VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset

Anonymous Author(s)
Affiliation
Address
email

Appendix

A More about VAST Foundation Model

A.1 Pretraining Settings

Specific pretraining configurations of VAST including training corpora, training steps for each corpus (i.e., dataset mix ratio), and training objectives on each corpus are presented in Table 1. To enhance data quality, we use trained vision captioner to generate new captions for CC12M and LAION datasets and replace original captions with them. It is noted that VAST have been trained for relatively small steps (205K steps), but have already shown excellent performances on various types of downstream tasks, and we believe that training by more steps can further increase the model capabilities.

Table 1: Model configurations and pretraining settings of VAST. It is noted that 400M Web data used in CLIP [1] and LAION-400M [2] used in EVAClip [3] are also counted for training samples statics. LAION-102M and LAION-110M are both random sampled subsets from LAION-400M. Regarding training objectives, 'ret' represents for the combination of VCC and VCM, while 'cap' denotes VCG, and differnet modality groups are separeted by '%'.

Model	Param	Sample	Training Corpus	Batch Size	Steps	Epoch	Objectives
VAST	1.3B	442M	VAST-27M	1024	60000	2.3	ret%vast%vat%vst%vt%at + cap%vast%vat%vst%vt%at
			VALOR-1M	1024	25000	25	ret%vat%vt%at + cap%vat%vt%at
			WavCaps	1024	15000	38	ret%at + cap%at
			CC4M	2048	30000	12	ret%vt + cap%vt
			CC12M	2048	20000	4	ret%vt + cap%vt
			LAION-110M	2048	55000	1	ret%vt + cap%vt

A.2 Downstream Datasets Descriptions

We evaluate VAST on multiple popular domnstream datasets, including MSRVTT, VATEX, YouCook2, VALOR-32K, MSVD, LSMDC, DiDeMo, ActivityNet Caption, TGIF, MUSIC-AVQA, TVC, Clotho, AudioCaps, MSCOCO, Flickr30K and VQAv2. Specific train/val/test splits of those benchmarks can be found in Table 2 and specific descriptions of them are as follows.

MSRVTT [4] contains 10K video clips and 200K captions. The videos cover a wide range of topics and scenes, including human activities, sports, natural landscapes, and more. We evaluate text-to-video retrieval, video captioning and video QA on this dataset. Following methods presented in Table 4, we use the '1K-A split' for retrieval evaluation. For captioning and QA, we use the standard split.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 2: Downstream dataset splits.

Task Type	Modal Type	Benchmark	#Videos/#Images			#Captions/#QA-pairs		
			Train	Val	Test	Train	Val	Test
Retrieval	V-T(SM)	MSCOCO	113287	5000	5000	566747	25010	25010
		Flickr30K	29000	1014	1000	145000	5070	5000
	A-T	ClothoV1	2893	1045	-	14465	5225	-
		ClothoV2	3839	1045	-	19195	5225	-
		AudioCaps	49291	428	816	49291	2140	4080
	V-T(MM)	MSRVTT	9000	-	1000	180000	-	1000
		YouCook2	10337	3492	-	10337	3492	-
		VALOR-32K	25000	3500	3500	25000	3500	3500
		VATEX	25991	1500	1500	259910	1500	1500
		DiDeMo	8394	1065	1003	8394	1065	1003
		ANET	10009	-	4917	10009	-	4917
		LSMDC	101046	7408	1000	101046	7408	1000
Caption	V-T(SM)	MSCOCO	113287	5000	5000	566747	25010	25010
		MSVD	1200	100	670	48774	4290	27763
	A-T	ClothoV1	2893	1045	-	14465	5225	-
		ClothoV2	3839	1045	-	19195	5225	-
		AudioCaps	49838	495	975	49438	2475	4875
	V-T(MM)	MSRVTT	6513	497	2990	130260	9940	59800
		YouCook2	10337	3492	-	10337	3492	-
		VALOR-32K	25000	3500	3500	25000	3500	3500
		VATEX	25991	3000	6000	259910	30000	60000
		TVC	86603	10841	-	174350	43580	-
QA	V-T(SM)	MSVD-QA	1200	250	520	30933	6,415	13157
		TGIF-FrameQA	32345	-	7132	39389	-	13691
		VQAv2	82783	40504	37K/81K	4437570	2143540	1.1M/4.5M
	V-T(MM)	MSRVTT-QA	6513	497	2990	158581	12278	72821
		MUSIC-AVQA	9277	3815	6399	32087	4595	9185
		ANET-QA	3200	1800	800	32000	18000	8000

VATEX [5] contains 41,250 video clips sourced from Kinetics-600 dataset [6] and 825,000 sentence-level descriptions. We evaluate text-to-video retrieval and video captioning on this dataset. For captioning, we use the official split. For retrieval. we follow the HGR [7] split protocol.

YouCook2 [8] consists of 14K video clips from 2K instructional cooking videos from YouTube. Each video includes multiple actions performed by the chef, along with corresponding textual descriptions and temporal annotations. We evaluate text-to-video retrieval and video captioning on this dataset with official splits.

VALOR-32K [9] is an audiovisual video-language benchmark that contains 32K 10 seconds long audible video clips sourced from AudioSet [10]. Each video clip is annotated with an audiovisual caption which simultaneously describes both visual and audio contents in videos. We evaluate text-to-video retrieval and video captioning on this dataset with official splits.

