M3-JEPA:

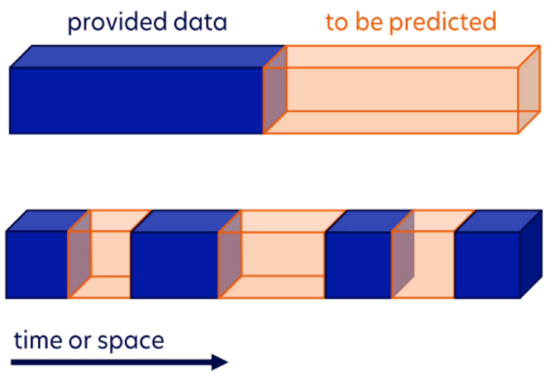
Multimodal Alignment via Multi-gate MoE based on the Joint-Embedding Predictive Architecture

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang,
Huazhen Huang , Qingqing Gu, Yetao Wu, Luo Ji

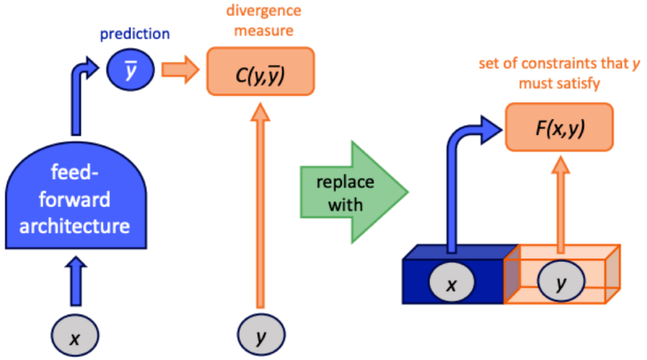
Geely AI Lab
(Luo.Ji1@geely.com)

ICML
International Conference
On Machine Learning

Motivations



- Human perception has a multimodal nature
- Ubiquitous unannotated data => self-supervised learning (**SSL**): *mask some, predict the other*
- Information bias when aligning on the token space => energy-based model (**EBM**)
- Alignment on the latent space
- use an embedding predictor to avoid representation collapse => **JEPA**



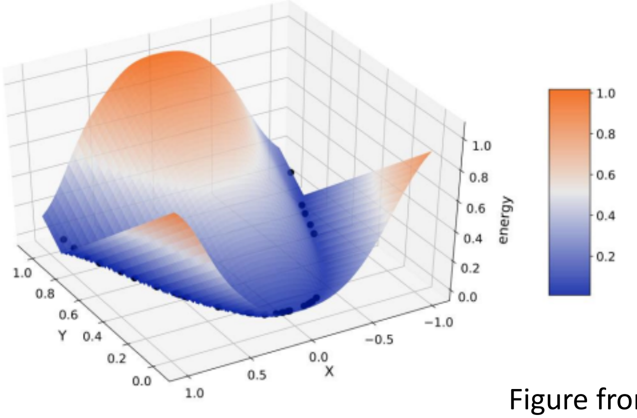
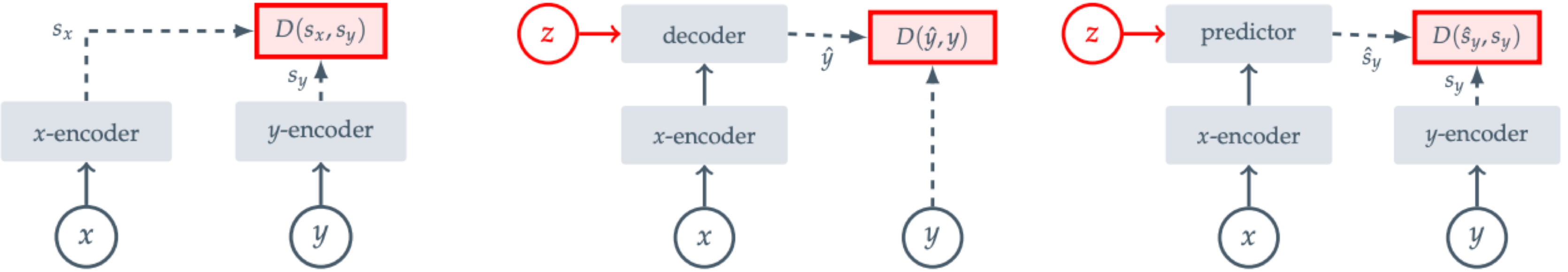


