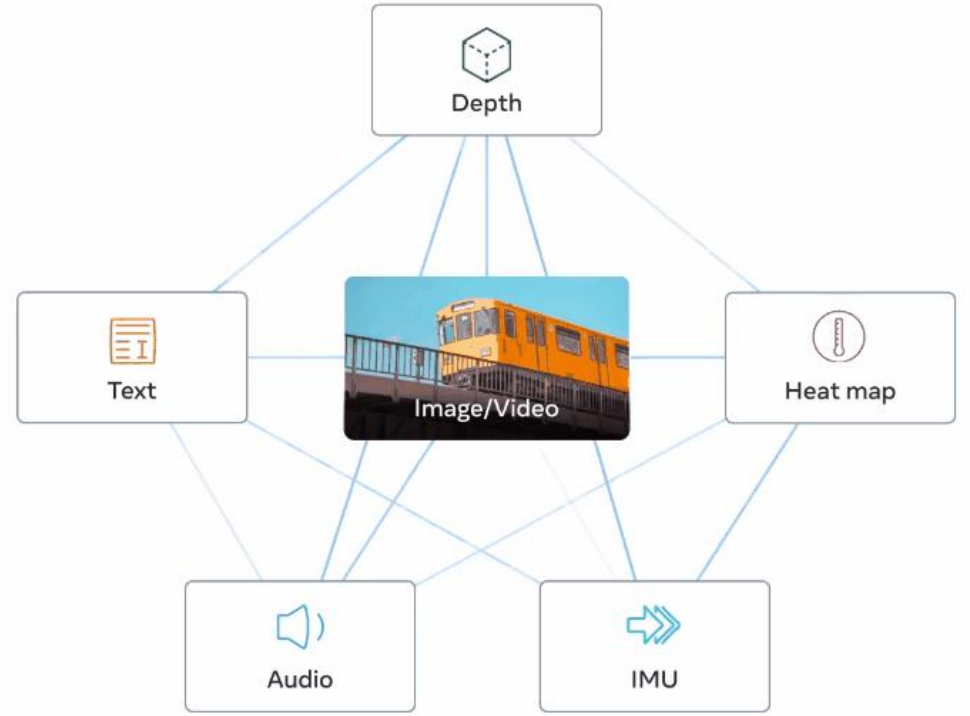
IMAGEBIND: One Embedding Space To Bind Them All

Presenter: Zitao Shuai (ztshuai@umich.edu)

Overview

- **Background**: towards generalizing to various multi-modality tasks
- **Motivation**: Binding all to the most informative modality
- **Method:** Emergent alignment only using image-based pairs
- **Experiment**
- **Quiz**



UNIVERSITY OF MICHIGAN

Background

towards generalizing to various multi-modality tasks

Towards ensemble more modalities

Traditional multi-modal learning:

- Often image-text pairs only
- Feed with humorous data
- Large computation cost

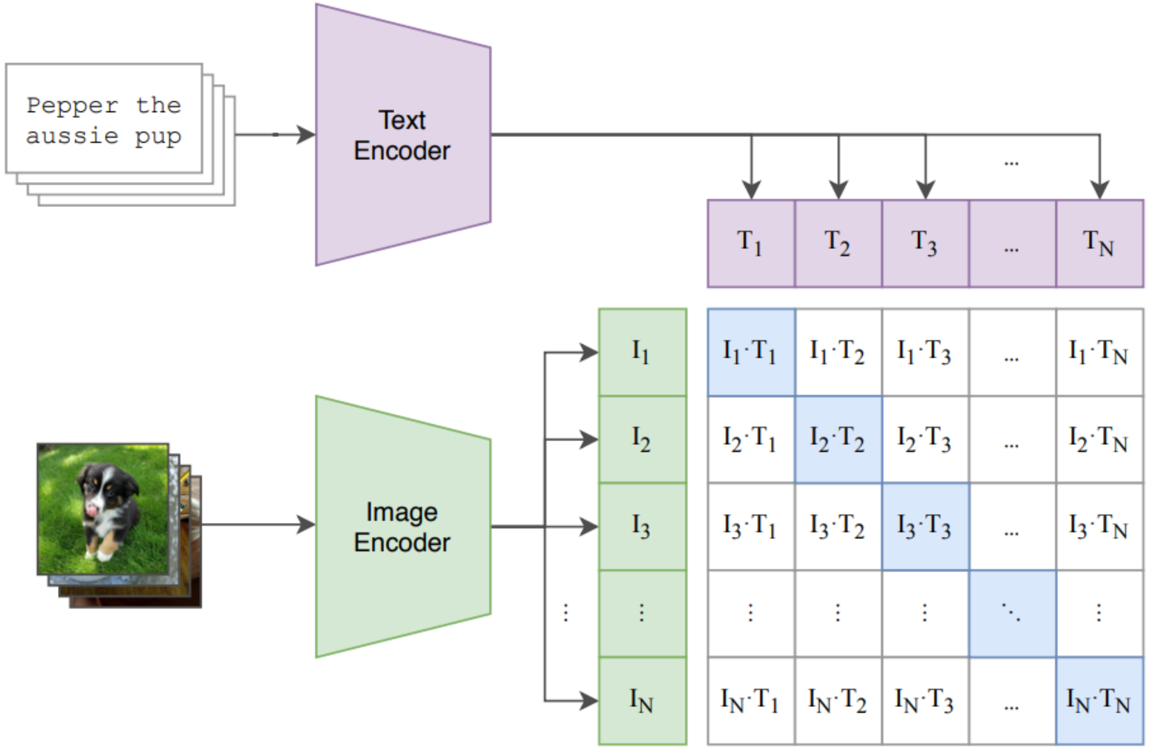
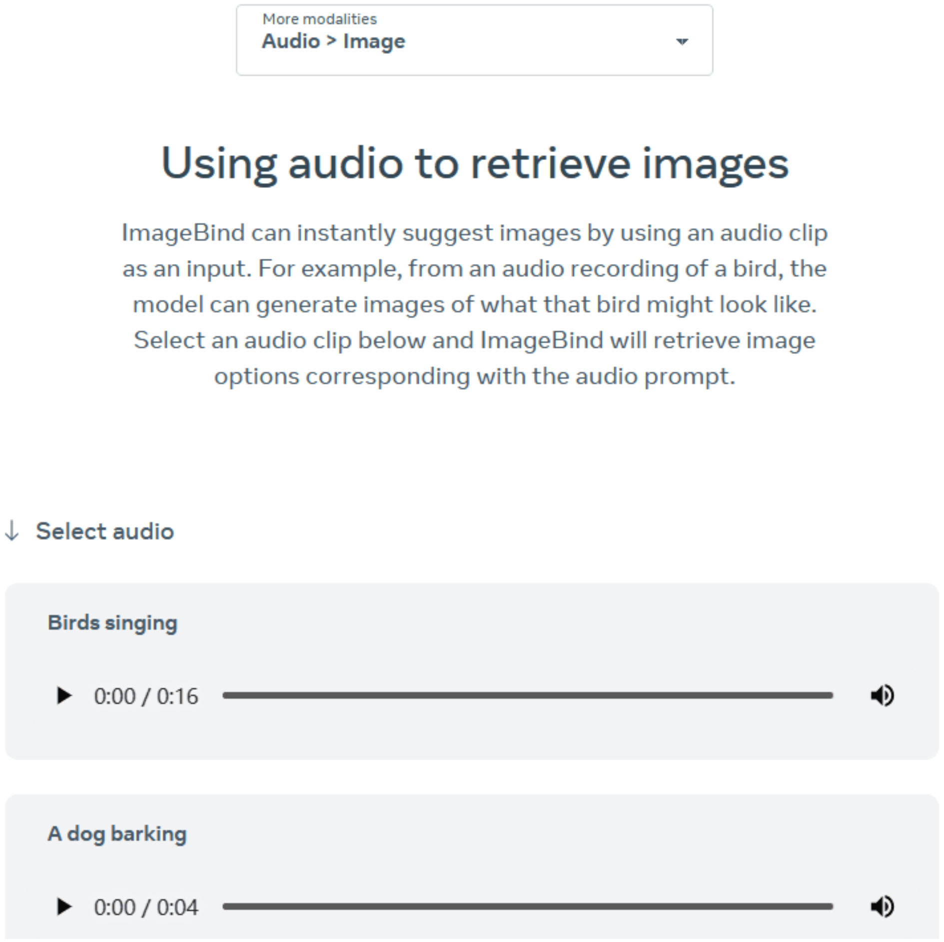


