OmniVL: One Foundation Model for Image-Language and Video-Language Tasks

Junke Wang[1,2], Dongdong Chen[3], Zuxuan Wu[1,2†], Chong Luo[4], Luowei Zhou[3],
Yucheng Zhao[4], Yujia Xie[3], Ce Liu[3], Yu-Gang Jiang[1,2†], Lu Yuan[3]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Microsoft Cloud + AI, [4]Microsoft Research Asia.
(† denotes corresponding authors)

OmniVL unifies the foundation models in three dimensions:

- ***Modality***: spatial-temporal transformer-based visual encoder to support both image and video inputs.
- ***Functionality***: encoder-decoder structure with two decoders for cross-modal alignment and text generation, respectively,
- ***Pretraining Data:*** joint visual-label-text space to unify labelled data and web-crawled data for vision-language pretraining.

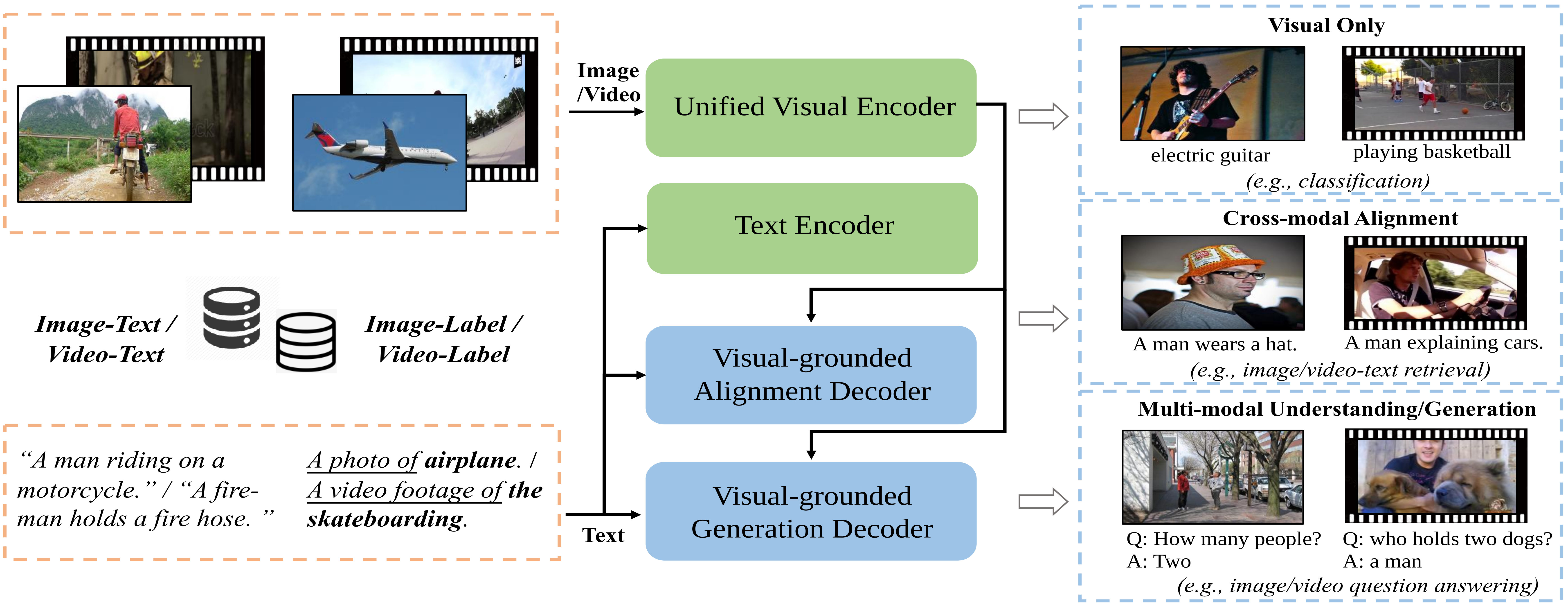
Image /Video
Unified Visual Encoder
Text Encoder
Visual-grounded Alignment Decoder
Visual-grounded Generation Decoder
Text
Image-Text / Video-Text
Image-Label / Video-Label
"A man riding on a motorcycle." / "A fire-man holds a fire hose. "
A photo of airplane. / A video footage of the skateboarding.
Visual Only
electric guitar
playing basketball
(e.g., classification)
Cross-modal Alignment
A man wears a hat.
A man explaining cars.
(e.g., image/video-text retrieval)
Multi-modal Understanding/Generation
Q: How many people?
A: Two
Q: who holds two dogs?
A: a man
(e.g., image/video question answering)

Pretraining Corpus Unification **Modality Unification** **Functionality Unification**

Paradigms: first perform image-language pretraining and then jointly pretrain with video-language data. Two potential benefits: 1) applying the image data to learn spatial representation first is more efficient. 2) The decoupled pattern makes the multimodal representation learning more effective to make image-language and video-language benefit each other.