Figure from [1]



(a) **Joint-Embedding Architecture**
can collapse

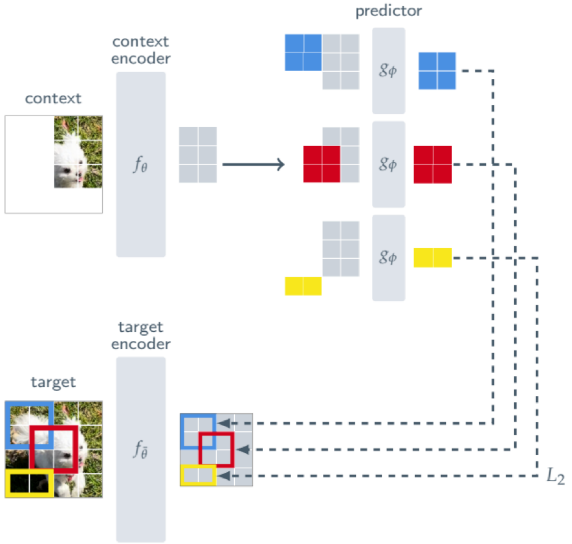
(b) **Generative Architecture**
no collapse
higher biase

(c) **Joint-Embedding Predictive Architecture**
no collapse

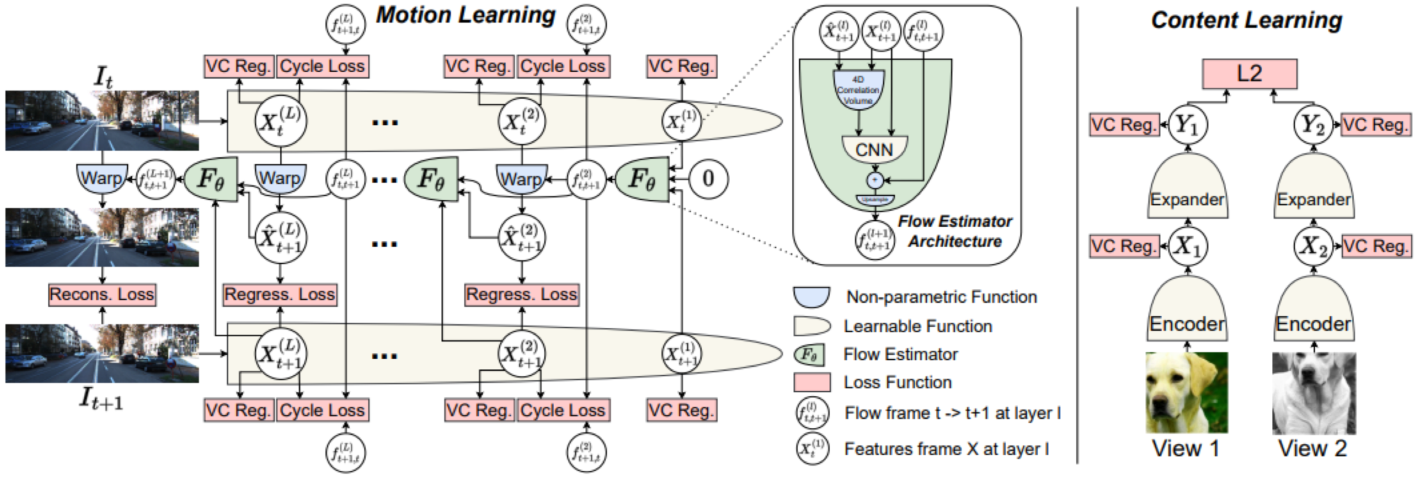
Figure from [2]