MSVD [11] contains 1,970 videos, each of which is paired with around 40 captions. We evaluate video QA on this dataset and use the split proposed by Xu et al. [12].

LSMDC [13] consists of 118K clips form 202 movies, each of which is paird with one caption. We evaluate text-to-video retrieval on this dataset with official split.

DiDeMo [14] contains 10K long-form videos from Flickr and for each video, four short sentences are annotated in temporal order. We follow methods in Table 4 to concatenate those short sentences and evaluate 'paragraph-to-video' retrieval on this benchmark. The official split is used.

ActivityNet Caption [15] contains 20K long-form videos (180s as average length) from YouTube and 100K captions. We evaluate text-to-video retrieval and video QA on this dataset. For retrieval we use official split and for video QA, split proposed by Yu et al. [16] is used.

TGIF [17] contains three video QA benchmarks including TGIF-Action, TGIF-transition and TGIF-Frame, and the first two are multiple-choice QA while the last is open-ended QA. We evaluate VAST on TGIF-frame benchmark with official split.

Table 3: Downstream task finetuning settings. Lr, Bs, Epo, Obj and Res denote learning rate, batch size, epoch, training objectives and resolution, respectively. Vf(Tr), Vf(Te), Ac(Tr), Ac(Te) denotes sampled video frames (Vf) or audio clips (Ac) in training (Tr) and testing (Te), respectively. The marks in Obj are the same as those in Table 1. Most hyperparameters in the table are not precisely tuned.

Task	Modality	Benchmark	Lr	Bs	Epo	Obj	Vf(Tr)	Vf(Te)	Ac(Tr)	Ac(Te)	Res
RET	V-T(SM)	MSCOCO	1e-5	256	5	ret%vt	-	-	-	-	384
		Flickr	1e-5	256	5	ret%vt	-	-	-	-	384
	A-T	ClothoV1/V2	2e-5	64	10	ret%at	-	-	3	3	-
		AudioCaps	2e-5	64	10	ret%at	-	-	1	1	-
	V-T(MM)	MSRVTT	2e-5	64	3.6	ret%vast	8	16	1	1	224
		YouCook2	3e-5	64	30	ret%vast	8	16	1	1	224
		VALOR-32K	2e-5	64	10	ret%vat	8	8	1	1	224
		VATEX	2e-5	64	2.5	ret%vast	8	16	1	1	224
		DiDeMo	2e-5	64	40	ret%vat	8	32	2	2	224
		ANET	2e-5	64	20	ret%vat	8	32	2	2	224
		LSMDC	2e-5	64	5	ret%vat	8	32	1	1	224
CAP	V-T(SM)	MSCOCO	1e-5	64	5	cap%vt	-	-	-	-	480
		MSCOCO(SCST)	2.5e-6	64	2.5	cap%vt	-	-	-	-	480
	A-T	ClothoV1/V2	2e-5	64	10	cap%at	-	-	3	3	-
		AudioCaps	2e-5	64	10	cap%at	-	-	1	1	-
	V-T(MM)	MSRVTT	2e-5	128	10	cap%vast	8	8	1	1	224
		YouCook2	3e-5	64	30	cap%vast	8	16	1	1	224
		VALOR-32K	1e-5	64	10	cap%vat	8	12	1	1	224
		VATEX	2e-5	64	10	cap%vast	8	20	1	1	224
		VATEX(SCST)	7e-6	64	5	cap%vast	8	20	1	1	224
		TVC	3e-5	64	40	cap%vst	8	8	-	-	224
QA	V-T(SM)	MSVD-QA	1e-5	64	10	qa%vt	8	14	-	-	224
		TGIF-FrameQA	2e-5	64	10	qa%vt	4	4	-	-	224
		VQAv2	2e-5	128	20	qa%vt	-	-	-	-	384
	V-T(MM)	MSRVTT-QA	2e-5	64	4.5	qa%vast	8	8	1	1	224
		MUSIC-AVQA	2e-5	64	20	qa%vat	8	8	2	2	224
		ANET-QA	2e-5	64	10	qa%vat	8	16	2	2	224

MUSIC-AVQA [18] is a audiovisual video QA benchmark containing more than 45K Q-A pairs covering 33 different question templates spanning over different modalities and question types. The offcial split is used.

TVC [19] is a multi-channel video captioning dataset containing 108K video moments and 262K paired captions. Video subtitles can be used as additional input. We evaluate video captioning on this benchmark with official split.

Clotho [20] contains 15-30 second audio clips and has two versions. The original (v1) has 4981 audios, while an expanded version (v2) includes 6974 audios, enlarging solely the training set. We evaluate text-to-audio retrieval and audio captioning on those benchmarks with official split.

AudioCaps [21] contains 51K 10-second clips, with one caption in the training set and five in the validation and test sets. We evaluate text-to-audio retrieval and audio captioning on it. For captioning, we use the official split, and for retrieval we follow the Sophia et al. [22] split protocol.

MSCOCO [23] contains 123K images each of which is paired with 5 annotated captions, We evaluate text-to-image retrieval and image captioning on this dataset with Karpathy split [24].

Flickr30K [25] contains 31K images each of which is paired with 5 annotated captions, We evaluate text-to-image retrieval on this dataset with Karpathy split [24].