Illustration of the Clip: **The boring graph that appears everywhere in our daily life.**

Towards ensemble more modalities

Ensemble multiple modalities:

Can we learn an MM model performs well on various types of downstream tasks **more than image-text**?



Towards ensemble more modalities

Challenges

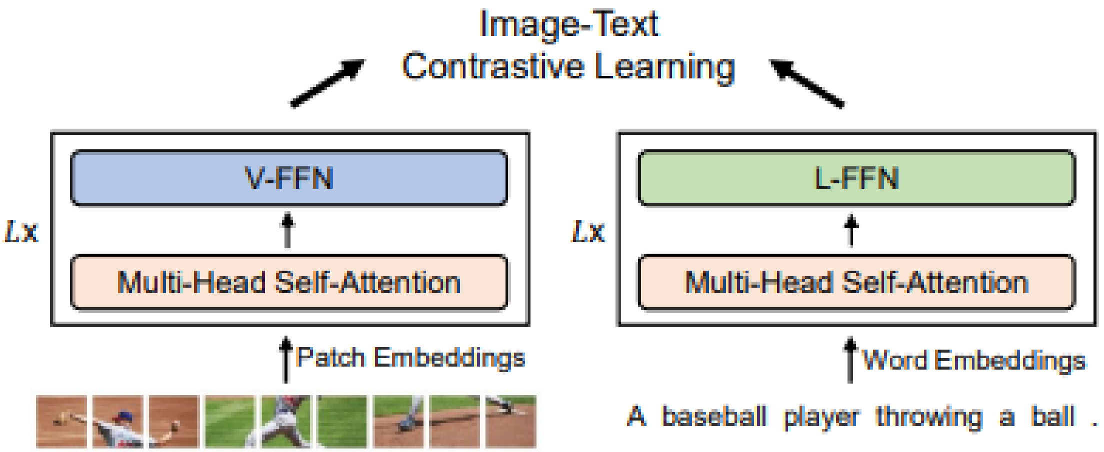
If we align different modalities in traditional ways:

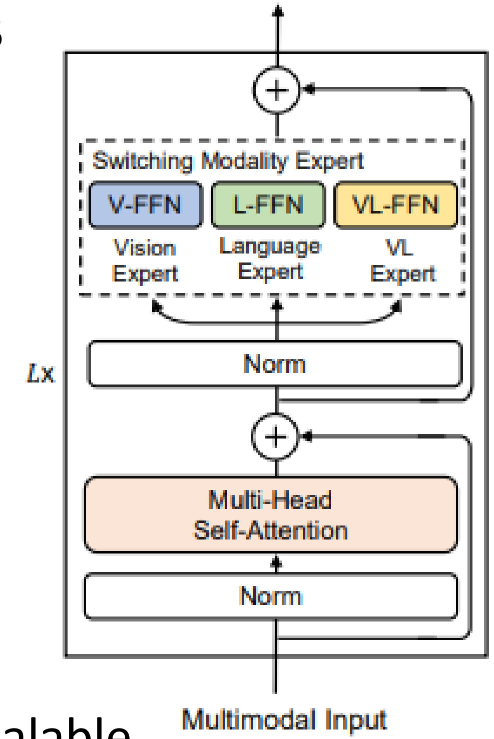
1. Given N modalities, we have $O(N^2)$ multi-modal tasks and each task needs corresponding paired data
2. Some types of paired data is not sufficient

Does anyone like to record the temperature when there is a piece of music?