[1] Dawid, LeCun. Introduction to latent variable energy-based models: a path toward autonomous machine intelligence. JSTAT 2024
[2] Assran, et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. ICCV 2023

Transfer JEPA from single to multiple modalities



i-JEPA [1]



MC-JEPA [2]

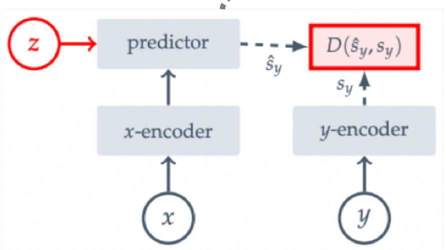
?

MM + JEPA ?

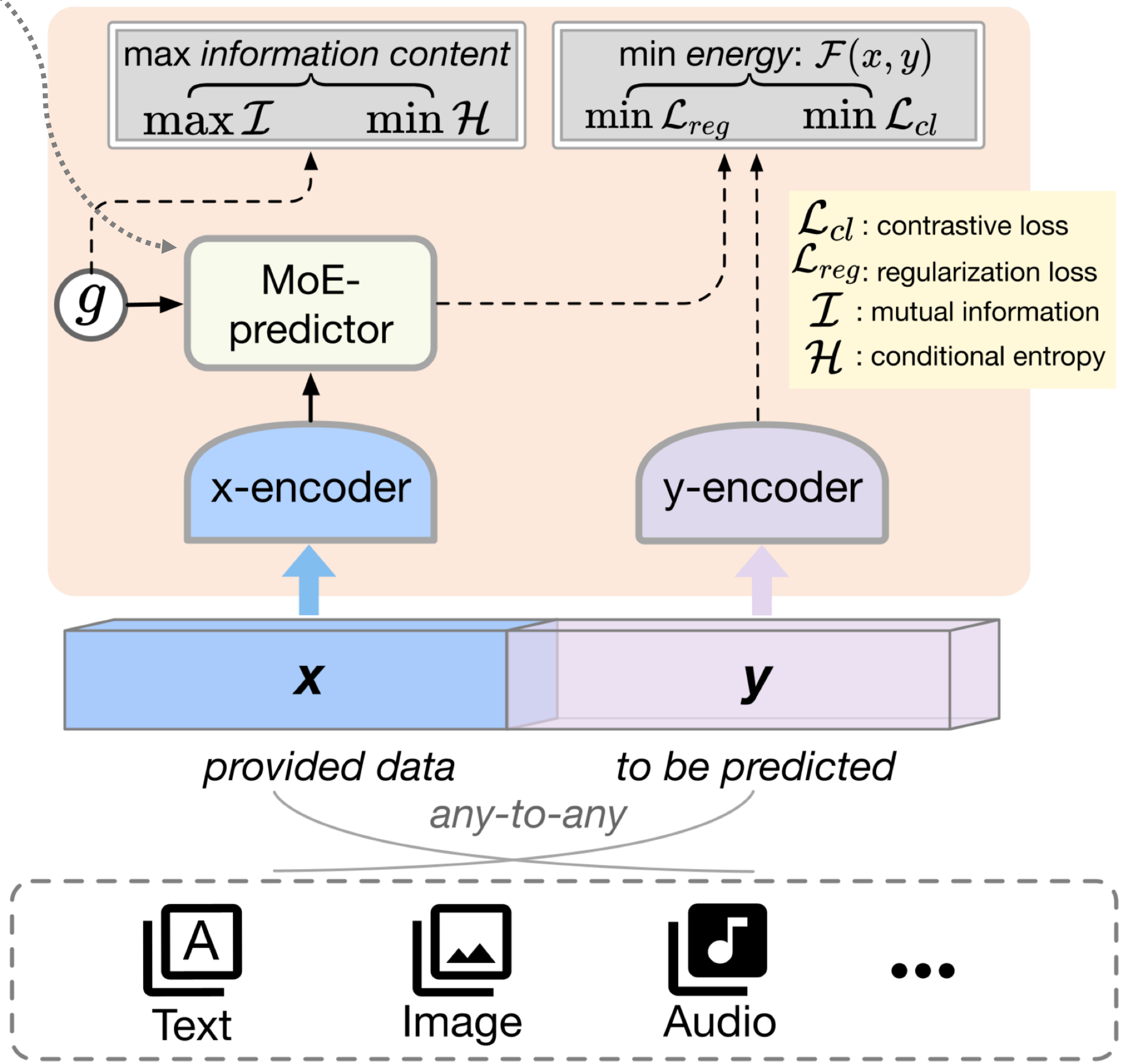
- i-JEPA studies the image classification task
- MC-JEPA expands to motion-content learning
- Here we leverage JEPA on broader multimodal (MM) scenarios
  - Vision, Text, Audio, Others
  - Various masking strategies (different combinations of modalities *w.r.t* input or output)

