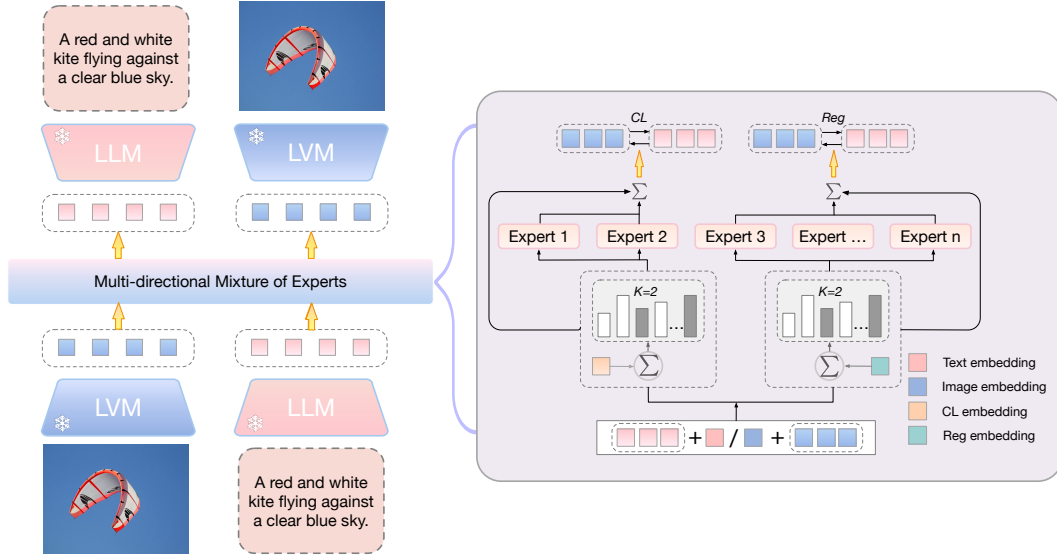

Figure 2: Architecture of M3-Jepa: input and output are encoded by modality encoders and aligned on the latent space. A predictor consisting of a multi-directional MoE is employed to project the input latent vector to the output space. The optimization is alternated between different tasks step by step, and both contrastive learning (CL) and regularization learning (Reg) are implemented by loss components. The text-vision tasks are depicted as an example in the figure.

2 PRELIMINARY

2.1 JOINT-PREDICTIVE EMBEDDING ARCHITECTURE

Given an input-output pair (I, O), we employ different encoders, denoted by $e_i(\cdot)$ and $e_o(\cdot)$. The encoding can be formally expressed as:

$$z_i = e_i(I), \quad z_o = e_o(O), \quad z_i, z_t \in \mathbb{R}^i, \mathbb{R}^o \tag{1}$$

where z_i, z_t are input and output's latent representations, with dimensions of i and o, respectively.

A joint-embedding predictive architecture (JEPA) (Dawid & LeCun, 2023; Assran et al., 2023) framework first converts the input encoding into the output's dimensional space by a predictor $p(\cdot)$, then minimizes the loss:

$$\min \mathcal{L}(z_{i \to o}, z_o), \quad z_{i \to o} = p(z_i) \in \mathbb{R}^o \tag{2}$$

2.2 MIXTURE-OF-EXPERTS

Mixture-of-Experts (MoE) architecture (Shazeer et al., 2017; Fedus et al., 2021; Lepikhin et al., 2021; Zuo et al., 2021) uses K feed-forward networks (FFN), namely "experts" denoted as $E_{k=1}^K$, each with its own set of learnable weights. An additional gating network G whose output is a sparse N-dimensional vector to route each token via a few of these experts.

Consider x_s as the input token representation in the sth position to the MOE layer comprising of the $\{E\}_{k=1}^K$ expert FFNs. Also, consider w_k^{in} and w_k^{out} to be the input and output projection matrices for i^{th} expert. Expert output $E_k(x_s)$ is given by:

$$E_k(x_s) = w_k^{out} \cdot GeLU(w_k^{in} \cdot x_s) \tag{3}$$

Consider $G(x_s)$ to be the softmax output of the gating network. The output of the MoE layer is given by:

$$MOE(x_s) = \sum_k^K \text{top-k}(G(x_s)_k)E_k(x_s) \tag{4}$$

where $G(x_s)_k$ denotes the probability of selecting expert E_k for x_s.

