



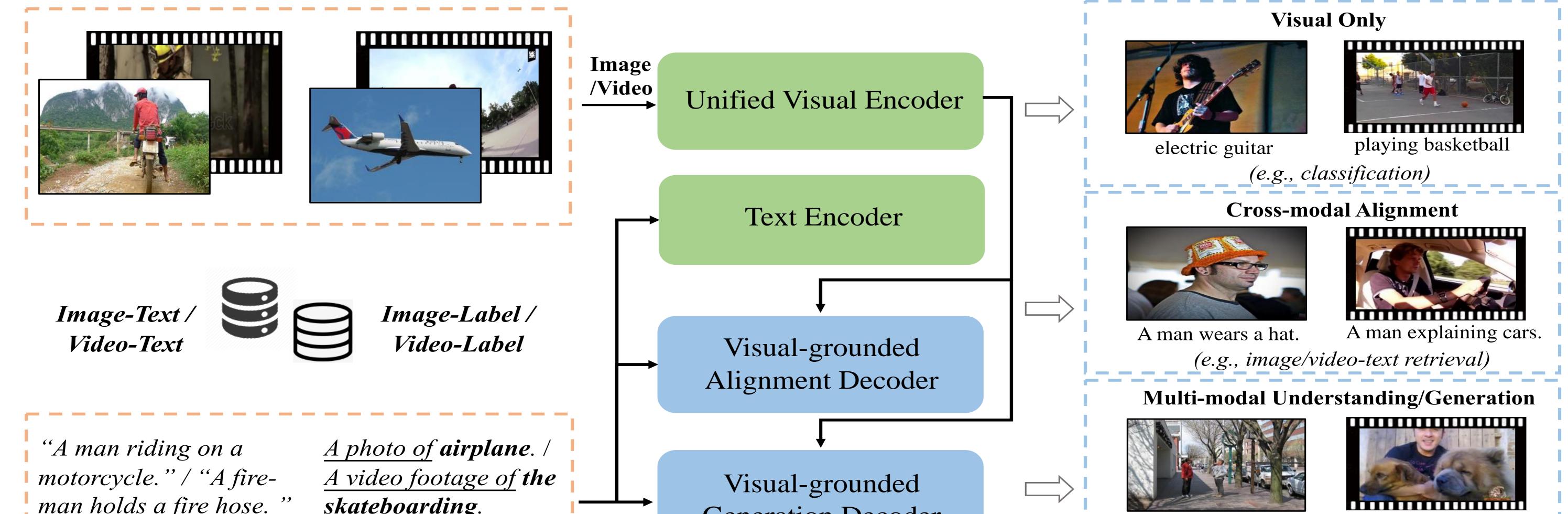
## **OmniVL: One Foundation Model for Image-Language and** Video-Language Tasks

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

<sup>1</sup>Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University <sup>2</sup>Shanghai Collaborative Innovation Center on Intelligent Visual Computing <sup>3</sup>Microsoft Cloud + AI, <sup>4</sup>Microsoft Research Asia. (<sup>†</sup> 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.



man notas a fire nose.	skalebbar ung.	Text	Generation Decoder	Q: How many people?	Q: who holds two dogs?	L
				A: Two	A: a man	

## **Pretraining Corpus Unification**

## **Modality Unification**

## **Functionality Unification**

(e.g., image/video question answering)