[1] Assran, et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. ICCV 2023
[2] Bardes, Ponce, LeCun. MC-JEPA: A Joint-Embedding Predictive Architecture for Self-Supervised Learning of Motion and Content Features. Ārxiv 2024

M3-JEPA



- **Multi-modal**
  - Input (x) and output (y) can be any modality or combination of modalities
- **Multi-gate**
  - Gate output for contrastive loss
  - Gate output for regularization loss
- **Mixture-of-expert**
  - Implement the latent predictor by the MoE structure



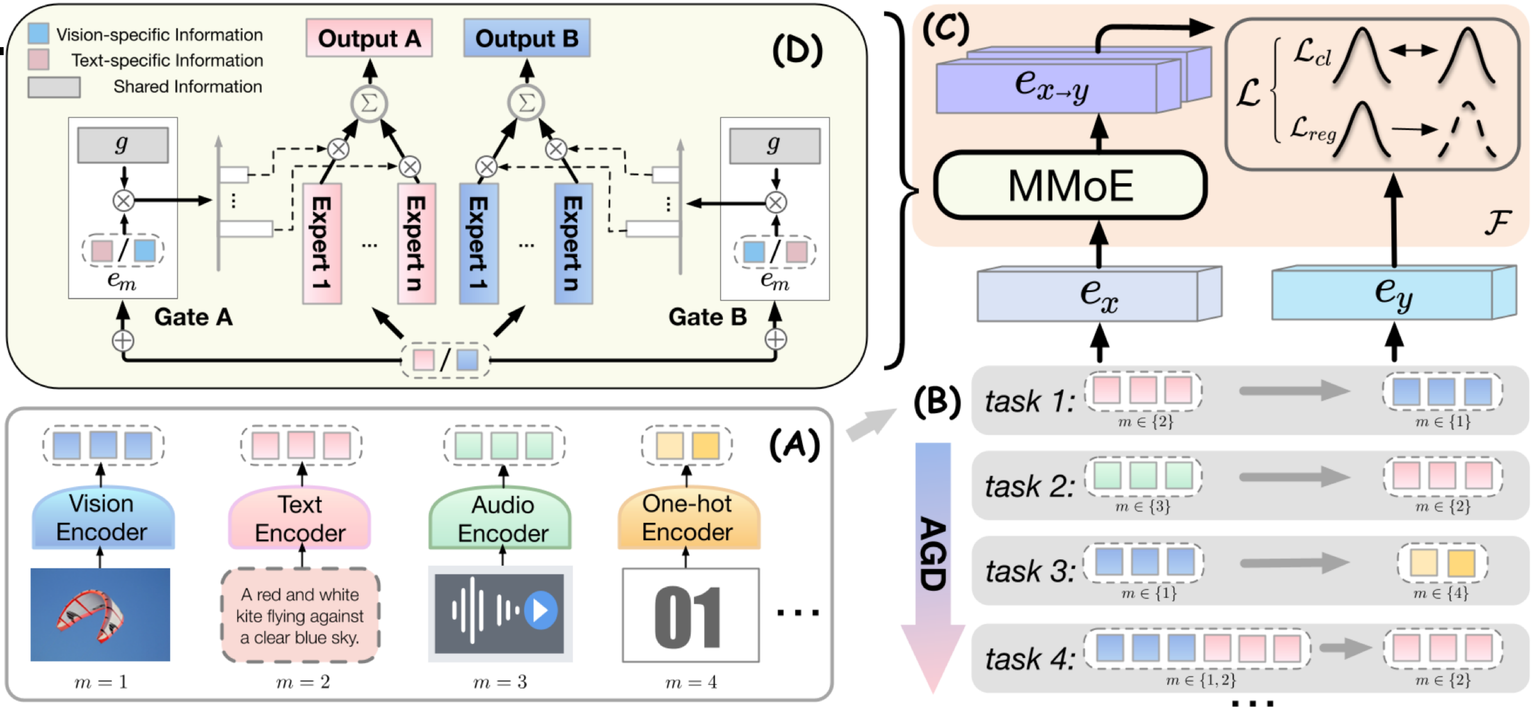
M3-JEPA: detailed architecture

MMoE-like Predictor

- Totally n experts
- Top-k selection
- 2 gates

Losses

- Contrastive loss (cl)
- Regularization loss (reg)
- Total Loss: $L = (1 - \alpha) * L_{cl} + \alpha * L_{reg}$



Modality Encoder

- Vision: DinoV2-Large [1]
- Text: Llama3-8b
- Audio: LanguageBind [2]

Task optimization

- Alternative Gradient Descent (AGD)
- Interleaved optimization: task1, task2, …

[1] Oquab et al. Dinov2: Learning robust visual features without supervision. TMLR 2024
[2] Zhan, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. ICLR 2024

Information-Theoretic Analysis

$\mathcal{L}_{cl}$: contrastive loss
$\mathcal{L}_{reg}$: regularization loss
$\mathcal{I}$: mutual information
$\mathcal{H}$: conditional entropy

