Dataset Name	X Modality	#.X	#.T	#.X-T
ALIGN (Jia et al., 2021)	Image	1.8B	1.8B	1.8B
LTIP (Alayrac et al., 2022)	Image	312M	312M	312M
MS-COCO (Lin et al., 2014)	Image	124K	620K	620K
Visual Genome (Krishna et al., 2017)	Image	108K	4.5M	4.5M
CC3M (Sharma et al., 2018)	Image	3.3M	3.3M	3.3M
CC12M (Changpinyo et al., 2021)	Image	12.4M	12.4M	12.4M
SBU (Ordonez et al., 2011)	Image	1M	1M	1M
LAION-5B (Schuhmann et al., 2022)	Image	5.9B	5.9B	5.9B
LAION-400M (Schuhmann et al., 2021)	Image	400M	400M	400M
LAION-en (Schuhmann et al., 2022)	Image	2.3B	2.3B	2.3B
LAION-zh (Schuhmann et al., 2022)	Image	142M	142M	142M
LAION-COCO (Schuhmann et al., 2022b)	Image	600M	600M	600M
Flickr30k (Young et al., 2014)	Image	31K	158K	158K
AI Challenger Captions (Wu et al., 2017)	Image	300K	1.5M	1.5M
COYO (Byeon et al., 2022)	Image	747M	747M	747M
Wukong (Gu et al., 2022)	Image	101M	101M	101M
COCO Caption (Chen et al., 2015)	Image	164K	1M	1M
WebLI (Chen et al., 2022b)	Image	10B	12B	12B
Episodic WebLI (Chen et al., 2023h)	Image	400M	400M	400M
CC595k (Liu et al., 2023e)	Image	595K	595K	595K
RefCOCO (Kazemzadeh et al., 2014)	Image	20K	142K	142K
RefCOCO+ (Yu et al., 2016)	Image	20K	142K	142K
Visual-7W (Zhu et al., 2016)	Image	47.3K	328K	328K
OCR-VQA (Mishra et al., 2019)	Image	207K	1M	1M
ST-VQA (Biten et al., 2022)	Image	23K	32K	32K
DocVQA (Mathew et al., 2021)	Image	12K	50K	50K
TextVQA (Singh et al., 2019)	Image	28.4K	45.3K	45.3K
DataComp (Gadre et al., 2023)	Image	1.4B	1.4B	1.4B
GQA (Hudson and Manning, 2019)	Image	113K	22M	22M
VGQA (Krishna et al., 2017)	Image	108K	1.7M	1.7M
VQA^{v2} (Goyal et al., 2017)	Image	265K	1.4M	1.4M
DVQA (Kafle et al., 2018)	Image	300K	3.5M	3.5M
OK-VQA (Schwenk et al., 2022)	Image	14K	14K	14K
A-OKVQA (Schwenk et al., 2022)	Image	23.7K	24.9K	24.9K
Text Captions (Sidorov et al., 2020)	Image	28K	145K	145K
Multimodal Arxiv (Li et al., 2024b)	Image	32K	16.6K	16.6K
M3W (Interleaved) (Alayrac et al., 2022)	Image	185M	182GB	43.3M (Instances)
MMC4 (Interleaved) (Zhu et al., 2023c)	Image	571M	43B	101.2M (Instances)
Obelics (Interleaved) (Laurençon et al., 2023)	Image	353M	115M	141M (Instances)
MSRVTT (Xu et al., 2016)	Video	10K	200K	200K
WebVid (Bain et al., 2021)	Video	10M	10M	10M
VTP (Alayrac et al., 2022)	Video	27M	27M	27M
AISHELL-1 (Chen et al., 2023b)	Audio	–	–	128K
AISHELL-2 (Chen et al., 2023b)	Audio	–	–	1M
WaveCaps (Mei et al., 2023)	Audio	403K	403K	403K
VSDial-CN (Ni et al., 2024)	Image, Audio	120K (Image), 1.2M(Audio)	120K	1.2M

Table 3: The statistics for MM PT datasets. **#.X** represents the quantity of X, **#.T** represents the quantity of Text, and **#.X-T** represents the quantity of X-Text pairs, where X can be Image, Video, or Audio.