**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.

	# Img-Tex	xt		coo	CO (5	5K tes	t set)			Flick	r30K (	1K te	est set)			# Img-Tex	•			aps				coco	Caption	
Method	Pairs			TR	IR				TR	-		IR		Method	Pairs	in-doi	main S	near-d C	omain S	out-de	domain S	ove C	verall Kar S B@4		thy test C	
VirTex [46]	-		-	-	_	38.1	62.8	-	-	_	-	35.1	64.6	-	Enc-Dec [11]	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1		110.9
UNITER [13]	4M	6	5.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8	VinVL [71]	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
OSCAR [40]	4M	7	0.0	91.1	95.5	54.0	80.8	88.5	_	_	-	-	-	_	LEMON [28]	12M	104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-
UNIMO [39]	4M		_	_	_	_	_	_		98.9	99.8	78.0	94.2	97.1	BLIP [35]	14M	111.3	15.1	104.5	14.4	102.4	13.7	105.1	14.4	38.6	129.7
VLMO [60]	4M	7.	48	93.1	96.9	57.2	82.6	89.8					95.7		SIMVLM [61]	1.8B	-	-	-	-	-	-	94.8	13.1	39.0	134.8
OmniVL	4M*	-											97.0		$OFA_{14M}$ [58]	14M	-	-	-	-	-	-	-	-	38.7	130.5
Omm vL	4141	- 17	0.0	23.0	97.5	30.3	02.0	09.5	34.5	<b>77.</b> 0	<i></i>	0.5.4	97.0	20.0	OFA [58] OmniVL	21.4M 14M*	104.6	- 15.0	- 108.3	- 14.9	- 106.3	- 14.2	- 107.5	14.7	41.0 39.8	138.2 133.9
FLAVA [53]	70 <b>M</b>	6	1.5	82.1	89.6	50.1	74.4	83.2	85.4	95.7	98.3	73.2	92.7	95.5		14141	104.0	15.0	100.5	14.2	100.5	14.2	107.5	1-1.7	57.0	155.7
METER [21]	404M	7	6.2	93.2	96.8	57.1	82.7	90.1	94.3	99.6	99.9	82.2	96.3	98.4	Method		B@3	F	3@4	Μ	IETEC	OR	R	DUGE	-L	CIDEr
ALIGN [29]	1.8B	7	7.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	100.0	84.9	97.4	98.6	D: I STM (72)				0.07		0 15					
ALBEF [36]	14M	7	7.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	100.0	85.6	97.5	98.9	Bi-LSTM [73]		-		0.87		8.15			-		-
BLIP [35]	14 <b>M</b>	8	0.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8	EMT [74]		-		4.38		11.55			27.44		0.38
Florence [69]	900M	8	1.8	95.2	-	63.2	85.7	' <u>-</u>	97.2	99.9	-	87.9	98.1	-	VideoBERT [5		6.80		4.04		11.01			27.50		0.49
OmniVL	14M*	8	2.1	95.9	98.1						100.0				ActBERT [75]		8.66	-	5.41		13.30	)		30.56		0.65
0	1 1111	10									10010	0/15	27.0		AT [27]		-		8.55		16.93			35.54		1.06
	I		m		Video	Details				7.	na ahat	Datai			UniVL [45]		16.46	1	1.17		17.57			40.09		1.27
Method		1	MSR		video	Retrie D	iDeMo	.	Ν	1SRVT	ro-shot T		vai DiDeM	D	OmniVL	1	12.87		8.72		14.83			36.09		1.16
ClipBERT [33]	I	22.0	46.	.8 59	9.9	20.4	48.0	60.8	_	-	_	_	_	_	Method	# Img-1	- Text Pairs	te	st-dev	test-	std M	lethod	-		SRVTT	Menn
TT-CE+ [14]		29.6	61.	6 74	4.2	21.6	48.6	62.9	-	-	-	-	-	-								lethou		IVIS	5K 1 1	MSVD
VideoCLIP [64]		30.9	55.	4 66	5.8	-	-	-	10.4	22.2	30.0	16.6	46.9	-	FLAVA [53] OSCAR [40]		8M M		72.80	73.4	44 CI	lipBER	T [33]		37.4	-
FiT [6]		32.5	61.		1.2	31.0	59.8	72.4	18.7	39.5	51.6	21.1	46.0	56.2	ALBEF [36]		4M		73.16 75.84	76.	1111	ıstAsk [	66]	i	41.5	46.3
TT-CE+ (+QB-N	IORM) [9]	33.3	63.				50.8	64.4	=	-	-	-	=	=	BLIP [35]		4M		77.54	77.0	· A 1	LPRO [	34]	i	42.1	45.9
ALPRO [34]	,	33.9	60.		3.2	35.9	67.5	78.8	24.1	44.7	55.4	23.8	47.3	57.9	METER [21]		4M		77.68	77.		ERLO	Г [ <mark>70</mark> ]	i	43.1	-
VIOLET [22]		34.5	63.	0 73			62.8	74.7	25.9	49.5	59.7	23.5	49.8	59.8	SimVLM [61]		8B		77.87	78.	14 V	IOLET	[22]	ł	43.9	47.9
OmniVL		47.8	74.				79.5	85.4	42.0	63.0	73.0	40.6			OFA [58]		.4M		78.00	78.	10 O	mniVL			44.1	51.0
															OmniVL	14	$M^*$		78.33	78.	35 -					