VQAv2 [26] was used as the basis of the 2017 VQA Challenge2, it contains 1.1M questions with 11.1M answers relating to MSCOCO images. The official split is used.

A.3 Finetuning Settings

Specific finetuning hyperparameters of VAST for different benchmarks are presented in Table 3.

Table 4: Performance comparison on Text-to-Video Retrieval benchmarks. For fair comparisons, performances before employing post-processing such as dual-softmax [27] are reported and compared. All benchmarks are multi-modal benchmarks (containing audio and subtitle tracks). Methods utilizing audio or subtitle modalities besides vision for video representation are marked with gray background color.

Method	Sample	MSRVTT			DiDeMo			ActivityNet		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Singularity [28]	17M	41.5	68.7	77.0	53.9	79.4	86.9	47.1	75.5	85.5
OmniVL [29]	17M	47.8	74.2	83.8	52.4	79.5	85.4	-	-	-
HiTeA [30]	17M	46.8	71.2	81.9	56.5	81.7	89.7	49.7	77.1	86.7
VINDLU-L [31]	25M	48.8	72.4	82.2	59.8	86.6	91.5	55.9	82.3	90.9
LAVENDER [32]	30M	40.7	66.9	77.6	53.4	78.6	85.3	-	-	-
All-in-one [33]	138M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
CLIP4Clip [34]	400M	44.5	71.4	81.6	43.4	70.2	80.6	40.5	72.4	-
X-CLIP [35]	400M	49.3	75.8	84.8	47.8	79.3	-	46.2	75.5	-
mPLUG-2 [36]	417M	53.1	77.6	84.7	56.4	79.1	85.2	-	-	-
UMT-L [37]	425M	58.8	81.0	87.1	70.4	**90.1**	**93.5**	66.8	89.1	94.9
CLIP-VIP [38]	500M	54.2	77.2	84.8	50.5	78.4	87.1	53.4	81.4	90.0
MMT [39]	136M	26.6	57.1	69.6	-	-	-	28.7	61.4	-
AVLNet [40]	136M	22.5	50.5	64.1	-	-	-	-	-	-
Gabeur et al. [41]	136M	28.7	59.5	70.3	-	-	-	29.0	61.7	-
ECLIPSE [42]	400M	-	-	-	44.2	-	-	45.3	75.7	86.2
VALOR-L [9]	433.5M	54.4	79.8	87.6	57.6	83.3	88.8	63.4	87.8	94.1
VAST	**442M**	**63.9**	**84.3**	**89.6**	**72.0**	89.0	91.4	**70.5**	**90.9**	**95.5**

Method	VATEX		
	R@1	R@5	R@10
Support-set [43]	44.9	82.1	89.7
CLIP4Clip [34]	55.9	89.2	95.0
DCR [44]	65.7	92.6	96.7
VALOR-L	76.9	96.7	98.6
VAST	**83.0**	**98.2**	**99.2**

Method	VALOR-32K		
	R@1	R@5	R@10
Frozen [45]	32.9	60.4	71.2
CLIP4Clip [34]	43.4	69.9	79.7
AVLNet [40]	21.6	47.2	59.8
VALOR-L [9]	73.2	91.6	95.4
VAST	**80.0**	**93.7**	**96.6**

Method	YouCook2		
	R@1	R@5	R@10
UniVL [46]	28.9	57.6	70.0
MELTR [47]	33.7	63.1	74.8
VLM [48]	27.1	56.9	69.4
VALUE [49]	31.3	53.0	62.2
VAST	**50.4**	**74.3**	**80.8**

Table 5: Performance comparison on zero-shot Text-to-Video Retrieval benchmarks. Methods utilizing audio or subtitle modalities besides vision for video representation are marked with gray background color.

Method	Sample	MSRVTT			DiDeMo		
		R@1	R@5	R@10	R@1	R@5	
Frozen [45]	5M	18.7	39.5	51.6	21.1	46.0	56.2
ALPRO [50]	5M	24.1	44.7	55.4	23.8	47.3	57.9
Singularity [28]	5M	28.4	50.2	59.5	36.9	61.6	69.3
HiTeA [30]	17M	34.4	60.0	69.9	43.2	69.3	79.0
OmniVL [29]	18M	42.0	63.0	73.0	40.6	64.6	74.3
VIOLET [51]	183M	25.9	49.5	59.7	23.5	49.8	59.8
UMT-L [37]	425M	40.7	63.4	71.8	48.6	72.9	79.0
Florence [52]	900M	37.6	63.8	72.6	-	-	-
VAST	**443M**	**49.3**	**68.3**	**73.9**	**55.5**	**74.3**	**79.6**

A.4 Detailed Comparisons to State-of-the-Art Methods

Text-to-Video Retrieval. We compare VAST to SOTA methods on six multi-modal text-to-video retrieval benchmarks. As shown in Table 4, VAST improves previous SOTA methods by 5.1, 1.6, 3.7, 6.1 points on MSRVTT, DiDeMo, ActivityNet, VATEX benchmarks, respectively. Besides above mentioned vision-oriented benchmarks, VAST outperforms VALOR-L [9] by 6.8 points on the audio-oriented benchmark VALOR-32K, and surpass MELTR [47] by 16.7 points on the subtitle-oriented benchmark YouCook2, which demonstrate the strong generalization capabilities of VAST towards different types of downstream datasets. In addition, the zero-shot retrieval performance comparison is

Table 6: Performance comparison on Video QA benchmarks. MSVD-QA and TGIF-QA are vision-only benchmarks while the others are multi-modal benchmarks. Methods utilizing audio or subtitle modalities besides vision for video representation are marked with gray background color.