$$(1-\alpha)\mathcal{L}_{\text{cl}} + \alpha\mathcal{L}_{\text{reg}} \iff \underbrace{-\mathcal{I}(x;y)}_{\text{redundancy reduction}} + \alpha\underbrace{\left(\mathcal{H}(y|x) + \mathcal{H}(x|y)\right)}_{\text{uncertainty reduction}}$$

Formulation of Losses

cl loss

reg loss

proved by [1]

[1] Lin, Gou, et al. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. CVPR 2021
[2] Jain & Kar. Non-convex optimization for machine learning. Found. Trends Mach. Learn. 2017

Information-Theoretic Analysis

$\mathcal{L}_{cl}$: contrastive loss
$\mathcal{L}_{reg}$: regularization loss
$\mathcal{I}$: mutual information
$\mathcal{H}$: conditional entropy

$$(1-\alpha)\mathcal{L}_{\mathrm{cl}} + \alpha\mathcal{L}_{\mathrm{reg}} \Longleftrightarrow \underbrace{-\mathcal{I}(x;y)}_{\text{redundancy reduction}} + \alpha\underbrace{\left(\mathcal{H}(y|x) + \mathcal{H}(x|y)\right)}_{\text{uncertainty reduction}}$$

Formulation of Losses

cl loss

reg loss

proved by [1]

Derivation of Loss Weight

Connection with Free Energy:

$$F = \underbrace{U}_{\text{Internal Energy}} - T\underbrace{S}_{\text{Entropy}}$$

Critical Temperature $T_c = \alpha = 0.5$

energy/entropy balance occurs

[1] Lin, Gou, et al. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. CVPR 2021
[2] Jain & Kar. Non-convex optimization for machine learning. Found. Trends Mach. Learn. 2017