Number of the second	Mada	# Img-Tex			co	CO (5	5K tes	t set)			Flick	r30K (	1K te	st set)			# Img-Text				NoC	aps					CO Caption	
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Method				TR			IR			TR		•	IR		Method		in-do C		near-d C	omain S	out-de C	omain S	ove C		-	thy test C	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	VirTex [46]	-		-	-	-	38.1	62.8	-	-	-	-	35.1	64.6	-	Enc-Dec [11]	15M	92.6	12.5	88.3	12.1	94.5	11.9	90.2	12.1	-	110.9	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	UNITER [13]	4M	6	5.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8	VinVL [71]	5.7M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3	
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	OSCAR [40]	4M	7	0.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-			104.5	14.6	100.7	14.0	96.7	12.4	100.4	13.8	-	-	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	UNIMO [39]	4M		-	-	_	-	-	-	89.4	98.9	99.8	78.0	94.2	97.1			111.3	15.1	104.5	14.4	102.4	13.7				129.7	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $			7	4.8	93.1	96.9	57.2	82.6	89.8									-	-	-	-	-	-	94.8	13.1		134.8	
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 95.9 99.8 100.0 84.9 97.4 98.6   BLIP [35] 14M 80.6 95.2 97.6 63.1 85.3 91.0 87.9 98.1 - <td></td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td></td> <td></td>																		-	-	-	-	-	-	-	-			
FLAVA [53] 70M 61.5 82.1 89.6 50.1 74.4 83.2 85.4 97.7 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 94.3 97.2 99.8 100.0 84.9 97.4 98.6   BLIP [35] 14M 80.6 95.2 97.6 63.1 85.3 99.8 100.0 87.2 97.5 98.8 - 0.87 8.15 - - 0.87 8.15 - - 0.87 8.15 - - 0.87 8.15 - - 0.87 8.15 - - 0.87 8.15 - - 0.67 8.15 - - 0.87 8.15 - - 0.87 8.15 - - 0.87 9.81 - - 8.66 5.41 13.30 30.56 0.0   OmmiVL14.4482.662.9 </td <td></td> <td>104.6</td> <td>15.0</td> <td>- 108.3</td> <td>- 14.9</td> <td>- 106.3</td> <td>14.2</td> <td>- 107.5</td> <td>14.7</td> <td></td> <td>133.9</td>																		104.6	15.0	- 108.3	- 14.9	- 106.3	14.2	- 107.5	14.7		133.9	
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 84.9 97.4 98.6   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   CommiVL 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 VideoCLIP [64] Text-to-Video Retrieval VideoCLIP [64] Zero-shot Retrieval MSRVTT DiDeMo MSRVTT DiDeMo MSRVTT Method Taxt-to-Video Retrieval VideoCLIP [64] Gale $a$																												
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   OmniVL 14M* 82.1 95.9 98.1 64.8 86.1 91.6 97.2 99.9 87.9 98.1 - 4.38 11.55 27.44 0.3   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 MSRVTT Zero-shot Retrieval MSRVTT DiDeMo MSRVTT DiDeMo MSRVTT DiDeMo 11.48 77.6 44.80 60.93 35.54 14.43   VideoCLIP [64] 30.9 50.4 66.8 $               -$ </td <td></td> <td>Method</td> <td>  1</td> <td>3@3</td> <td>E</td> <td>3@4</td> <td>N</td> <td><b>IETEC</b></td> <td>DR</td> <td>R</td> <td>DUGE</td> <td>-L</td> <td>CIDEr</td>																Method	1	3@3	E	3@4	N	<b>IETEC</b>	DR	R	DUGE	-L	CIDEr	