Method	# Img-Text Pairs	COCO (5K test set)						Flickr30K (1K test set)					
		TR			IR			TR			IR		
VirTex [46]	-	-	-	-	38.1	62.8	-	-	-	-	35.1	64.6	-
UNITER [13]	4M	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
OSCAR [40]	4M	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
UNIMO [39]	4M	-	-	-	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1
VLMO [60]	4M	74.8	93.1	96.9	57.2	82.6	**89.8**	92.3	99.4	99.9	79.3	95.7	97.8
OmniVL	4M*	**76.8**	**93.6**	**97.3**	**58.5**	**82.6**	89.5	**94.9**	**99.6**	**99.9**	**83.4**	**97.0**	**98.6**
FLAVA [53]	70M	61.5	82.1	89.6	50.1	74.4	83.2	85.4	95.7	98.3	73.2	92.7	95.5
METER [21]	404M	76.2	93.2	96.8	57.1	82.7	90.1	94.3	99.6	99.9	82.2	96.3	98.4
ALIGN [29]	1.8B	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6
ALBEF [36]	14M	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9
BLIP [35]	14M	80.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8
Florence [69]	900M	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	**98.1**	-
OmniVL	14M*	**82.1**	**95.9**	**98.1**	**64.8**	**86.1**	**91.6**	**97.3**	**99.9**	**100.0**	**87.9**	97.8	**99.1**

Method	Text-to-Video Retrieval						Zero-shot Retrieval					
	MSRVTT			DiDeMo			MSRVTT			DiDeMo		
ClipBERT [33]	22.0	46.8	59.9	20.4	48.0	60.8	-	-	-	-	-	-
TT-CE+ [14]	29.6	61.6	74.2	21.6	48.6	62.9	-	-	-	-	-	-
VideoCLIP [64]	30.9	55.4	66.8	-	-	-	10.4	22.2	30.0	16.6	46.9	-
FiT [6]	32.5	61.5	71.2	31.0	59.8	72.4	18.7	39.5	51.6	21.1	46.0	56.2
TT-CE+ (+QB-NORM) [9]	33.3	63.7	76.3	24.2	50.8	64.4	-	-	-	-	-	-
ALPRO [34]	33.9	60.7	73.2	35.9	67.5	78.8	24.1	44.7	55.4	23.8	47.3	57.9
VIOLET [22]	34.5	63.0	73.4	32.6	62.8	74.7	25.9	49.5	59.7	23.5	49.8	59.8
OmniVL	**47.8**	**74.2**	**83.8**	**52.4**	**79.5**	**85.4**	**42.0**	**63.0**	**73.0**	**40.6**	**64.6**	**74.3**

Method	# Img-Text Pairs	NoCaps								COCO Caption Karpathy test	
		in-domain		near-domain		out-domain		overall			
		C	S	C	S	C	S	C	S	B@4	C
Enc-Dec [11]	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9
VinVL [71]	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
LEMON [28]	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
BLIP [35]	14M	**111.3**	**15.1**	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
SIMVLM [61]	1.8B	-	-	-	-	-	-	94.8	13.1	39.0	134.8
OFA$_{14M}$ [58]	14M	-	-	-	-	-	-	-	-	38.7	130.5
OFA [58]	21.4M	-	-	-	-	-	-	-	-	**41.0**	**138.2**
OmniVL	14M*	104.6	15.0	**108.3**	**14.9**	**106.3**	**14.2**	**107.5**	**14.7**	39.8	133.9

Method	B@3	B@4	METEOR	ROUGE-L	CIDEr
Bi-LSTM [73]	-	0.87	8.15	-	-
EMT [74]	-	4.38	11.55	27.44	0.38
VideoBERT [55]	6.80	4.04	11.01	27.50	0.49
ActBERT [75]	8.66	5.41	13.30	30.56	0.65
AT [27]	-	8.55	16.93	35.54	1.06
UniVL [45]	16.46	11.17	17.57	40.09	1.27
OmniVL	**12.87**	**8.72**	**14.83**	**36.09**	**1.16**

Method	# Img-Text Pairs	test-dev	test-std
FLAVA [53]	68M	72.80	-
OSCAR [40]	4M	73.16	73.44
ALBEF [36]	14M	75.84	76.04
BLIP [35]	14M	77.54	77.62
METER [21]	404M	77.68	77.64
SimVLM [61]	1.8B	77.87	78.14
OFA [58]	21.4M	78.00	78.10
OmniVL	14M*	**78.33**	**78.35**

Method	MSRVTT	MSVD
ClipBERT [33]	37.4	-
JustAsk [66]	41.5	46.3
ALPRO [34]	42.1	45.9
MERLOT [70]	43.1	-
VIOLET [22]	43.9	47.9
OmniVL	**44.1**	**51.0**

OmniVL achieves new state-of-the-art or at least competitive results on a wide scope of downstream tasks. When using ViT-Base scale model to pretrain on a moderate data scale (e.g., ~ 14M image-text, ~2.5M video-text), we achieve state-of-the-art performance on image-text retrieval (82.1/64.8 R@1 on COCO for image-to-text / text-to-image), image captioning (39.8 BLEU@4 on COCO), text-to-video retrieval (47.8 R@1 on MSRVTT), and video question answering (51.9% accuracy on MSVD).