LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment

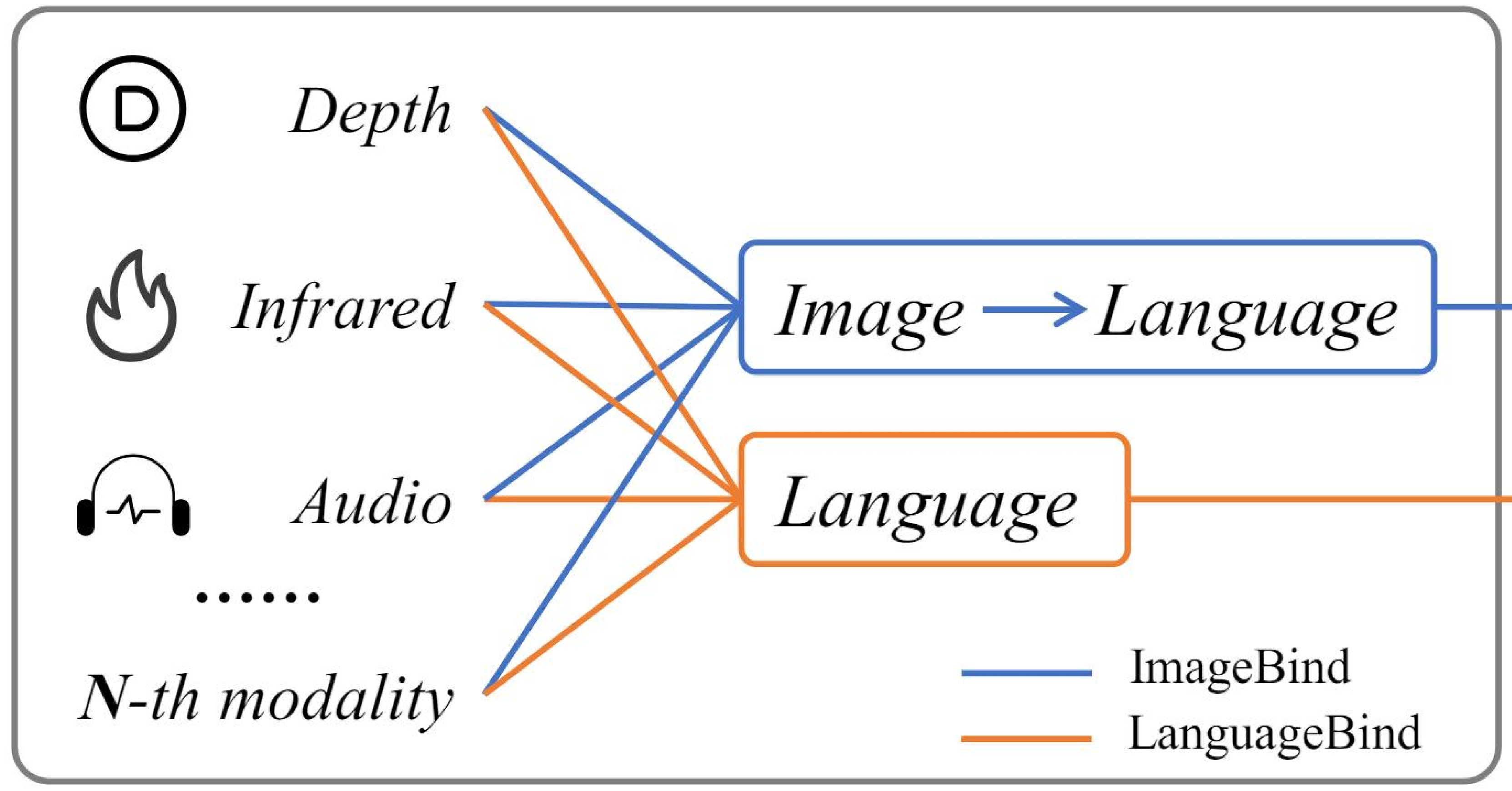
Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, **Li Yuan**