ALDEP [35] 14M 77.5 94.3 97.2 90.7 99.3 99.3 90.0 87.3 97.3 99.7 98.1 - 4.38 11.55 27.44 0.3   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 - 4.38 11.55 27.44 0.3   CommivL 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 MSRVTT DiDeMo Zero-shot Retrieval MSRVTT DiDeMo MSRVTT DiDeMo 66.8 4.04 11.17 17.57 40.09 1.2   VideoCLIP [64] 30.9 55.4 66.8 - - 10.4 22.2 30.0 16.6 46.9 -		· · · · · · · · · · · · · · · · · · ·														Bi-LSTM [73]		_	(	0.87		8.15			-		_	
BLIP [55] 14M 80.6 95.2 97.6 05.3 91.1 96.6 99.8 100.0 87.2 97.3 98.1 97.2 99.9 87.9 98.1 99.1   OmniVL 14M* 82.1 95.9 98.1 64.8 86.1 91.6 97.2 99.9 87.9 98.1 - 8.66 5.41 13.30 30.56 0.0   Method Text-to-Video Retrieval MSRVTT Zero-shot Retrieval MSRVTT DiDeMo Zero-shot Retrieval MSRVTT DiDeMo MSRVTT DiDeMo 12.87 8.72 14.83 36.09 1.4   ClipBERT [33] TI-CE+ [14] 22.0 46.8 59.9 20.4 48.0 60.8 -																		_							27 44		0.38	
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	BLIP [35]	14 <b>M</b>	8	0.6	95.2	97.6	63.1	85.3	91.1	96.6	99.8	100.0	87.2	97.5	98.8		51										0.49	
Text-to-Video Retrieval MSRVTT   Zero-shot Retrieval MSRVTT   Zero-shot Retrieval MSRVTT   AT [27]   -   8.55   16.93   35.54   1.1     Method   Text-to-Video Retrieval MSRVTT   DiDeMo   MSRVTT   DiDeMo   MSRVTT   DiDeMo   16.46   11.17   17.57   40.09   1.1     ClipBERT [33] TT-CE+ [14]   29.6   61.6   74.2   21.6   48.6   62.9   -   -   -   -   8.72   14.83   36.09   1.1     VideoCLIP [64]   30.9   55.4   66.8   - <td>Florence [69]</td> <td>900M</td> <td>8</td> <td>1.8</td> <td>95.2</td> <td>-</td> <td>63.2</td> <td>85.7</td> <td>-</td> <td>97.2</td> <td>99.9</td> <td>-</td> <td>87.9</td> <td>98.1</td> <td>-</td> <td>•</td> <td>-</td> <td></td>	Florence [69]	900M	8	1.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-	•	-											
Method   Text-to-Video Retrieval MSRVTT   Zero-shot Retrieval MSRVTT   DiDeMo   UniVL [45]   16.46   11.17   17.57   40.09   1.17     ClipBERT [33] TT-CE+ [14]   22.0   46.8   59.9   20.4   48.0   60.8   -	OmniVL	$14M^*$	8	2.1	95.9	98.1	64.8	86.1	91.6	97.3	99.9	100.0	87.9	97.8	99.1			0.00										
Method Text-to-Video Retrieval MSRVTT DiDeMo DiD			•							1							1	-									1.06	
Method MSRVTT DiDeMo MSRVTT DiDeMo OmniVL 12.87 8.72 14.83 36.09 1.   ClipBERT [33] 22.0 46.8 59.9 20.4 48.0 60.8 - <td< td=""><td></td><td>1</td><td></td><td>Т</td><td>`ext-to</td><td>-Video</td><td>Retrie</td><td>eval</td><td></td><td></td><td>Ze</td><td>ro-shot</td><td>Retrie</td><td>val</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>1.27</td></td<>		1		Т	`ext-to	-Video	Retrie	eval			Ze	ro-shot	Retrie	val													1.27	
TT-CE+ [14] 29.6 61.6 74.2 21.6 48.6 62.9 - <t< td=""><td>Method</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Ν</td><td>-</td><td></td><td></td><td></td><td>0</td><td>OmniVL</td><td></td><td>2.87</td><td></td><td>8.72</td><td></td><td>14.83</td><td>; </td><td></td><td>36.09</td><td></td><td>1.16</td></t<>	Method									Ν	-				0	OmniVL		2.87		8.72		14.83	; 		36.09		1.16	