3 METHODOLOGY

This section illustrates the methodology of M3-Jepa. Figure 2 provides an example of the M3-Jepa architecture when applying on text-vision tasks.

3.1 LOSS

We implement the loss of JEPA by two terms: the regularization loss and the contrastive loss. By their definitions, when $z_{i \to o} \to z_o$, we ensure $\mathcal{L}$ also converge to zero, and $\mathcal{L}$ can be considered as the energy between I and O. As a result, our framework belongs to the energy-based model (EBM) with Equation 2 aims to minized the energy between input and output.

Regularization Loss. The regularization loss can be defined by the L-2 distance:
$$\mathcal{L}_{\text{Reg}} = |z_{i \to o} - z_o|_2^2 \tag{5}$$
where $|\cdot|_2$ denotes the L-2 norm.

Contrastive Loss. We conduct the contrastive learning by in-batch negatives:
$$\mathcal{L}_{\text{CL}} = \frac{1}{N} \sum_{n=1}^N \left[-\log \frac{\exp(\text{sim}(z_{i \to o}^n, z_o^n)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(z_{i \to o}^j, z_o^j)/\tau)} \right] \tag{6}$$
where N is the batch size, the superscripts n and j represents the n-th or the j-th sample within the batch. $\text{sim}(\cdot, \cdot)$ is the cosine similarity between two latent vectors, 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:
$$\mathcal{L}(z_{i \to o}, z_o) = \alpha \mathcal{L}_{\text{Reg}} + (1 - \alpha)\mathcal{L}_{\text{CL}} \tag{7}$$
in which α is the loss weight which determines the weighting balance between two loss terms.

3.2 PREDICTOR

We implement the predictor by a multi-directional MOE architecture.
$$p(z_i) = \text{MOE}_{\mathbb{T}}(z_i + e^{Reg} + e^{CL}) \tag{8}$$
in which $\mathbb{T}$ labels the task, while e^{Reg} and e^{CL} are learnable embeddings for two types of losses. For each expert, its learnable matrices has the following dimensions
$$w_k^{in} \in \mathbb{R}^{i_{\mathbb{T}} \times i_{\mathbb{T}}}, w_k^{out} \in \mathbb{R}^{o_{\mathbb{T}} \times i_{\mathbb{T}}} \tag{9}$$

3.3 IMPLEMENTATION OF MULTI-MODAL TASKS

We develop M3-Jepa as an any-to-any modality framework. Assume there are totally $(M_1, \ldots, M_m)$ modalities and m corresponding uni-modality encoders:

$$z_i^M = e^M(M_i) \tag{10}$$

and the framework is dealing with N multi-modal tasks

$$\mathbb{T}_n := \{M_{in}\} \to \{M_{out}\}, \quad in, out \in 1, \ldots, m, \quad n \in 1, \ldots, N \tag{11}$$

then for the n-th task, the latent vectors of inputs and outputs can be defined as

$$z_i = \text{concat}(z_{\{in\}}^M) \tag{12}$$

$$z_o = \text{concat}(z_{\{out\}}^M) \tag{13}$$

At different training steps, we switch between N tasks and conduct the alternating gradient descent (AGD):

$$\theta^{t+1} \longleftarrow \theta^t - \eta \nabla_{\theta^t} \mathcal{L}^t(\text{MOE}_{\mathbb{T}}(z_i^t), z_o^t), \quad \text{if } \text{mod}(t, N) = n \tag{14}$$

in which t is the current optimization step, and $\text{mod}(\cdot, \cdot)$ is the reminder after the integer division.

4 THE INFORMATION-THEORETIC ANALYSIS

In this section, we show that M3-Jepa maximizes mutual information and minimizes conditional entropy to obtain independent and shared information for unidirectional alignment.

4.1 ALTERNATING UNIDIRECTIONAL ALIGNMENT

Here we delineate the optimization objectives and parameter update procedures across various time steps t. By introducing AGD, we can alternately optimize unidirectional alignment at each time step t with the goals of maximizing mutual information, denoted as $I(I; O)$, and minimizing conditional entropies, denoted as $H(O|I)$ and $H(I|O)$. Given time step t, the objective function can be viewed as the following objective:

$$\mathcal{L} = -I(I; O) + H(O|I) \tag{15}$$

4.2 INFORMATION DECOMPOSITION AND ALIGNMENT

For cross-modal alignment, conditional entropy $H(O|I)$ and $H(I|O)$ represent modality-specific information in text and image respectively, measuring uncertainty in one modality after observing the other. Mutual information $I(I; O)$ quantifies shared information between image and text modalities, indicating how much knowing one reduces uncertainty about the other. For good alignment, high mutual information (more shared content) and low conditional entropy (less modality-specific information) are desirable, ensuring strong semantic coupling between modalities.