Method	Sample	MSRVTT-QA	MSVD-QA	TGIF-QA	ActivityNet-QA	MUSIC-AVQA
ClipBERT [53]	5.4M	37.4	-	60.3		-
ALPRO [50]	5M	42.1	45.9	-	-	-
VIOLETv2 [54]	5M	44.5	54.7	72.8	-	-
Clover [55]	5M	43.9	51.9	71.4	-	-
OmniVL [29]	17M	44.1	51.0			-
HiTeA [30]	17M	45.9	55.3	73.2	46.4	-
SINGULARITY [28]	17M	43.5	-	-	43.1	-
VINDLU-B [31]	17M	43.8	-	-	44.6	-
LAVENDER [32]	30M	45.0	56.6	73.5	-	-
JustAsk [56]	69M	41.5	46.3	-	38.9	-
MERLOT [57]	180M	43.1	-	69.5	41.4	-
All-in-one [33]	228.5M	46.8	48.3	66.3	-	-
FrozenBiLM [58]	410M	47.0	54.8	68.6	43.2	-
mPLUG-2 [36]	417M	48.0	58.1	75.4	-	-
UMT-L [37]	425M	47.1	55.2	-	-	-
InternVideo [59]	646M	47.1	55.5	72.2	-	-
GIT [60]	1.7B	43.2	56.8	72.8	-	-
MaMMUT [61]	2B	49.5	60.2	-	-	-
Flamingo (80B) [62]	2.3B	47.4	-	-	-	-
VideoCoCa (2.1B) [63]	4.8B	46.0	56.9	-	-	-
GIT2 (5.1B) [60]	12.9B	45.6	58.2	74.9	-	-
VALOR-L [9]	433.5M	49.2	60.0	78.7	48.6	78.9
VAST(1.3B)	**442M**	**50.1**	**60.2**	**79.1**	**50.4**	**80.7**

Table 7: Performance comparison on Video Captioning benchmarks. All benchmarks are multi-modal benchmarks. BLEU@4 and CIDEr (C) metrics are reported. On VATEX benchmark, we follow most state-of-the-art methods [60; 9; 64] employing SCST finetuning [65] after cross-entropy training, and corresponding results are marked with '*'. Methods utilizing audio or subtitle modalities besides vision for video representation are marked with gray background color.

Method	Sample	MSRVTT		VATEX		YouCook2		TVC		VALOR-32K	
		B@4	C	B@4	C	B@4	C	B@4	C	B@4	C
SwinBERT [66]	-	41.9	53.8	38.7	73.0	9.0	109.0	14.5	55.4	5.4	27.3
VIOLETv2 [54]	5M	-	58.0	-	-	-	-	-	-	-	-
HiTeA [30]	5M	-	62.5	-	-	-	-	-	-	-	-
LAVENDER [32]	30M	-	60.1	-	-	-	-	-	-	-	-
MaMMUT [61]	2B	-	73.6	-	-	-	-	-	-	-	-
GIT [60]	1.7B	53.8	73.9	41.6*	91.5*	10.3	129.8	16.2	63.0		
GIT2(5.1B) [60]	12.9B	54.8	75.9	42.7*	94.5*	9.4	131.2	16.9	66.1	-	-
SMPFF [67]	-	48.4	58.5	39.7	70.5	-	-	-	-	7.5	37.1
VALUE [49]	136M	-	-	-	58.1	12.4	130.3	11.6	50.5	-	-
UniVL [46]	136M	41.8	50.0	-	-	17.4	181.0	-	-	-	-
MELTR [47]	136M	44.2	52.8	-	-	17.9	190.0	-	-	-	-
CLIP4Caption++ [64]	400M	-	-	40.6*	85.7*	-	-	15.0	66.0	-	-
VALOR-L [9]	433.5M	54.4	74.0	**45.6***	95.8*	-	-	-	-	9.6	61.5
VAST(1.3B)	**442M**	**56.7**	**78.0**	45.0*	**99.5***	**18.2**	**198.8**	**19.9**	**74.1**	**9.9**	**62.2**

Table 8: Performance comparison on Text-to-Audio Retrieval benchmarks.

Method	Sample	ClothoV1			ClothoV2			AudioCaps		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Oncescu et al. [22]	-	9.6	-	40.1	-	-	-	25.1	-	73.2
Nagrani et al. [68]	1M	12.6	-	45.4	-	-	-	35.5	-	84.5
LAION [69]	0.63M	-	-	-	16.1	38.3	51.1	36.1	71.8	83.9
CNN14-BERT [70]	0.4M	-	-	-	21.5	47.9	61.9	35.1	70.0	82.1
HTSAT-BERT [70]	0.4M	-		-	19.7	45.7	59.4	42.2	76.5	**87.1**
VALOR-B [9]	1M	17.5	42.7	55.3	-	-	-	40.1	73.9	83.1
VAST	**28.4M**	**25.1**	**51.5**	**64.0**	**26.9**	**53.2**	**66.1**	**52.0**	**76.8**	82.9

Table 9: Performance comparison on Audio Captioning benchmarks.