Towards ensemble more modalities

Solution: parameters shared across modalities





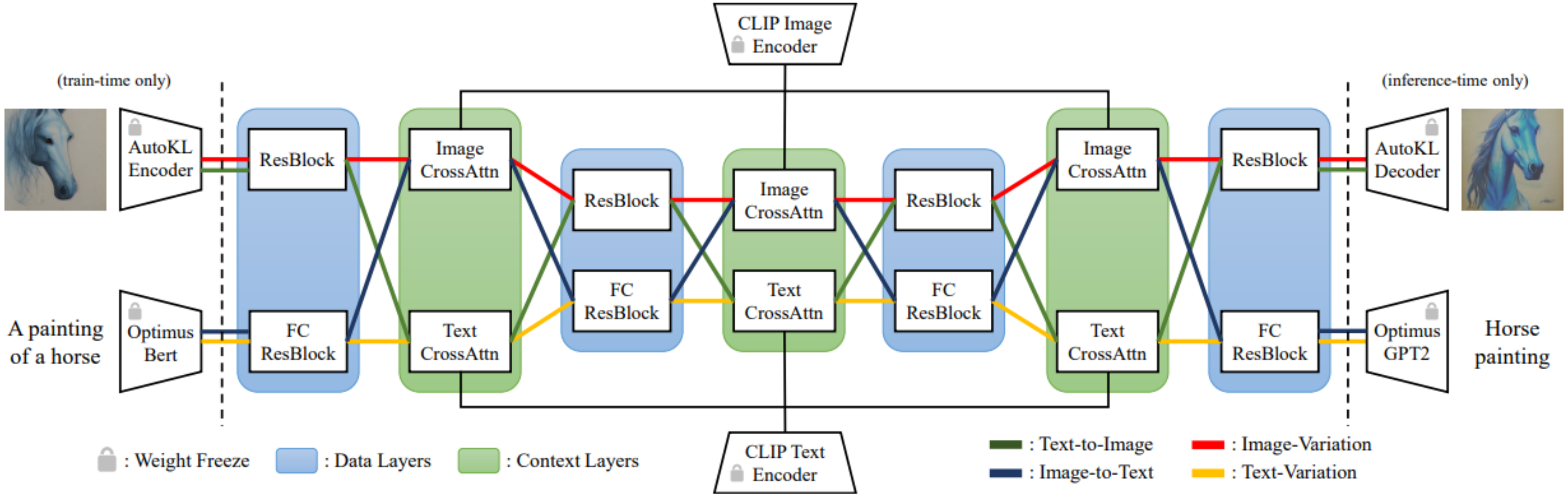
MoME Transformer with Shared Parameters

Problem: Only consider image and text and hard to be scalable

[1] Bao H, Wang W, Dong L, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts[J]. Advances in Neural Information Processing Systems, 2022, 35: 32897-32912.

Towards ensemble more modalities

Solution: parameters shared across tasks / multi-flow network



Problem: only for generation task, not for pretraining

[2] Xu X, Wang Z, Zhang G, et al. Versatile diffusion: Text, images, and variations all in one diffusion model[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 7754-7765.

Motivation

Bind to the most informative modality

Bind to the most informative modality

Something still remains unsolved:

- Parameter sharing is a good way to fusion modalities but we still need $O(N^2)$ contrastive losses.
- Multi-flow network might reduce the size of model but it requires $O(N^2)$ data-flows and feed-forwards
- Current solutions still need paired data or context models pre-trained with the pairs

Bind to the most informative modality

To unify different modalities, we might expect:

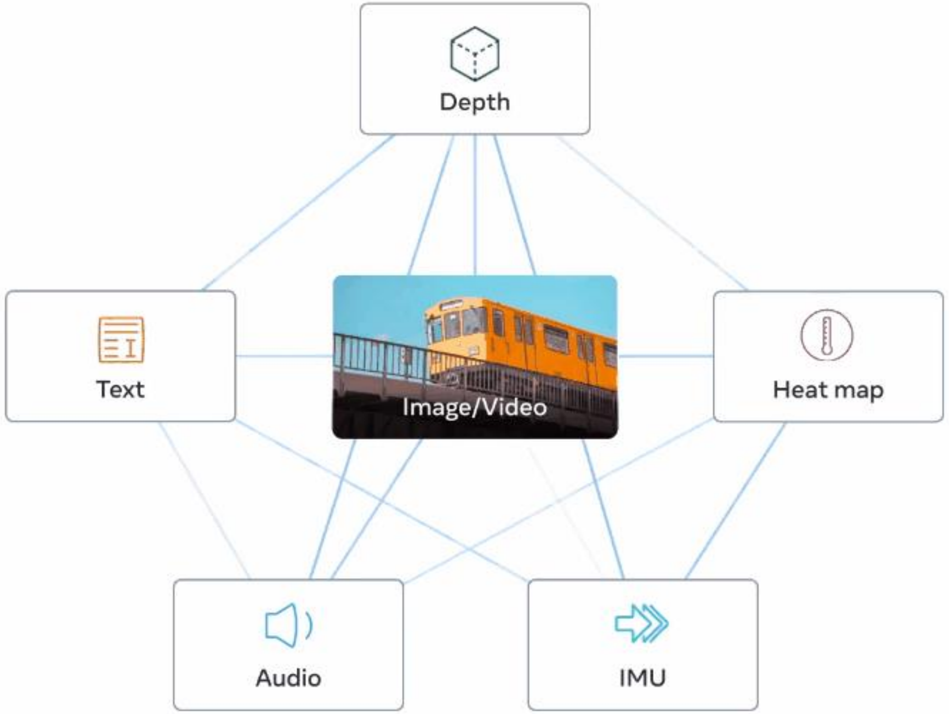
- $\text{Each modality is aligned with other modalities}$
- $O(N)$ Contrastive losses
- Each modality could only appear in one combination of paired modalities