The multi-directional leverages the MoE router to automatically select different experts, optimizing for these two objectives. This approach helps decouple modality-specific information from shared information, potentially improving the balance between capturing unique modal features and cross-modal relationships.

5 EXPERIMENT

To explore a reasonable framework on multimodal modeling, we design a series of experiments to address the following research questions:

RQ1: Can M3-Jepa perform well on typical cross-modal alignment tasks?
RQ2: Can M3-Jepa be arbitrarily expanded to more modality and preserve somewhat generalization performance?
RQ3: How well can M3-Jepa generalize to unseen data and domains?

RQ4: Can more forms of information (other than well-studied text, image and audio) be learned and encoded by M3-Jepa?

RQ5: Can M3-Jepa take multiple-modality as input (or output), instead of single-modality?

RQ6: If multi-directional MoE, the alternative optimization and the finetuning approach are all reasonable components of M3-Jepa.

RQ7: If both contrastive and predictive losses are necessary to achieve the optimal performance, and what is the reasonable weight factor between them?

RQ8: Is M3-Jepa an 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 a multi-directional MoE. We use LLama3-8B, Dinov2-Large and LanguageBind(Zhu et al., 2023) to encode text, image and audio modalities, respectively. For uni-modal classification tasks such as image classification, we consider the label also as a modality and encode it by one-hot. Further implementation details can be found in Appendix B.

Different datasets and benchmarks are summarized by different tasks in the Appendix A. 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.

Table 1: Image-text retrieval results.

| Model | # Trainable Params | Flickr30K (1K test set) | | | | | | COCO Fine-tuned (5K test set) | | | | | |
| | | 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
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., 2021)	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., 2022b)	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 ViT-L (Li et al., 2023b)	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 ViT-g (Li et al., 2023b)	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**

5.2 VISION-LANGUAGE RETRIEVAL

To answer **RQ1**, we first validate M3-Jepa on image-text retrieval tasks, including COCO(Lin et al., 2014) 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 **RQ8**, 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. To Further illustrate the studied image-text similarity, we also visualize the similarity matrix in Appendix C.

5.3 AUDIO-LANGUAGE RETRIEVAL

In address **RQ2** and **RQ3**, 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 Zhu et al. (2023), *i.e.*, training M3-Jepa on a held-out audio-text dataset and

evaluate on Clotho (Drossos et al., 2020) and Audiocaps (Kim et al., 2019) in a zero-shot manner[1]. Baselines including AVFIC, ImageBind, VALOR and LanguageBind are considered here which are originally summarized by Zhu et al. (2023). Since Zhu et al. (2023) provide limited results of metrics (R@1 and R@10 on text-to-audio), we re-run the evaluation test of LanguageBind to obtain full metric results on both text-to-audio and audio-to-text tasks[2]. As shown in the table 2, M3-Jepa still demonstrates superior performance over all baselines, showcasing its alignment capability on audio modality and exceptional level of generalization.

Table 2: Audio-text retrieval results. Results of AVFIC, ImageBind and VALOR are obtained from Zhu et al. (2023) directly. We evaluate LanguageBind by ourselves in order to compare on more 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 (Chen et al., 2023)	-	-	-	8.4	-	-	-	-	-	-	-	-
LanguageBind (Zhu et al., 2023)	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**

5.4 Vision Task with Labels

This subsection further expands the modality by incorporating the image classification tasks, in which we consider the classified labels as another modality, with the one-hot encoder. To validate the hypothesis proposed by **RQ4**, we train and evaluate M3-Jepa on ImageNet-1K (Deng et al., 2009) and compare it to DinoV2 (Oquab et al., 2023) and CLIP-ViT (Radford et al., 2021), with classification heads added. 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.

As shown in Table 3, M3-Jepa outperforms DinoV2 and CLIP-ViT on all classification metrics, highlighting that M3-Jepa has the potential to study the inherent model of the world, not limited within the traditional modalities (text, image, audio, etc), in a self-supervised manner. Moreover, the vision-only DinoV2 performs better than the text-vision model CLIP-ViT, suggesting that the latter has a suboptimal cross-modal alignment on this scenario, and simply incrementing the modality might not be an optimal solution.