Method	Sample	ClothoV1				ClothoV2				AudioCaps			
		B@4	M	R	C	B@4	M	R	C	B@4	M	R	C
Xu et al. [71]	-	15.9	16.9	36.8	37.7	-	-	-	-	23.1	22.9	46.7	66.0
CNN14-BART [70]	0.4M	-	-	-	-	18.0	18.5	40.0	48.8	27.2	24.7	49.9	75.6
HTSAT-BART [70]	0.4M	-	-	-	-	16.8	18.4	38.3	46.2	28.3	**25.0**	50.7	**78.7**
VALOR-B [9]	1M	16.2	17.4	38.2	42.3	-	-	-	-	27.0	23.1	49.4	74.1
VAST	**28.4M**	**18.5**	**18.9**	**39.9**	**50.7**	**19.0**	**19.3**	**40.8**	**51.9**	**29.5**	24.7	**50.9**	78.1

shown in Table 5, VAST achieves 49.3 and 55.5 zero-shot R@1 performance that surpasses previous SOTA by 7.3 and 6.9 points, respectively.

Video QA. We evaluate VAST on five open-ended video QA benchmarks. As shown in Table 6, VAST have achieved new SOTA performances on all benchmarks, and outperform recent proposed large-scale foundation models such as GIT [60], MaMMUT [61], Flamingo [62] and CoCa [63]. In addition, on the audiovisual video QA benchmark MUSIC-AVQA, VAST surpasses VALOR by 1.8 points, demonstrating its better capabilities to answer both visual and audio questions.

Video Captioning. In Table 7, we compare VAST to state-of-the-art methods on five multi-modal video captioning benchmarks. According to the results, VAST have achieved new state-of-the-art CIDEr score on all five benchmarks with evident margins. Compared to previous vision-language modal SOTA method GIT [60] which takes a 5.1B DaViT [81] as vision encoder and conduct pretraining on 12.9B private image-text corpus, VAST surpass it with only 22.5% parameters and 3.4% training data, demonstrating the high efficiency of our method. Compared to previous multi-modal video-language SOTA method VALOR [60], VAST can additionally process subtitle-oriented

Table 10: Performance comparison on Image-Text downstream tasks. CIDEr (C) and SPICE (S) metrics are reported for captioning. On MSCOCO caption benchmark, we follow SOTA methods [60; 9; 72] employing SCST finetuning [65], and corresponding results are marked with '*'.

Method	Sample	MSCOCO-Ret			Flickr30K-Ret			MSCOCO-Cap		VQAv2	
		R@1	R@5	R@10	R@1	R@5	R@10	C	S	dev	std
ALBEF [73]	14M	60.7	84.3	90.5	85.6	97.5	98.9	-	-	75.84	76.04
OFA [72]	18M	-	-	-	-	-	-	**154.9***	26.6*	82.0	82.0
BEiT-3 [74]	21M	67.2	87.7	92.8	90.3	**98.7**	99.5	147.6	25.4	84.19	84.03
BLIP [75]	129M	65.1	86.3	91.8	87.6	97.7	99.0	136.7	-	78.25	78.32
BLIP-2 [76]	129M	**68.3**	87.7	92.6	-	-	-	145.8	-	82.19	82.30
mPLUG-2 [36]	417M	65.7	87.1	92.6	88.1	97.6	99 .1	137.7	23.7	81.11	81.13
VALOR-L [9]	433.5M	61.4	84.4	90.9	-	-	-	152.5*	25.7*	78.46	78.62
Florence [52]	900M	63.2	85.7	-	87.9	98.1	-	-	-	80.16	80.36
PaLI [77]	1.6B	-	-	-	-	-	-	149.1	-	**84.3**	**84.3**
GIT [60]	1.7B	-	-	-	-	-	-	151.1*	26.3*	78.6	78.8
SimVLM [78]	1.8B	-	-	-	-	-	-	143.3	25.4	80.03	80.34
ALIGN [79]	1.8B	59.9	83.3	89.8	84.9	97.4	98.6	-	-	-	-
Flamingo (80B) [62]	2.3B	-	-	-	-	-	-	138.1	-	82.0	82.1
CoCa(2.1B) [80]	4.8B	-	-	-	-	-	-	143.6	24.7	82.3	82.3
GIT2(5.1B) [60]	12.9B	-	-	-	-	-	-	152.7*	26.4*	81.7	81.9
VAST	442M	68.0	**87.7**	**92.8**	**91.0**	98.5	**99.5**	149.0*	**27.0***	80.23	80.19

benchmarks such as YouCook2 and TVC, and achieves better results due to that it jointly models the relations between text and omni-modalities in videos.

Text-to-Audio Retrieval and Audio Captioning. As shown in Table 8, VAST have largely improved previous SOTA methods on three text-to-audio retrieval benchmarks, by 7.6, 5.4 and 9.8 R@1 points, respectively. and for audio captioning task, VAST achieves new SOTA performances on Clotho benchmark (both V1 and V2), and comparable performance on AudioCaps benchmark to WavCaps [70]. It is noted that WavCaps explored four model architectures with different audio encoder and text encoders targeting at different benchmarks, while VAST takes a unified architecture without targeted optimizations for specific downstream benchmarks.