Information-Theoretic Analysis

$\mathcal{L}_{cl}$: contrastive loss
$\mathcal{L}_{reg}$: regularization loss
$\mathcal{I}$: mutual information
$\mathcal{H}$: conditional entropy

$$(1-\alpha)\mathcal{L}_{\mathrm{cl}} + \alpha\mathcal{L}_{\mathrm{reg}} \Longleftrightarrow \underbrace{-\mathcal{I}(x;y)}_{\text{redundancy reduction}} + \alpha\underbrace{\left(\mathcal{H}(y|x) + \mathcal{H}(x|y)\right)}_{\text{uncertainty reduction}}$$

Formulation of Losses

cl loss

reg loss

proved by [1]

Derivation of Loss Weight

Connection with Free Energy:

$$F = \underbrace{U}_{\text{Internal Energy}} - T\underbrace{S}_{\text{Entropy}}$$

Critical Temperature $T_c = \alpha = 0.5$

energy/entropy balance occurs

From convergence assumption:

$$\mathcal{L} \to \frac{1}{2}\underbrace{(\mathcal{L}(x\to y) + \mathcal{L}(y \to x))}_{\text{consecutive steps}} \to \frac{1}{2}\underbrace{\left(-\mathcal{I}(x;y) + \mathcal{H}(y|x) - \mathcal{I}(y;x) + \mathcal{H}(x|y)\right)}_{\text{converges to the same term}} = -\mathcal{I}(x;y) + \frac{1}{2}\left(\mathcal{H}(y|x) + \mathcal{H}(x|y)\right)$$

$\alpha = 0.5$

convergence theorem of alternating optimization

proved by [2]

[1] Lin, Gou, et al. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. CVPR 2021
[2] Jain & Kar. Non-convex optimization for machine learning. Found. Trends Mach. Learn. 2017

M3-JEPA performs well on cross-modality alignment

Vision-Language Retrieval

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**

- M3-Jepa achieves SOTA performance on Filicker30K and COCO
- M3-Jepa has good computation efficiency (140M trainable parameters)

M3-JEPA adapts to new modalities and generalizes well to different domains

Audio-Language Retrieval

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**

- M3-Jepa has good zero-shot performance on audio-language retrieval
- Generalized on unseen datasets (Clotho and Audiocaps)

Image Classification

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**

- For image classification, we treat the **image label** as a new modality
- Encode by one-hot
- M3-Jepa still surpasses the baseline

M3-JEPA can deal with multiple modalities on input or output

VQA

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>

- For VQA, we simply concatenate the embedding of vision and text as the input encoding
- M3-JEPA performs the second best on VQA and NLVR-2
- Better encoding fusion approaches should be explored

A typical case:

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

M3-Jepa correctly answer the questions of horse, but fail to recognize the staircase accurately

Ablation & sensitivity

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

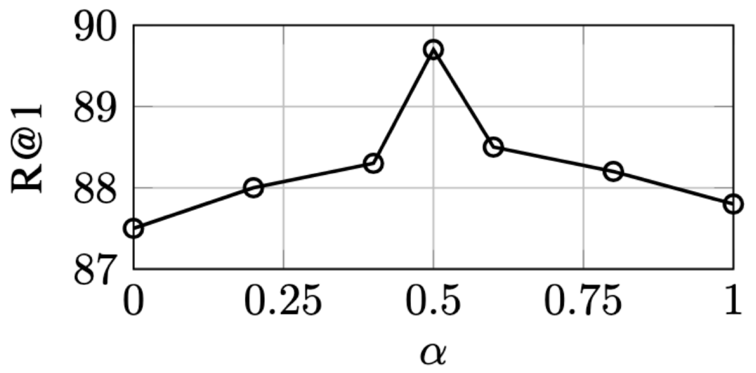
MoE	AGD	Image → Text R@1	Image → Text R@5	Image → Text R@10	Text → Image R@1	Text → Image R@5	Text → Image R@10
×	✓	74.4	86.0	92.2	82.3	89.5	92.6
✓	×	68.2	68.7	81.1	74.2	88.7	92.4
✓	✓	**87.7**	**99.6**	**99.9**	**89.7**	**99.7**	**99.9**

- Both MoE and AGD contribute positively to the performance

Table 6. Ablation of modality encoder finetuning on COCO.