Bind to the most informative modality

Insight: connected graph

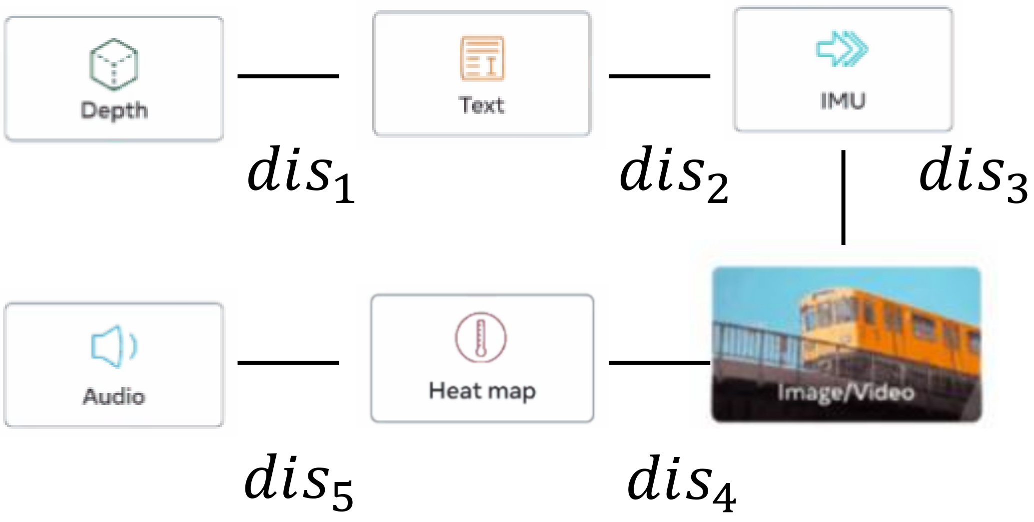
If two modalities are aligned with a loss term, we add an edge between them.

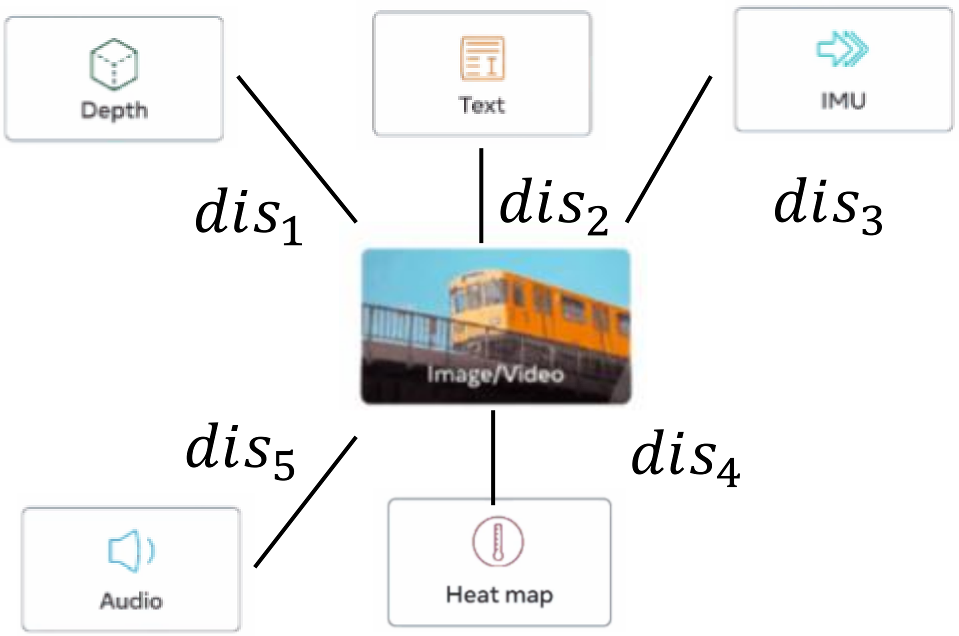
The min number of edges could be N-1.



Bind to the most informative modality

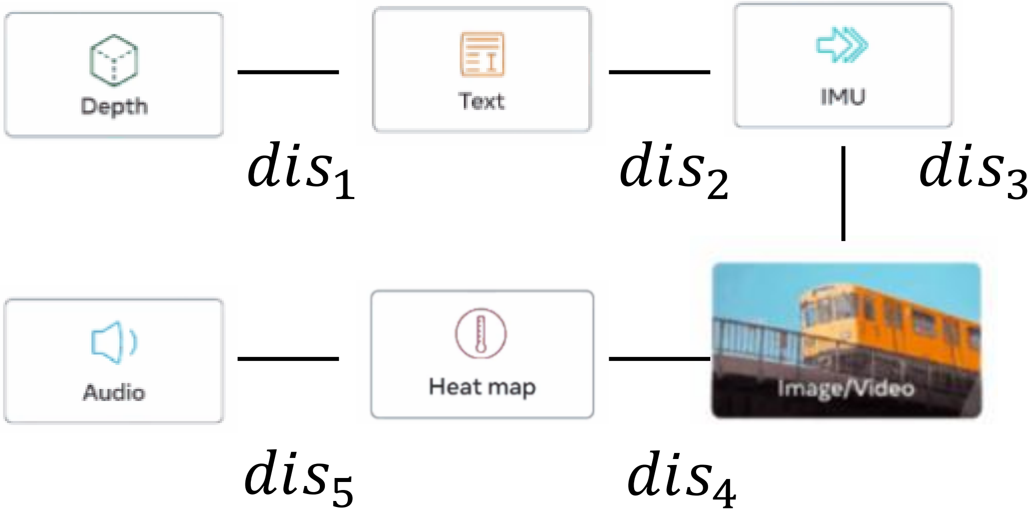
Which one would be better:



