APPENDIX

A STATISTICS OF VIDAL-10M DATASET

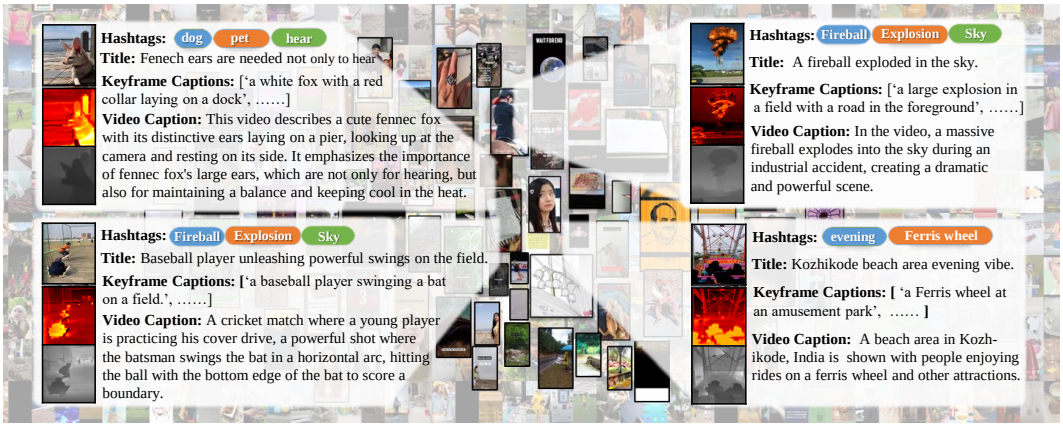

Figure 6: Examples of video-audio-text-depth-infrared pairs in VIDAL-10M, with the text components comprising hashtags, title, keyframe captions, and video caption. Examples are taken from 4 distinct clusters, corresponding to Sports, Pets & Animals, News & Politics, and Education.

In order to build a video dataset with rich visual concepts and diversity, we develop a unique but simple search term acquisition strategy. This strategy involves obtaining search terms from various visual datasets (as shown in Table 6). Subsequently, we use these search terms to gather videos from the YouTube Shorts platform, which has become a popular source for video data due to its abundance and diverse content. We collect videos in various categories, including sports, animals, nature, etc., resulting in a large and diverse dataset. Examples of video-audio-text-depth-infrared pairs in the VIDAL-10M dataset are shown in Figure 6. Moreover, to ensure data quality, we manually design a list of stop words that are filtered from our datasets. These words include terms such as "bts", "bmw", and "nfl", among others, that are not relevant to our research.

Video categories and duration Furthermore, we analyze the distribution of video categories with varying durations in our datasets, as illustrated in Figure 7. The normal distribution pattern observed in this analysis indicates that our dataset covers a wide range of concepts. Besides, we show the proportions of each category across different duration grades in the VIDAL-10M dataset in Figure 8.

Table 6: Examples of textual descriptions from various datasets as search terms.

Dataset	Search terms
YouTube-8M	How to make a delicious chocolate cake. Learn to dance salsa in 10 easy steps. ……
Howto100M	How to play chess. How to make pizza. ……
ImageNet	lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens, coon bear killer whale, killer, grampus, sea wolf, Orcinus orca, giant panda, panda, panda bear ……
COCO	A small boat floating on a body of water with a city skyline in the background. A man with a red helmet on a small moped on a dirt road. ……
Others	……

Table 7: Stop words in our datasets.

viral	funny	love	fashion	subscribe	nature
asmr	motivation	fitness	art	satisfying	foryou
music	india	fun	bts	amazing	edit
life	roblox	vlog	minecraft	design	marvel
explore	dubai	foryoupage	comedy	outfit	ootd
share	indian	lol	creative	relaxing	tattoo
random	instagram	quotes	workout	sad	ideas
views	bgmi	yummy	respect	easy	usa
ronaldo	jawellery	memes	happy	nfl	song
mlb	reel	support	nba	wow	status
gree	meme	gameplay	top	blackpink	whatsappstatus
follow	homedecor	history	tutorial	bodybuilding	japan
interiordesign	freefire	stunt	foodie	animation	recipe
skills	tips	crazy	pov	editing	aesthetic
style	view	london	reaction	story	pubg
construction	challenge	healthy	bmw	uk	free
hairstyle	enjoy	motivational	messi	capcut	nailart
entertainment	fifa	attitude	europe	health	geography
gta	unboxing	adventure	whatsapp	fail	btsarny
god	inspiration	relatable	comment	tattoos	fy
highlights	amazon	illustration	fortnite	ntb	avaiation
interior	decor	travelvlog	canada	btsarmy	tranding
time	mtb	luxury	vlogs	picsart	reels
photoshoot	business	photography	...	...	...