Approach	Image → Text R@1	Image → Text R@5	Image → Text R@10	Text → Image R@1	Text → Image R@5	Text → Image 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**

- Training with full-layer LoRA on encoders further improves the result
- Formally use N=3 layers LoRA considering the efficiency



- Empirical results indicate equal weights of CL and Pred losses are optimal
- *And this is also consistent with the theoretical result!*

Recall:

$L = L_{CL} + \alpha * L_{Pred}$

- Sensitivity of n and k
- We choose n=12 and 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 Image → Text	Flickr30K Text → Image	COCO Image → Text	COCO 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

The efficiency analysis

Training

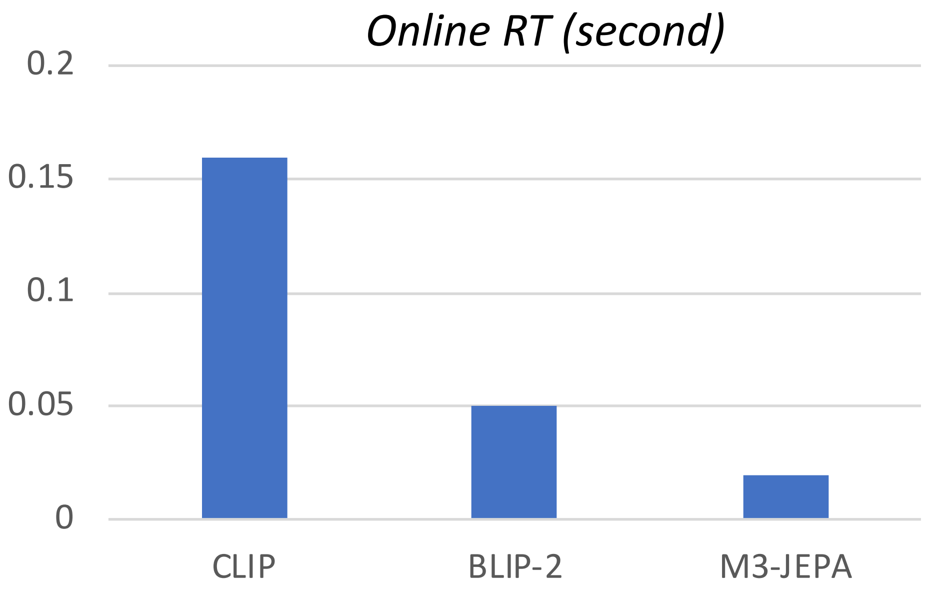
- M3-JEPA has relatively large sizes of modality encoders
- But we only finetune the MoE predictor (maybe also several layers of modality encoders
- M3-JEPA has much fewer # of trainable than baselines

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

Inference

- M3-JEPA can be fast if modality precomputing and online cache are allowed
- For dynamic inputs (e.g. user-provided input), M3-JEPA's latency is dominated by the modality encoder inference



Takeaways & Future Works

Our Paper

Conclusion

- **M3-JEPA** applies JEPA on multi-modal learning by implementing a multi-gate MoE aligner
- **M3-JEPA** achieves SOTA performance with *vision, language & audio* related tasks
- **M3-JEPA** can generalize to broader scenarios

- Scalable: various modalities with a unified architecture
- Generalizable: consistent performance across unseen tasks and domains
- Efficient: small amount of trainable parameters

Future work

- Better modality fusion encoding strategy
- Expand to generative tasks
- Embodied tasks; robotics; world model