Image-Text Benchmarks. We evaluate VAST on text-to-image retrieval, image captioning and image QA benchmarks. The results are presented in Table 10, from which we can find that even though VAST is designed as a omni-modality video-language understanding and generation model, it also shows strong capabilities on image-text benchmarks, demonstrating its generalization capabilities towards tasks of various modality types. Specifically, VAST achieves new SOTA performance on R@1 score of Flicker30K and R@5, R@10 scores of MSCOCO dataset, which outperforms image-text pretrained foundation models such as BLIP-2 [76] and BEiT-3 [74]. On COCO caption benchmark, VAST achieves 27.0 SPICE score which outperforms all previous methods such as OFA [72] and GIT2 [60]. On image QA benchmark, VAST achieves better performance than GIT [60], which is also a generative methods and predicts answers in a fully open way without any constraints.

B More about VAST-27M Dataset

B.1 Word cloud distribution

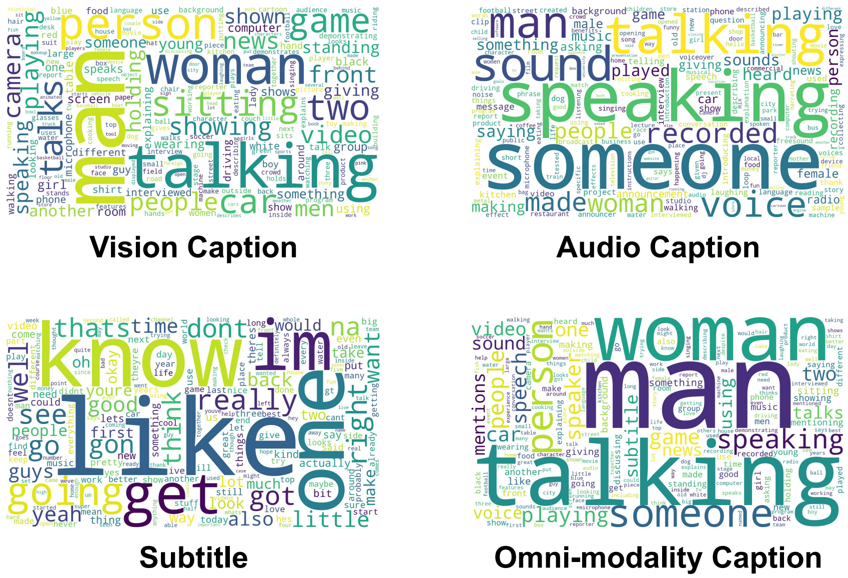

Figure 1: Word cloud map (Top-200) for vision, audio, omni-modality captions and raw subtitles of VAST-27M.

B.2 Prompts for Omni-Modality Caption Generation

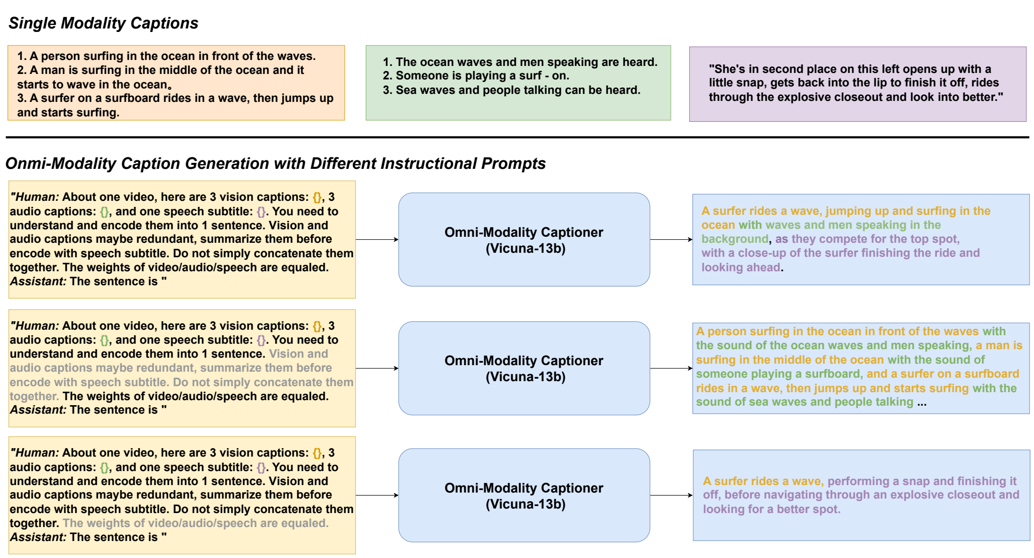

Figure 2: Ablation study for instructional prompt used for omni-modality video caption generation in VAST-27M.