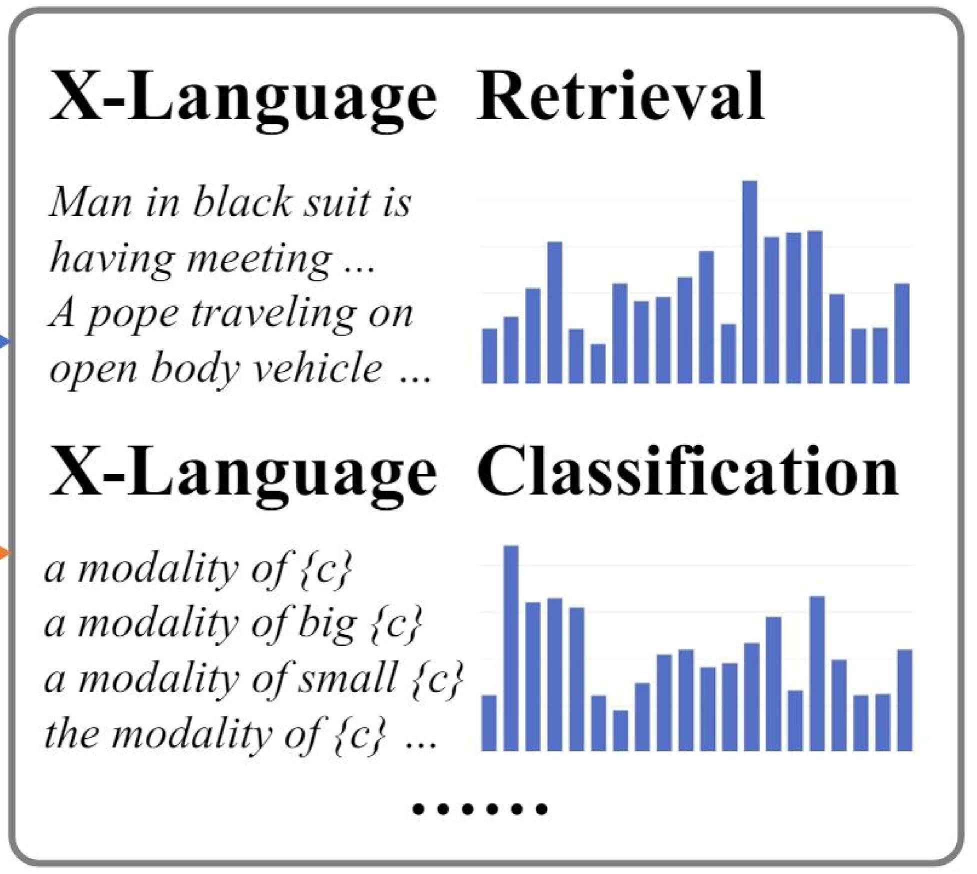
Peking University, Tencent Data Platform, National University of Singapore, Pengcheng Lab

Github

Paper

Core Challenge





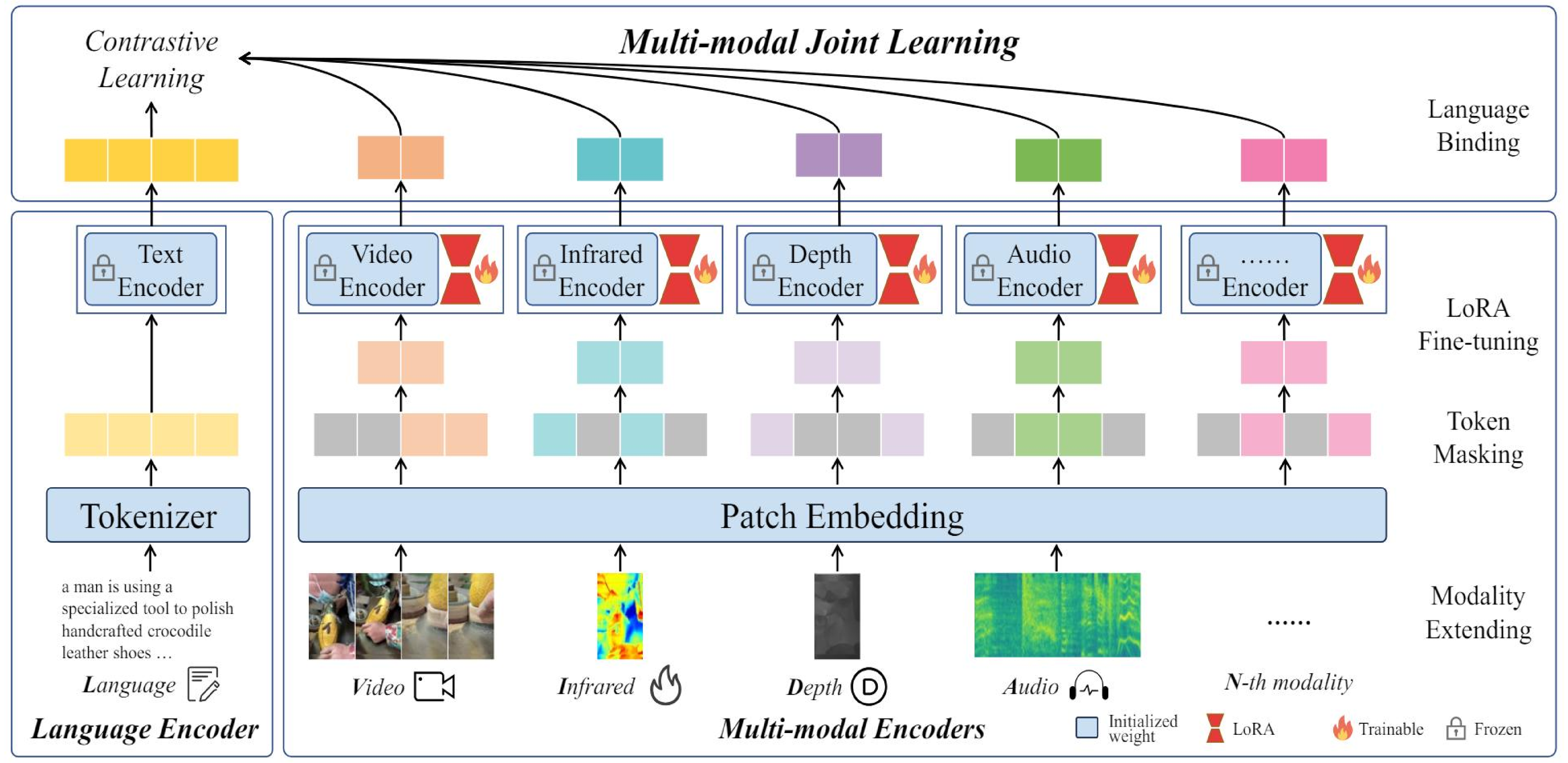
- Current VL pretraining frameworks are often limited to vision and language modalities. ImageBind (Girdhar et al., 2023) introduces an **indirect** alignment method for multi-modal pretraining. It aligns other modalities to images.
- In practical tasks such as retrieval and classification, the **alignment with language** modality is predominantly required for various modalities.