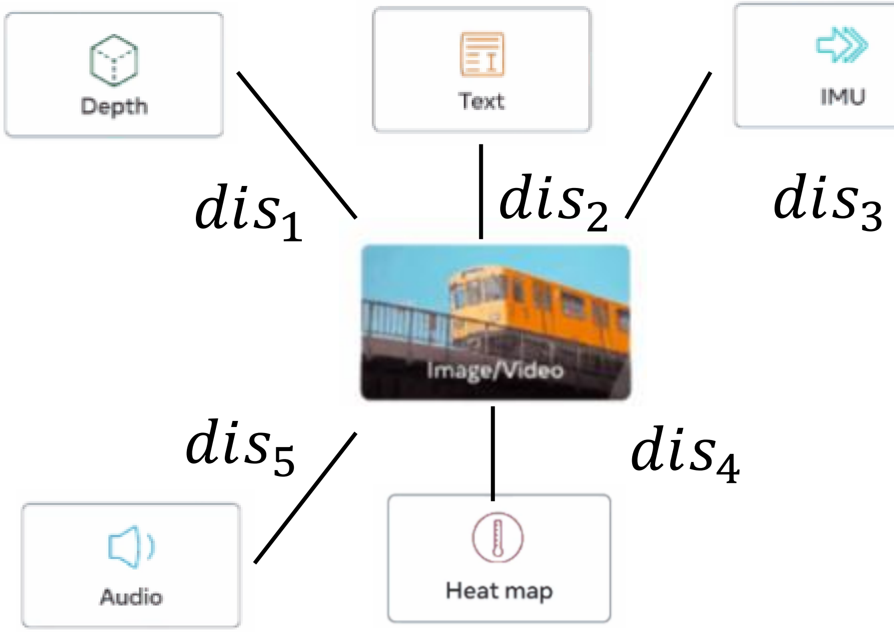


Bind to the most informative modality

Maybe the right one has a tighter upper bound



$$\boldsymbol{dis(D., A.) \leq \sum dis_i}$$



$$\boldsymbol{dis(D., A.) \leq dis_1 + dis_5}$$

Bind to the most informative modality

Remaining question: how to choose the anchor

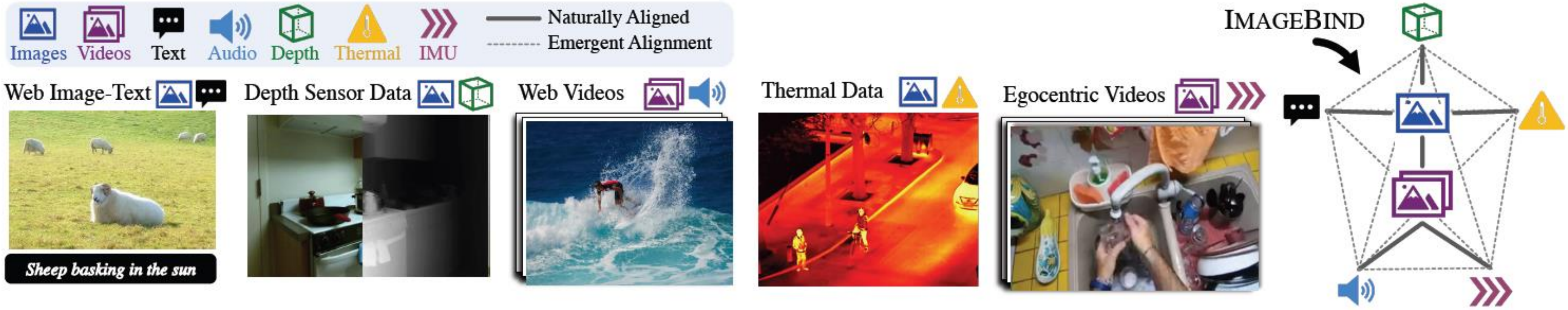
There should exist correlations between the anchor and other modalities:

- Paired data
- Connection of semantics of different modalities

We don't like the case that we have paired data and true relationships as shown above

Bind to the most informative modality

In this paper, image is used as anchor

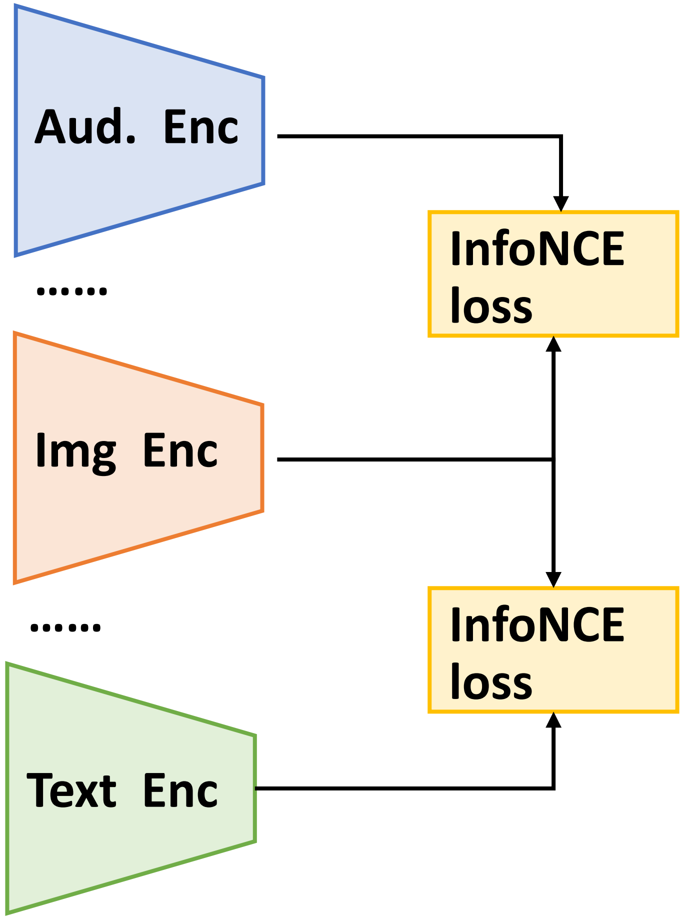


- Each considered modality is closely related to the image(video) modality
- Each modality has paired data with images

Method

Emergent alignment only using image-based pairs

Binding modalities with images



Encoders are not shared.