B.3 More Examples

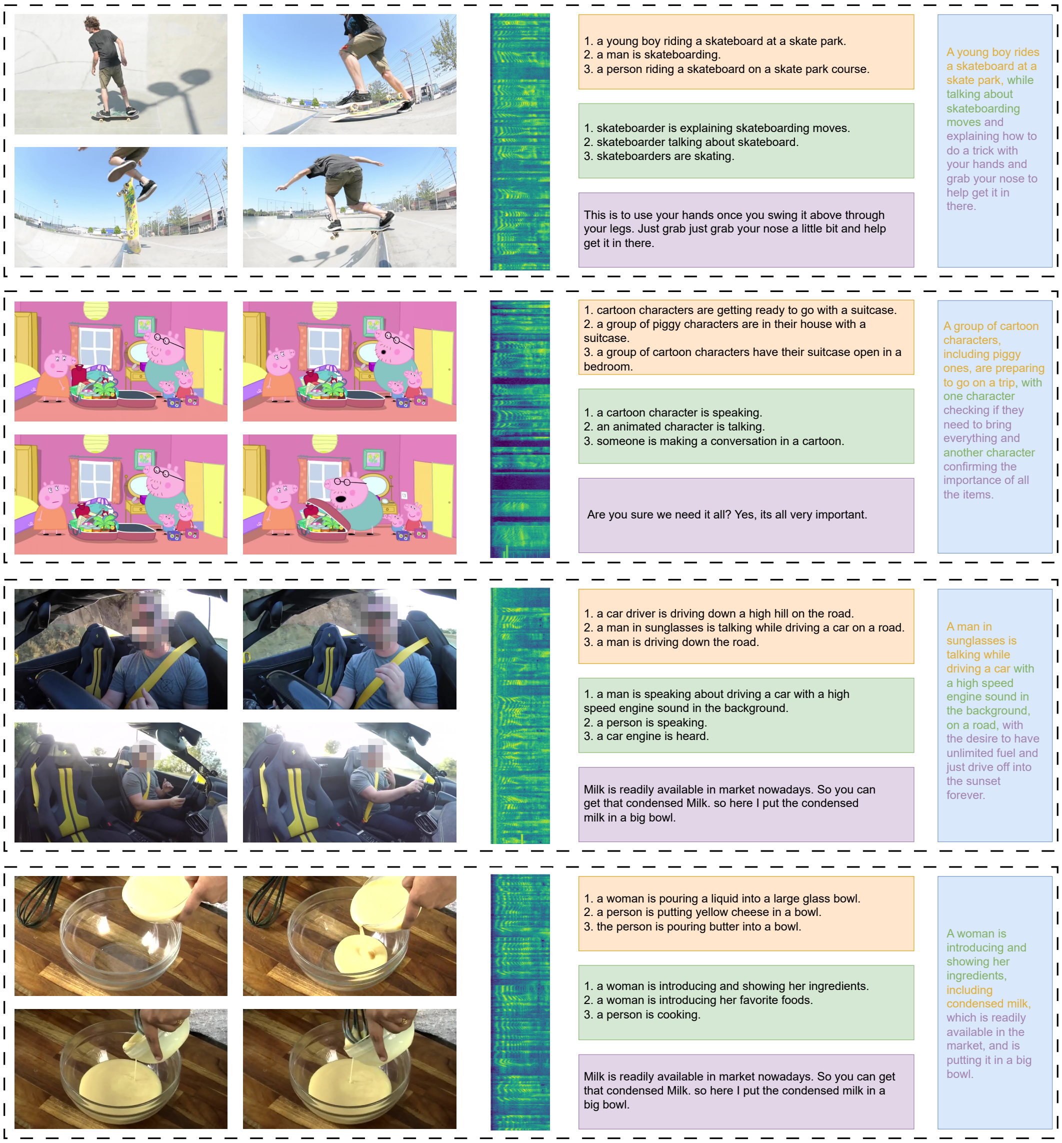

Figure 3: More samples in VAST-27M.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[2] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[3] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[5] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4581–4591.

[6] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[7] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 638–10 647.

[8] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *arXiv preprint arXiv:2304.08345*, 2023.

[10] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[11] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, pp. 190–200.

[12] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video question answering via gradually refined attention over appearance and motion," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1645–1653.

[13] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[14] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[15] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 706–715.

[16] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "Activitynet-qa: A dataset for understanding complex web videos via question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9127–9134.

[17] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, "Tgif-qa: Toward spatio-temporal reasoning in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2758–2766.

[18] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to answer questions in dynamic audio-visual scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 108–19 118.

[19] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 447–463.

[20] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[21] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[22] A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," *arXiv preprint arXiv:2105.02192*, 2021.

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[24] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[25] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[26] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[27] X. Cheng, H. Lin, X. Wu, F. Yang, and D. Shen, "Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss," *arXiv preprint arXiv:2109.04290*, 2021.

[28] J. Lei, T. L. Berg, and M. Bansal, "Revealing single frame bias for video-and-language learning," *arXiv preprint arXiv:2206.03428*, 2022.

[29] J. Wang, D. Chen, Z. Wu, C. Luo, L. Zhou, Y. Zhao, Y. Xie, C. Liu, Y.-G. Jiang, and L. Yuan, "Omnivl: One foundation model for image-language and video-language tasks," *arXiv preprint arXiv:2209.07526*, 2022.

[30] Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, and F. Huang, "Hitea: Hierarchical temporal-aware video-language pre-training," *arXiv preprint arXiv:2212.14546*, 2022.

[31] F. Cheng, X. Wang, J. Lei, D. Crandall, M. Bansal, and G. Bertasius, "Vindlu: A recipe for effective video-and-language pretraining," *arXiv preprint arXiv:2212.05051*, 2022.

[32] L. Li, Z. Gan, K. Lin, C.-C. Lin, Z. Liu, C. Liu, and L. Wang, "Lavender: Unifying video-language understanding as masked language modeling," *arXiv preprint arXiv:2206.07160*, 2022.