5.5 Vision-Language Understanding

To address **RQ5**, we examine M3-Jepa in a more challenging scenario, in which the framework input or output is multi-modality, 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 expected to provide a reasonable textual answer. In this situation, we integrate the encoding of the image and question by simple concatenation and feed it into the MoE 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 multi-modal input scenario, with the second-best result obtained for each test set split. Although results of M3-Jepa are lower than BEiT-3(Wang et al., 2022b), it might due to the large pretrained corpus of BEiT-3, including MSCOCO and Visual Genome. Furthermore, M3-Jepa might be further improved by attempting more smart integration of multi-modal encoding (*e.g.*, cross-attention), instead of simple concatenation. We will leave this further attempt to the future work.

5.6 Analysis and Discussions

We conduct further ablation and sensitivity studies, to address **RQ6** and **RQ7**:

---

[1]One can refer to Zhu et al. (2023) for more details of training datasets and other settings.

[2]For overlapping metrics, our results are close to the originally reported values in (Zhu et al., 2023) and do not change the conclusion.

Table 3: Image Classification Performance on ImageNet-1K.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CLIP-ViT (Radford et al., 2021)	82.1	82.4	82.0	82.0
DinoV2 (Oquab et al., 2023)	83.2	83.5	83.3	83.1
M3-Jepa (ours)	**86.6**	**86.9**	**86.6**	**86.5**

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

Method	VQA		NLVR-2	
	test-dev	test-std	dev	test-P
ALBEF (Li et al., 2021)	75.8	76.0	82.55	83.14
BLIP (Li et al., 2022)	78.25	78.32	82.15	82.24
X-VLM	78.22	78.37	84.41	84.76
SimVLM	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., 2023b)	82.2	82.3	-	-
BEiT-3 (Wang et al., 2022b)	**84.2**	**84.0**	**91.5**	**92.6**
M3-Jepa (ours)	82.3	82.5	86.8	87.6

Ablation on predictor Architecture. To validate the predictor design of multi-directional MoE, we replace it with an MLP, with results shown in the first row of Table 5. Compared with the last row (the formal M3-Jepa), a significant drop in performance can be observed, demonstrating the effectiveness of multi-directional MoE.

Table 5: Ablation of the M3-Jepa Approaches.

MoE	ALT	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
✓	✓	88.1	99.4	99.8	90.1	99.6	99.9

Ablation on the Alternative Training Mechanism. Additionally, we examined the impact of the alternating optimization approach (ALT). 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 Uni-Modal Encoder. During training, one can choose the uni-modal encoder to be completely frozen, or fine-tuned by LoRA (Low-Rank Adaptation) on full layers or N layers ($N = 3$). We also show this ablation result on Table 6. Results indicate that the full layer LoRA achieves the best result, , with R@1, R@5, and R@10 reaching 92.1%, 99.4%, and 99.9% for image-to-text, and 91.1%, 99.8%, and 99.9% for text-to-image. Nevertheless, full layer LoRA also has a higher time cost. As a result, we apply N-layer LoRA in all the aforementioned experiments, with only 3 layers LoRA obtaining the state-of-the-art performance on vision-language and audio-language tasks.

Sensitivity Analysis on the Loss Weight. We finally conduct the sensitivity study on an important hyper-parameter, the loss weight α between the contrastive loss and the prediction loss. Experiment is conducted on the text-to-image task of COCO, with the R@1 results shown in Figure 3. One can observe that the best performance is achieved when $\alpha = 0.5$, indicating that both CL and Pred losses are necessary. We also select 0.5 as the final choice in all aforementioned experiments.

Comparison on Retrieval Efficiency. According to **RQ8**, one should notice that M3-Jepa aligns cross-modality in latent space, supporting modality pre-computing and online caching, while the real-time retrieval only requires the computation of the lightweight predictor. As a result, M3-Jepa can be potentially applied as an efficient multi-modal retriever on massive documents. To illustrate this conclusion, we calculate the averaged retrieval time (RT) on image-text retrieval of M3-Jepa, compared to the classic dual-encoder method CLIP. By experimenting with the image-text retrieval task of COCO, we observe the RT of CLIP is 0.16 seconds while M3-Jepa has RT of 0.02 seconds, indicating M3-Jepa is a much more efficient retriever.

Table 6: Ablation of modality encoder finetuning.