Given a (img, M) pair, we calculate the following InfoNCE: $L_{I,M} + L_{M,I}$

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

Binding modalities with images

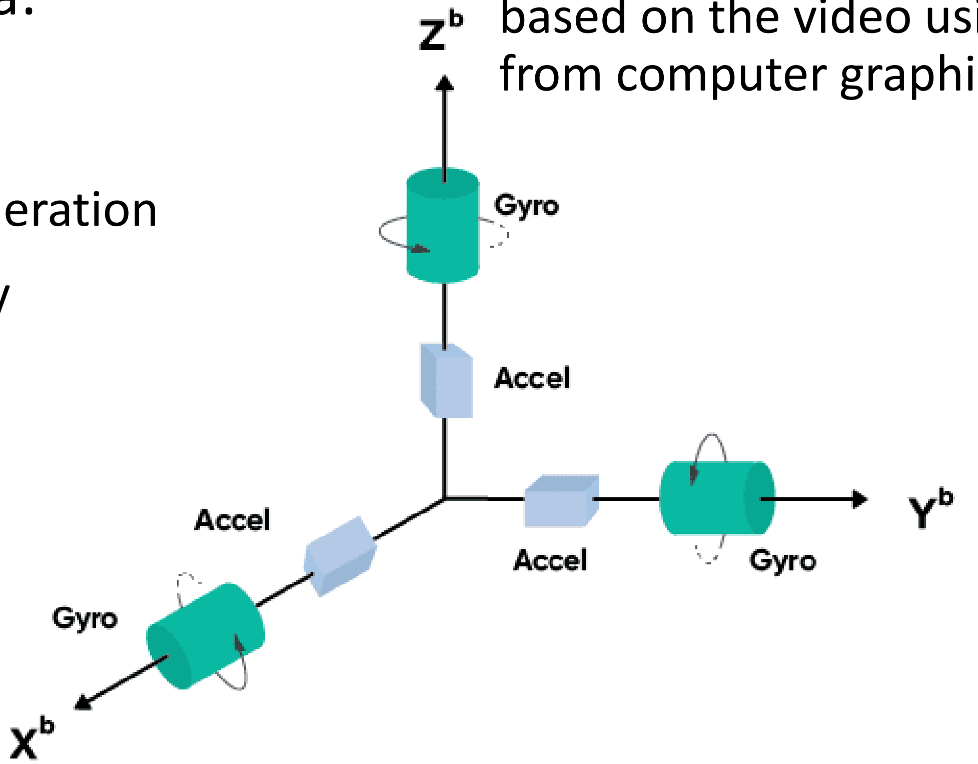
We might be interested in the data:

Inertial Measurement Unit (IMU)

- Accelerometers: Measure linear acceleration
- Gyroscopes: Measure angular velocity

They are 1-D signals, the paper uses 1-D conv and transformers to encode them.

These signals could be calculated based on the video using techniques from computer graphics.



https://www.advancednavigation.com/tech-articles/inertial-measurement-unit-imu-an-introduction/

Binding modalities with images

We might be interested in the data:

Depth data

- Distance information
- Viewpoint

It is related to some 3-D tasks.

It can be viewed as a 1-channel image with similar object semantics to the raw image.

Color

Raw depth

Improved depth

https://rgbd.cs.princeton.edu/

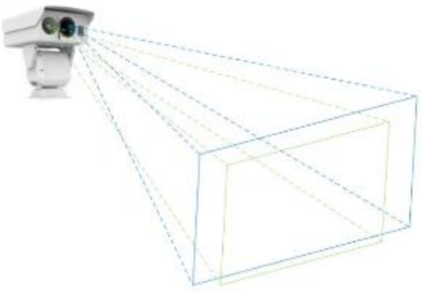
Binding modalities with images

We might be interested in the data:

Thermal data

- Temperature variations of objects or environments.

It can be viewed as a 1-channel image with similar object semantics to the raw image.



(a) dual-spectrum camera

(b) different field of views

(c) images after registration

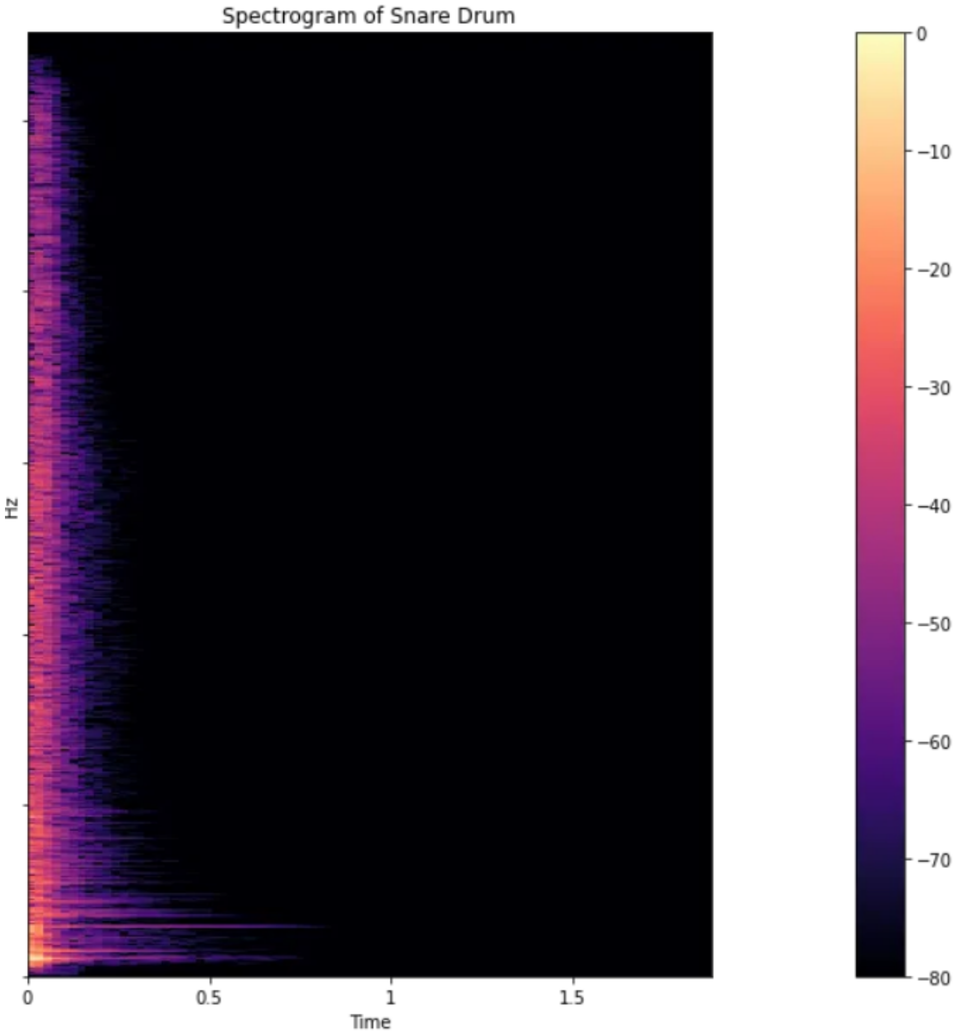
https://bupt-ai-cz.github.io/LLVIP/

Binding modalities with images

We might be interested in the data:

Audio data

- A sequential data, this paper converts a 2 second audio sampled at 16kHz into spectrograms



It can be viewed as a 1-channel image when training.

https://towardsdatascience.com/learning-from-audio-spectrograms-37df29dba98c

Experiment

Experiment

Backbones

- Image-text: Clip-text
- Video-audio: Vit-B
- Image-depth: Vit-S
- Image-thermal: Vit-B
- Image-IMU: 1-D conv + transformer

Dataset	Task	#cls	Metric	#test
Audioset Audio-only (AS-A) [18]	Audio cls.	527	mAP	19048
ESC 5-folds (ESC) [58]	Audio cls.	50	Acc	400
Clotho (Clotho) [16]	Retrieval	-	Recall	1045
AudioCaps (AudioCaps) [36]	Retrieval	-	Recall	796
VGGSound (VGGS) [8]	Audio cls.	309	Acc	14073
SUN Depth-only (SUN-D) [67]	Scene cls.	19	Acc	4660
NYU-v2 Depth-only (NYU-D) [64]	Scene cls.	10	Acc	653
LLVIP (LLVIP) [31]	Person cls.	2	Acc	15809
Ego4D (Ego4D) [22]	Scenario cls.	108	Acc	68865

Experiment

Task 1: zero-shot with text embeddings

	IN1K	P365	K400	MSR-VTT	NYU-D	SUN-D	AS-A	VGGS	ESC	LLVIP	Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9*	25.4*	28.4† [26]	-	68.6† [26]	-	-
Absolute SOTA	91.0 [80]	60.7 [65]	89.9 [78]	57.7 [77]	76.7 [20]	64.9 [20]	49.6 [38]	52.5 [35]	97.0 [9]	-	-

Random < Baseline < ImageBind < Supervised

Conclusion: it transfers the text supervision associated with images to other modalities.

Experiment

Task 2: zero-shot audio retrieval with text

	Emergent	Clotho R@1	Clotho R@10	AudioCaps R@1	AudioCaps R@10	ESC Top-1
Uses audio and text supervision						
AudioCLIP [26]	✗	–	–	–	–	**68.6**
Uses audio and text loss						
AVFIC [50]	✗	3.0	17.5	8.7	37.7	–
No audio and text supervision						
IMAGEBIND	✓	**6.0**	**28.4**	**9.3**	**42.3**	66.9
Supervised						
AVFIC finetuned [50]	✗	8.4	38.6	–	–	–
ARNLQ [52]	✗	12.6	45.4	24.3	72.1	–

Table 3. Emergent zero-shot audio retrieval and classification.

Experiment

Task 3: zero-shot video retrieval with text embeddings

	Modality	Emergent	MSR-VTT R@1	MSR-VTT R@5	MSR-VTT R@10
MIL-NCE [48]	V	✗	8.6	16.9	25.8
SupportSet [56]	V	✗	10.4	22.2	30.0
FIT [5]	V	✗	15.4	33.6	44.1
AVFIC [50]	A+V	✗	19.4	39.5	50.3
IMAGEBIND	A	✓	6.8	18.5	27.2
IMAGEBIND	A+V	✗	36.8	61.8	70.0

Table 4. Zero-shot text based retrieval on MSR-VTT 1K-A.

Experiment

Task 4: Few-shot classification on audio and depth

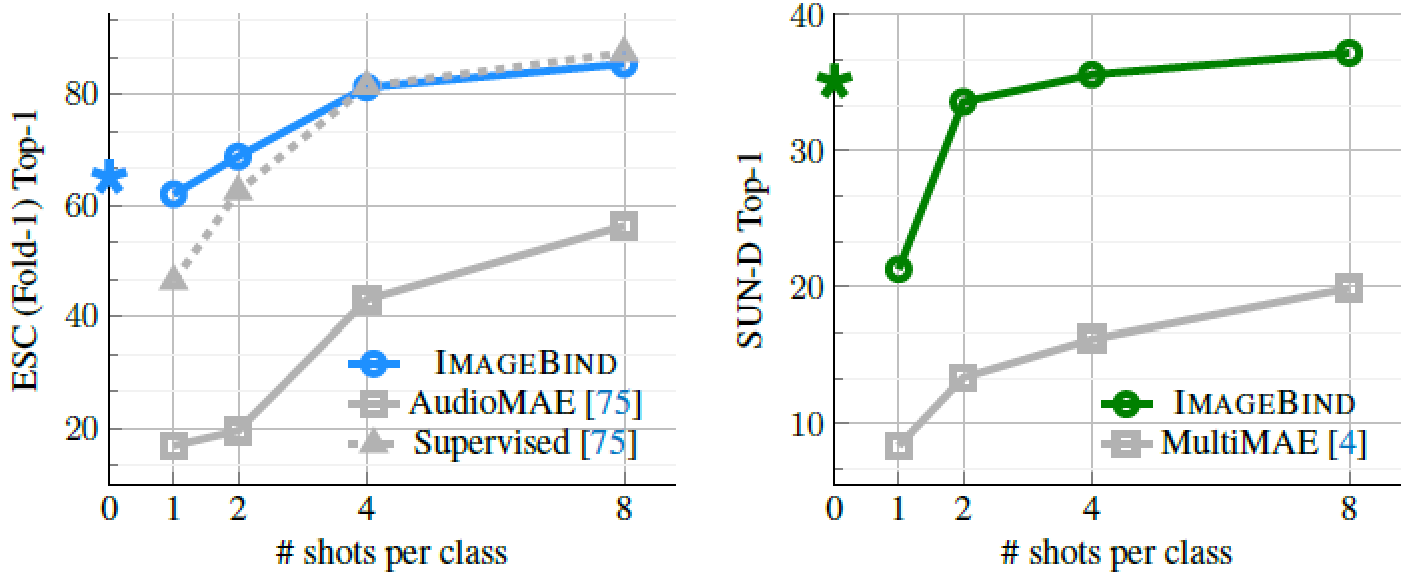


Figure 3. Few-shot classification on audio and depth.