[33] A. J. Wang, Y. Ge, R. Yan, Y. Ge, X. Lin, G. Cai, J. Wu, Y. Shan, X. Qie, and M. Z. Shou, "All in one: Exploring unified video-language pre-training," *arXiv preprint arXiv:2203.07303*, 2022.

[34] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[35] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 638–647.

[36] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang *et al.*, "mplug-2: A modularized multi-modal foundation model across text, image and video," *arXiv preprint arXiv:2302.00402*, 2023.

[37] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," *arXiv preprint arXiv:2303.16058*, 2023.

[38] H. Xue, Y. Sun, B. Liu, J. Fu, R. Song, H. Li, and J. Luo, "Clip-vip: Adapting pre-trained image-text model to video-language representation alignment," *arXiv preprint arXiv:2209.06430*, 2022.

[39] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *European Conference on Computer Vision*. Springer, 2020, pp. 214–229.

[40] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," *arXiv preprint arXiv:2006.09199*, 2020.

[41] V. Gabeur, A. Nagrani, C. Sun, K. Alahari, and C. Schmid, "Masking modalities for cross-modal video retrieval," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1766–1775.

[42] Y.-B. Lin, J. Lei, M. Bansal, and G. Bertasius, "Eclipse: Efficient long-range video retrieval using sight and sound," *arXiv preprint arXiv:2204.02874*, 2022.

[43] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. Hauptmann, J. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," *arXiv preprint arXiv:2010.02824*, 2020.

[44] Q. Wang, Y. Zhang, Y. Zheng, P. Pan, and X.-S. Hua, "Disentangled representation learning for text-video retrieval," *arXiv preprint arXiv:2203.07111*, 2022.

[45] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1728–1738.

[46] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *arXiv preprint arXiv:2002.06353*, 2020.

[47] D. Ko, J. Choi, H. K. Choi, K.-W. On, B. Roh, and H. J. Kim, "Meltr: Meta loss transformer for learning to fine-tune video foundation models," *arXiv preprint arXiv:2303.13009*, 2023.

[48] H. Xu, G. Ghosh, P.-Y. Huang, P. Arora, M. Aminzadeh, C. Feichtenhofer, F. Metze, and L. Zettlemoyer, "Vlm: Task-agnostic video-language model pre-training for video understanding," *arXiv preprint arXiv:2105.09996*, 2021.

[49] L. Li, J. Lei, Z. Gan, L. Yu, Y.-C. Chen, R. Pillai, Y. Cheng, L. Zhou, X. E. Wang, W. Y. Wang *et al.*, "Value: A multi-task benchmark for video-and-language understanding evaluation," *arXiv preprint arXiv:2106.04632*, 2021.

[50] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[51] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "Violet: End-to-end video-language transformers with masked visual-token modeling," *arXiv preprint arXiv:2111.12681*, 2021.

[52] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[53] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7331–7341.

[54] T.-J. Fu, L. Li, Z. Gan, K. Lin, W. Y. Wang, L. Wang, and Z. Liu, "An empirical study of end-to-end video-language transformers with masked visual modeling," *arXiv preprint arXiv:2209.01540*, 2022.

[55] J. Huang, Y. Li, J. Feng, X. Sun, and R. Ji, "Clover: Towards a unified video-language alignment and fusion model," *arXiv preprint arXiv:2207.07885*, 2022.

[56] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[57] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "Merlot: Multimodal neural script knowledge models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 634–23 651, 2021.

[58] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-shot video question answering via frozen bidirectional language models," *arXiv preprint arXiv:2206.08155*, 2022.

[59] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang *et al.*, "Internvideo: General video foundation models via generative and discriminative learning," *arXiv preprint arXiv:2212.03191*, 2022.

[60] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "Git: A generative image-to-text transformer for vision and language," *arXiv preprint arXiv:2205.14100*, 2022.

[61] W. Kuo, A. Piergiovanni, D. Kim, X. Luo, B. Caine, W. Li, A. Ogale, L. Zhou, A. Dai, Z. Chen *et al.*, "Mammut: A simple architecture for joint learning for multimodal tasks," *arXiv preprint arXiv:2303.16839*, 2023.

[62] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[63] S. Yan, T. Zhu, Z. Wang, Y. Cao, M. Zhang, S. Ghosh, Y. Wu, and J. Yu, "Video-text modeling with zero-shot transfer from contrastive captioners," *arXiv preprint arXiv:2212.04979*, 2022.

[64] M. Tang, Z. Wang, Z. Zeng, F. Rao, and D. Li, "Clip4caption++: Multi-clip for video caption," *arXiv preprint arXiv:2110.05204*, 2021.

[65] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[66] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "Swinbert: End-to-end transformers with sparse attention for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[67] S. Chen, X. Zhu, D. Hao, W. Liu, J. Liu, Z. Zhao, L. Guo, and J. Liu, "Mm21 pre-training for video understanding challenge: Video captioning with pretraining techniques," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4853–4857.

[68] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid, "Learning audio-video modalities from image captions," *arXiv preprint arXiv:2204.00679*, 2022.

[69] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[70] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[71] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, "Investigating local and global information for automated audio captioning with transfer learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 905–909.

[72] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 318–23 340.

[73] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[74] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[75] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[76] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[77] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[78] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[79] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[80] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[81] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "Davit: Dual attention vision transformers," *arXiv preprint arXiv:2204.03645*, 2022.