Approach	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
freeze	75.4	88.6	94.53	84.3	90.1	97.8
N layers LoRA	88.1	**99.4**	99.8	90.1	99.6	**99.9**
full layer LoRA	**92.1**	**99.4**	**99.9**	91.1	**99.8**	**99.9**
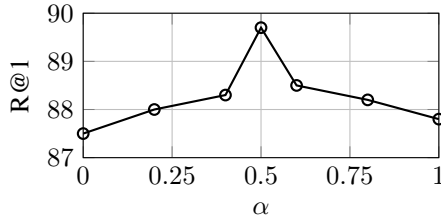
Figure 3: Sensitivity analysis of the loss weight α.

6 RELATED WORKS

6.1 MULTI-MODAL LEARNING WITH UNI-MODAL MODELS

Recent trends in multi-modal learning have increasingly focused on integrating high-performance uni-modal models to achieve multi-modal alignment. 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., 2023b) introduces an additional vision-to-language adaptation module, Q-former, and proposes a two-stage training process to mitigate the challenges associated with learning vision-language alignment. However, both methods require substantial parameters and multi-modal data for cross-modal alignment.

Recent studies indicate a convergence of representations across modalities, providing evidence for the feasibility of developing advanced multi-modal models by connecting high-performance uni-modal models with lightweight parameters and data-efficient techniques (Huh et al., 2024). For instance, LLaVA (Liu et al., 2024) achieved state-of-the-art performance by employing a two-layer multilayer perceptron (MLP). Similar architectures have subsequently proliferated across various domains (Li et al., 2024; Zhang et al., 2024; Lin et al., 2024; 2023). M3-Jepa further advances this concept by interconnecting diverse high-performance uni-modal models through a shared multi-directional MoE. We conducted extensive experiments focused on alignment efficacy, demonstrating the effectiveness of joint training across multiple modalities and directions.

6.2 MULTI-MODAL LEARNING WITH MOE AND AGD

Prior studies have investigated AGD-based multi-modal multi-task alternating training, revealing that the integration of diverse modalities, tasks, and resolutions can yield mutual benefits, thereby effectively enhancing the model's generalization capabilities and cross-domain performance (Akbari et al., 2023; Likhosherstov et al., 2021). we further extend this approach to integrate existing pre-trained high-performance uni-modal models, achieving overall alignment through alternating bidirectional alignment.

MoE-LLaVA (Lin et al., 2024) proposes MoE-Tuning, a strategy for Large Vision-Language Models that creates a sparse model with constant computational cost. While both MoE-LLaVA and M3-Jepa employ sparse MoE to connect high-performance uni-modal large models, M3-Jepa not only scales this approach to large audio-visual-language models but also achieves pairwise bidirectional alignment across modalities.

7 CONCLUSION

In this study, we introduced M3-Jepa, a novel modality-agnostic training strategy and architecture designed for multi-modal learning. M3-Jepa effectively integrates high-performance uni-modal models using bidirectional alignment module, facilitating the alignment of modality pairs in multiple directions and enabling generalization to new tasks and modalities. To validate the modality and task scalability of M3-Jepa, we conducted extensive experiments focused on alignment performance. The experimental results demonstrate that M3-Jepa can easily generalize to new modalities, tasks, and datasets while maintaining the same training strategy and architecture.

REFERENCES

Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *Advances in Neural Information Processing Systems*, 36:79142–79154, 2023.

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

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 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.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

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

Anna Dawid and Yann LeCun. Introduction to latent variable energy-based models: A path towards autonomous machine intelligence, 2023. URL https://arxiv.org/abs/2306.02572.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 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.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022.

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

William Fedus, Barret Zoph, and Noam M. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *ArXiv*, abs/2101.03961, 2021. URL `https://arxiv.org/pdf/2101.03961.pdf`.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Peng Jin, Bo Zhu, Li Yuan, and Shuicheng Yan. Moh: Multi-head attention as mixture-of-head attention. *arXiv preprint arXiv:2410.11842*, 2024.

Jiasheng Ye Yunhua Zhou Dong Zhang Zhigeng Liu Xin Zhang Ruibin Yuan Ge Zhang Linyang Li Hang Yan Jie Fu Tao Gui Tianxiang Sun Yugang Jiang Xipeng Qiu Jun Zhan, Junqi Dai. Anygpt: Unified multimodal llm with discrete sequence modeling, 2024. URL `https://arxiv.org/abs/2402.12226`.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (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`.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL `https://arxiv.org/abs/1602.07332`.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=qrwe7XHTmYb`.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, 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.

Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, 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*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pp. 407–426. Springer, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 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.

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

Noam M. Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ArXiv*, abs/1701.06538, 2017. URL `https://arxiv.org/pdf/1701.06538.pdf`.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. URL `https://arxiv.org/abs/1811.00491`.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 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.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023. URL `https://arxiv.org/abs/2305.11172`.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 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.

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

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts. *ArXiv*, abs/2110.04260, 2021. URL `https://arxiv.org/pdf/2110.04260.pdf`.

A DATASETS

1. **Image-text retrieval datasets:**
   - COCO (Lin et al., 2014): 330,000 images with object annotations and captions, providing a rich resource for multi-label image classification and visual understanding.
   - Flickr30K (Plummer et al., 2015): 30,000 images with descriptive sentences, capturing a diverse range of real-world scenarios.
2. **Zero-shot Audio-text Retrieval**:
   In the zero-shot audio-text task, we trained M3-Jepa on a pretraining dataset and evaluated its performance on unseen datasets. The details of the data 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., 2022): The WavText5K data was sourced from two main website: BigSoundBank and and SoundBible 3 (details can be found in (Deshmukh et al., 2022)). 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.

M3-Jepa was tested on the Clotho dataset and trained on AudioCaps, WavText5K, and Freesound. M3-Jepa was tested on the AudioCaps dataset and trained on Clotho, WavText5K, and Freesound.

3. **Image Classification**:

- ImageNet 1k (Deng et al., 2009): 1,281,167 training images, 50,000 validation images, and 100,000 test images across 1,000 classes, serving as a benchmark for image classification models.

4. **VQA**:

- VQA v2 (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.

B   IMPLEMENTATION DETAILS

B.1   DETAIL OF MULTI-DIRECTIONAL MOE

The Multi-directional MoE assigns k experts by applying a softmax function to the input cross-embedding and representation of input modality. Based on our experimental findings, we opted not to include a load-balancing loss in this framework. We tailored the number of experts and the value of k for each task, optimizing performance. Detailed configurations for different tasks are provided in the supplementary material. The final design of the multi-directional MoE enables the model to perform various tasks based on different modality and embedding cues.

To ensure stable training of sparse-gated MoE under alternating optimization, we experimented with several methods:

1. Random Selection Based on Fixed Probability: We set a probability of 0.5 for both image-to-text and text-to-image tasks, randomly selecting between them.

2. Dynamic Probability Selection: Probabilities were dynamically adjusted based on evaluation results from the validation set. For example, if image-to-text outperformed text-to-image, we increased the probability for text-to-image tasks.

3. Sequential Alternation: Image-to-text and text-to-image tasks were executed in a fixed sequential order. Experimental results showed that the sequential alternation method outperformed the others.

B.2   ZERO-SHOT AUDIO-TEXT RETRIEVAL

In the zero-shot audio-text retrieval task, we fine-tuned the audio and text encoders from Language-Bind (Zhu et al., 2023). M3-Jepa was trained on multiple datasets and tested on unseen data to evaluate its generalization capability. The multi-directional MoE uses 12 experts with a top-4 gating ($k = 4$) setting, a configured hidden size of 2048, and a configured dropout rate of 0.1. It is worth noting that we tested LanguageBind under the same settings, and the results are shown in the Table 2.

B.3   IMAGE CLASSIFICATION

To thoroughly evaluate the quality of the learned representations, we follow the widely adopted linear probing protocol, where a linear classifier is trained on top of the frozen features to perform classification tasks. We evaluate M3-Jepa, Dino V2, and CLIP-ViT under the same settings on ImageNet-1K. Specifically, M3-Jepa aggregates the representations of Dino V2 and CLIP-ViT using a multi-directional MoE mechanism and is evaluated via linear probing. Notably, with Dino V2 and CLIP-ViT kept frozen, M3-Jepa achieves the best performance. The MoE is configured with 8 expert heads, a top-k selection of 4, a dropout rate of 0.1, and a hidden size of 2056.