Experiment

Task 4: Few-shot classification on audio and depth

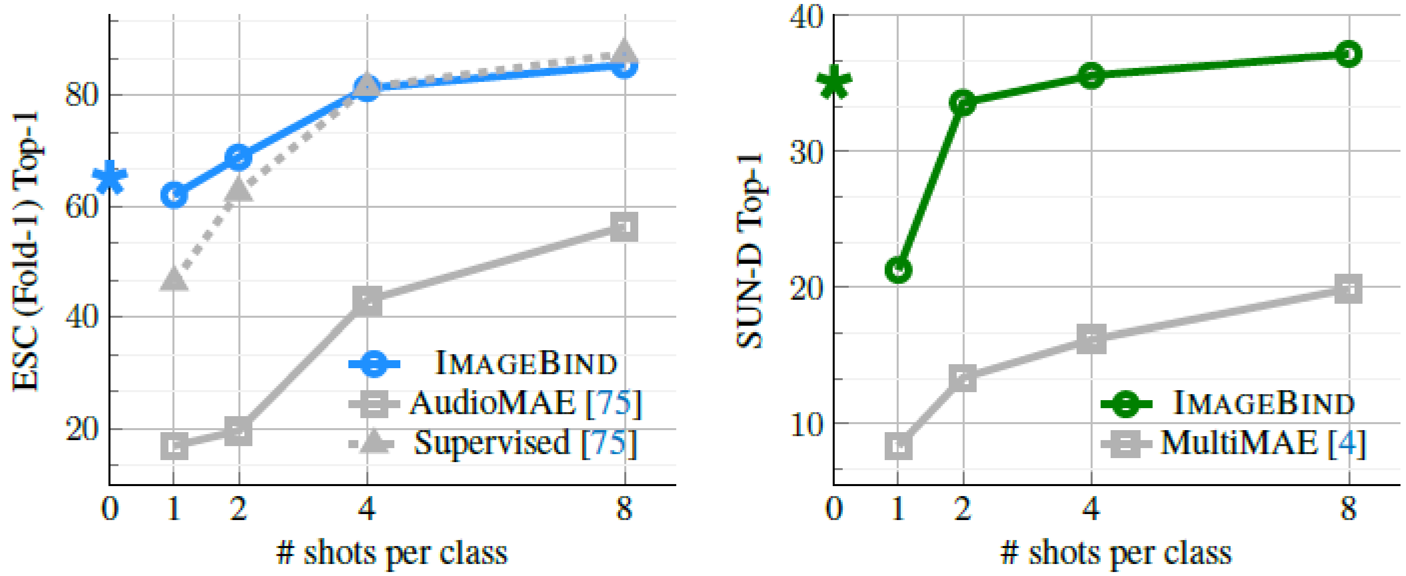
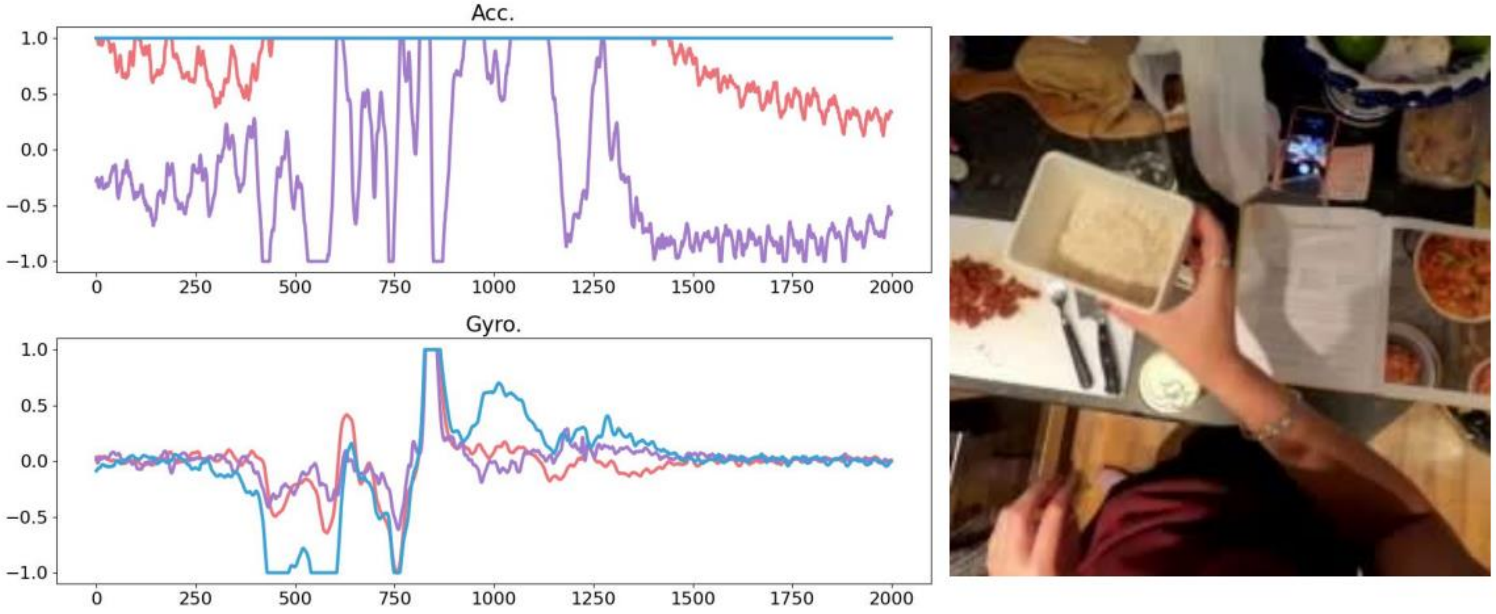


Figure 3. Few-shot classification on audio and depth.

Binding modalities with images

In the appendix of the paper, the proposed model can retrieval images based on IMU signals

Text query: *"Cooking a meal"*



https://arxiv.org/pdf/2305.05665.pdf

Quiz

Opening questions

Q1: Can biomedical data be bonded to the image?

Open questions

Q2: Can modalities bind to others like sensor data?

I think there is an existing work considering binding modalities to the text side.

Zhu B, Lin B, Ning M, et al. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment[J]. arXiv preprint arXiv:2310.01852, 2023.