TT-CE+ [14] 29.6 61.6 74.2 21.6 48.6 62.9 - <t< td=""><td>ClipBERT [33]</td><td></td><td>22.0</td><td>46</td><td>.8 5</td><td>9.9</td><td>20.4</td><td>48.0</td><td>60.8</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>_</td><td>Method</td><td># Img-T</td><td>ext Pairs</td><td>te</td><td>st-dev</td><td>test</td><td>std M</td><td>lethod</td><td>-</td><td>MS</td><td>RVTT</td><td>MSVD</td></t<>	ClipBERT [33]		22.0	46	.8 5	9.9	20.4	48.0	60.8	-	-	-	-	-	_	Method	# Img-T	ext Pairs	te	st-dev	test	std M	lethod	-	MS	RVTT	MSVD	
VideoCLIP [64] 30.9 55.4 66.8 - - 10.4 22.2 30.0 16.6 46.9 - OSCAR [40] 4M 73.16 73.44 CupBERT [33] 37.4   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 ALBEF [36] 14M 75.84 76.04 JustAsk [66] 41.5 44.   TT-CE+ (+QB-NORM) [9] 33.3 63.7 76.3 24.2 50.8 64.4 - - - BLIP [35] 14M 75.84 76.04 ALPRO [34] 42.1 44.   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 METER [21] 404M 77.68 77.64 MERLOT [70] 43.1   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 SimVLM [61] 1.8B 77.87 78.14 VIOLET [22] 43.9	TT-CE+ [14]		29.6	61	.6 7	4.2	21.6	48.6	62.9	-	-	-	-	-	-	FLAVA [53]	68	м		72.80	-							
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 ALBEF [36] 14M 75.84 76.04 JustAsk [66] 41.5 44.1   TT-CE+ (+QB-NORM) [9] 33.3 63.7 76.3 24.2 50.8 64.4 - - - BLIP [35] 14M 77.54 77.62 ALPRO [34] 42.1 44.5 44.7 44.7 55.4 23.8 47.3 57.9 METER [21] 404M 77.68 77.64 MERLOT [70] 43.1   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 SimVLM [61] 1.8B 77.87 78.14 VIOLET [22] 43.9 44.1 54.9 54.9 54.9 54.6 74.3 0FA [58] 21.4M 78.00 78.10 0mniVL 44.1 54.9   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	VideoCLIP [64]		30.9	55	.4 6	6.8	-	-	-	10.4	22.2	30.0	16.6	46.9	-						73.	44	· · ·				-	
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 METER [21] 404M 77.68 77.64 MERLOT [70] 43.1   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 SimVLM [61] 1.8B 77.87 78.14 VIOLET [22] 43.9 4'   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 0FA [58] 21.4M 78.00 78.10 0mniVL 44.1 51	FiT [6]		32.5	61	.5 7	1.2	31.0	59.8	72.4	18.7	39.5	51.6	21.1	46.0	56.2							04 <sup>Ju</sup>					46.3	
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 SimVLM [61] 1.8B 77.87 78.14 VIOLET [22] 43.9 4'   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 0FA [58] 21.4M 78.00 78.10 0mniVL 44.1 51	TT-CE+ (+QB-NC	ORM) [9]	33.3	63	.7 7	6.3	24.2	50.8	64.4	-	-	-	-	-	-	BLIP [35]	14	М		77.54	77.	94	-	-			45.9	
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	ALPRO [34]		33.9	60	.7 7	3.2	35.9	67.5	78.8	24.1	44.7	55.4	23.8	47.3	57.9	METER [21]				77.68							-	
	VIOLET [22]		34.5	63	.0 7	3.4	32.6	62.8	74.7	25.9	49.5	59.7	23.5	49.8	59.8												47.9	
OmniVL 14M* 78.33 78.35	OmniVL		47.8	74	.2 8	3.8	52.4	79.5	85.4	42.0	63.0	73.0	40.6	64.6	74.3	OFA [58] OmniVL				78.00 78.33			mniVL			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.,  $\sim 14M$  image-text,  $\sim 2.5M$  video-text), we achieve stateof-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).