B.4   VISUAL QUESTION ANSWER

Specifically, we follow previous work (Wang et al., 2022b; 2023; Bao et al., 2021) to conduct finetuning experiments on the VQA v2.0 dataset and formulate the task as a classification problem. The task requires the model to answer natural language questions about input images. M3-Jepa's training strategy adopts a retrieval-based VQA approach, predicting answers from the 3,129 most frequent answer candidates in the training set. For the VQA task, we first perform an alternating alignment task with M3-Jepa on the COCO dataset. We fine-tuned the pretrained multi-directional MoE and applied LoRA (Hu et al., 2021) to both the vision and language models separately. The vision model uses DinoV2-Large, while the language model is based on Qwen-2.5 (Hui et al., 2024). The multi-directional MoE uses 12 experts with a top-4 gating ($k = 4$) setting, a configured hidden size of 2048, and a configured dropout rate of 0.1. Our results on two tasks are lower than BEiT3 and ONE-PEACE (shown in Tabel 4). It is important to highlight that our model was trained exclusively on the VQA v2 training set, whereas other studies have incorporated additional datasets or pretrained on in-domain datasets to boost their model's performance (such as ONE-PEACE trained with Visual Genome(Krishna et al., 2016),and BEiT-3 pretrained on MSCOCO(Lin et al., 2015) and Visual Genome).
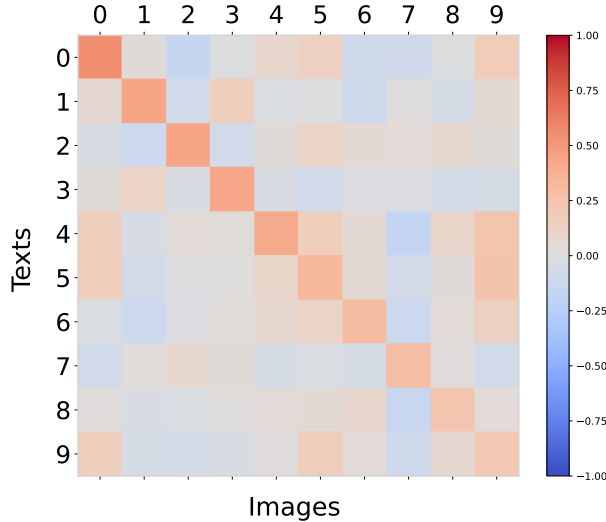
C   VISUALIZATION OF IMAGE-TEXT SIMILARITY

Figure 4: Visualization of the similarity matrix. The up-triangle part indicates text-to-image and the down-triangle part indicates image-to-text. The symmetric line corresponds to the ground truth pairs.

In the experiment, we conducted an image-text retrieval task. The results showed that M3-Jepa performed well in this task, effectively finding matching image-text pairs from a large dataset. To more intuitively demonstrate the model's performance, we visualized the similarity matrix for 10

text-image pairs from the COCO dataset, for the example purpose. Figure 2 shows our method can differentiate the positive and negative pairs well, which ensures the retrieval performance.

D   LIMITATION

In this experiment, we demonstrate the effectiveness of the model across three key tasks: cross-modal retrieval, cross-modal question answering, and image classification. However, as a general multi-modal alignment framework, M3-Jepa still faces the following limitations: 1. Multi-task Scaling: The current experiments focus primarily on cross-modal understanding and reasoning tasks, with limited exploration of generative tasks such as generative visual question answering. As a multi-directional predictor, M3-Jepa has the potential to integrate with various generative frameworks, including diffusion models and large language models (LLMs). Future work will aim to extend its capabilities to multi-modal generation. 2. Modality Expansion: Currently, the framework supports alignment between two modalities. With alternating training, M3-Jepa has the potential to align more than two modalities simultaneously. Further research will explore scaling to more complex multi-modal scenarios. 3. VQA: Although M3-Jepa demonstrates competitive performance in VQA tasks, it does not surpass the current state-of-the-art models. This limitation may stem from the Mixture-of-Experts mechanisms, which may struggle to capture more complex relationships, such as semantic associations and spatial arrangements. While we experimented with local attention mechanisms to aggregate visual semantics, these approaches may not effectively model the global dependencies required for tasks like VQA. Understanding the intricate interplay between questions and image content often requires mechanisms capable of balancing both local and global context. However, the flexibility of the M3-Jepa framework offers opportunities for further enhancements. For instance, incorporating Mixture-of-Heads (MoH) (Jin et al., 2024) attention mechanisms could improve the model's ability to represent cross-modal features and better capture semantic and spatial relationships. Future work will focus on implementing and evaluating such enhancements, including the use of MoH and other advanced connection structures, to further improve the model's performance in VQA and other multimodal tasks.