Framework

- By employing **contrastive learning** between language and other modalities, LanguageBind successfully achieved multimodal joint learning, thereby fostering semantic alignment across different modalities.

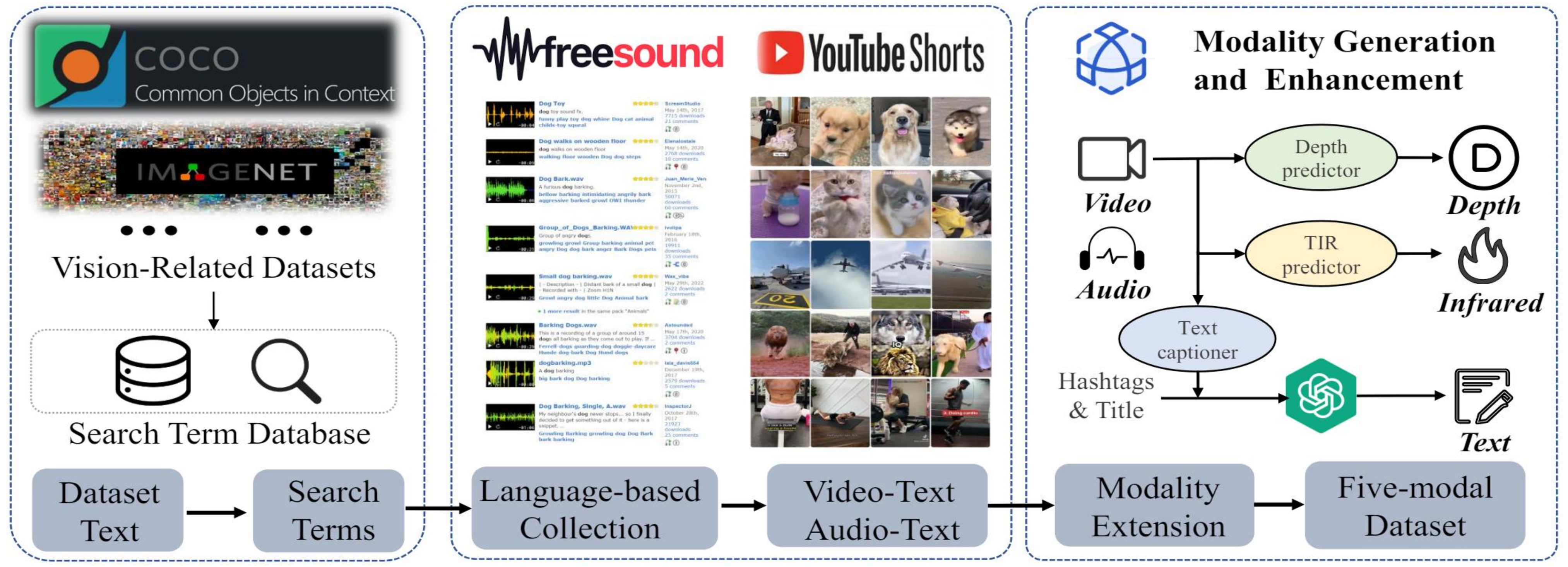
$$L_{M2T} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{K}\exp(x_i^\top y_j/\tau)}$$

$$L_{T2M} = -\frac{1}{K}\sum_{i=1}^{K}\log\frac{\exp(y_i^\top x_i/\tau)}{\sum_{j=1}^{K}\exp(y_i^\top x_j/\tau)}$$



Dataset

- **Search term database collection**
- **Video and audio collection and filtering**
- **Modality generation and enhancemnet**



VIDAL-10M construction. **(a)** Firstly, a search term database is generated by leveraging visually related datasets. **(b)** Subsequently, relevant videos and audios are collected from the internet and undergo a series of filtering processes. **(c)** Lastly, we perform infrared and depth modality generation, as well as multi-view text generation and enhancement.

Experiments

Zero-shot Video-Text retrieval

Method	Dataset	MSR-VTT R@1	MSR-VTT R@5	MSR-VTT R@10	MSVD R@1	MSVD R@5	MSVD R@10	DiDeMo R@1	DiDeMo R@5	DiDeMo R@10	ActivityNet R@1	ActivityNet R@5	ActivityNet R@10
Non-CLIP models													
OmniVL	14M	34.6	58.4	66.6	-	-	-	33.3	58.7	68.5	-	-	-
VideoCoCa	100M	34.3	57.8	67.0	-	-	-	-	-	-	34.5	63.2	76.6
CLIP-H/14													
ImageBind	-	36.8	61.8	70.0	-	-	-	-	-	-	-	-	-
CLIP-L/14													
UMT	5M	33.3	58.1	66.7	44.4	73.3	82.4	34.0	60.4	68.7	31.9	69.2	72.0
TVTSv2	8.5M	38.2	62.4	73.2	-	-	-	34.6	61.9	71.5	-	-	-
InternVideo	12.8M	40.7	-	-	43.4	-	-	31.5	-	-	30.7	-	-
LanguageBind	3M	42.6	65.4	75.5	52.2	79.4	87.3	37.8	63.2	73.4	**35.1**	63.4	76.6
LanguageBind*	3M	42.7	67.1	77.0	53.5	80.5	87.5	38.1	65.0	73.6	36.9	65.1	77.2
LanguageBind*	10M	**42.8**	67.5	76.0	**54.1**	**81.1**	**88.1**	39.7	65.5	73.8	38.4	66.6	77.9
CLIP-H/14													
LanguageBind*†	**10M**	**44.8**	**70.0**	**78.7**	53.9	80.4	87.8	**39.9**	**66.1**	**74.6**	**41.0**	**68.4**	**80.0**

Zero-shot Audio-Language retrieval

Method	Clotho R@1	Clotho R@10	Audiocaps R@1	Audiocaps R@10
AVFIC	3.0	17.5	8.7	37.7
ImageBind	6.0	28.4	9.3	42.3
VALOR	8.4	-	-	-
LanguageBind	**12.1**	**44.0**	**12.2**	**53.2**
LanguageBind*	**16.7**	**52.0**	**19.7**	**67.6**

Zero-shot X-Language classification

Method	Size	Video K400	Video K600	Infrared LLVIP	Infrared FLIR V1	Infrared FLIR V2	Depth NYU-D	Audio AS-A	Audio ESC-50	Audio VGGS
ImageBind	Huge	50.0	-	63.4	-	-	54.0	17.6	66.9	27.8
OpenCLIP	Large	60.7	59.0	82.2	81.2	42.6	45.4	-	-	-
LanguageBind	**Large**	**64.0**	**61.9**	**87.2**	**82.9**	**48.0**	**65.1**	**27.7**	**91.8**	**28.9**
LanguageBind*	**Large**	-	-	-	-	-	-	**30.0**	**94.0**	**38.6**

Emergent zero-shot retrieval

Dataset	Method	Task	Top-1
MSR	ImageBind	V→T	36.1*
		A+V→T	36.8 (+0.7)
	Ours	V→T	41.4
		A+V→T	42.0 (+0.6)
NYU	ImageBind	D→T	54.0
	Ours	D→T	65.1
		RGB→T	76.0
		D+RGB→T	77.4 (+1.4)
LLVIP	Ours	RGB†→T	62.4
		I+RGB†→T	79.3 (+16.9)

Dataset	Method	Task	Emergent	R@1
AVE†	Ours	RGB→A	✔	10.6
	ImageBind		✗	36.9
VGGS†	Ours	RGB→A	✔	10.0
	ImageBind		✗	28.7
LLVIP†	Ours	RGB→I	✔	7.5
		RGB+T→I	✗	9.1
		I→RGB	✔	9.3
		D+I→RGB	✗	10.6
NYU	Ours	RGB→D	✔	17.9
		RGB+T→D	✗	18.3
		D→RGB	✔	24.5
		D+T→RGB	✗	25.7