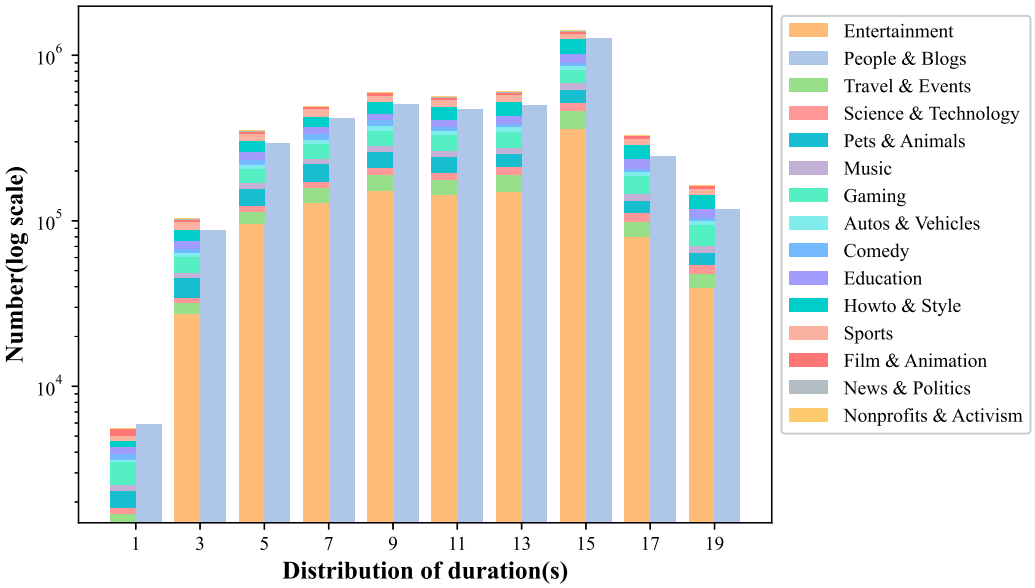

Figure 7: The number of 15 categories with different durations in our VIDAL-10M datasets. A wide range of concepts are covered.

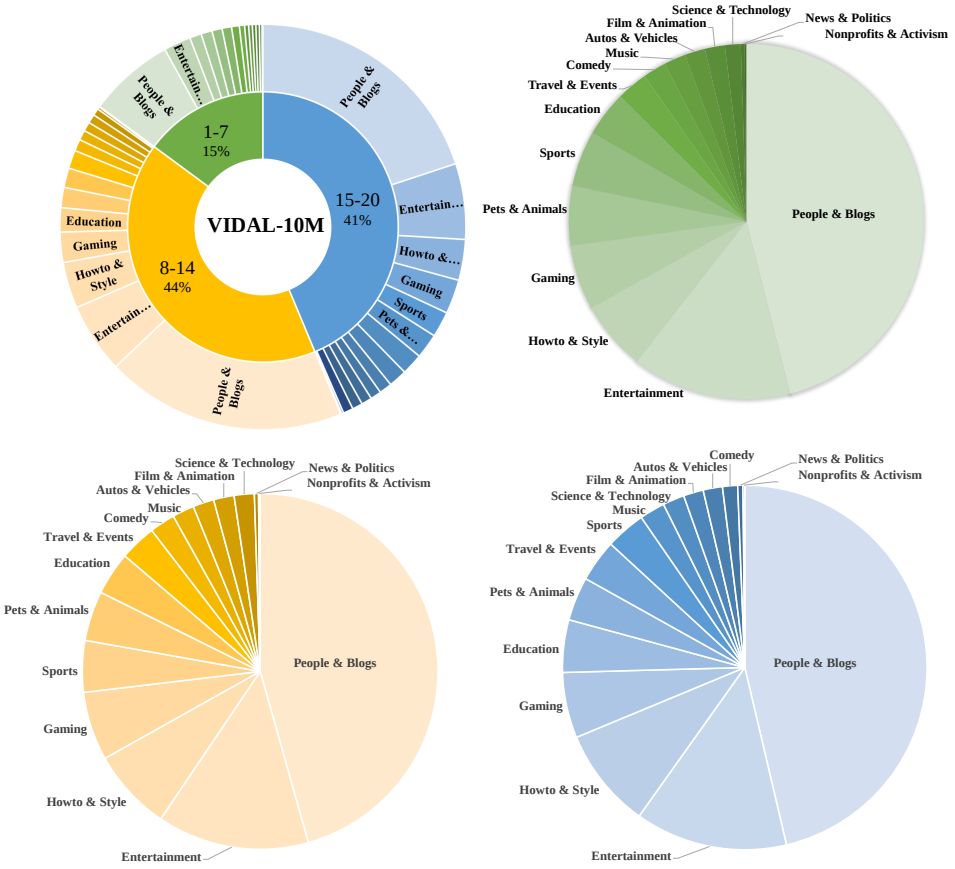

Figure 8: The statistical distribution of categories across the three duration grades in the VIDAL-10M dataset. The colors green, blue, and yellow represent video durations of 1-7, 8-14, and 15-20 s, respectively.

FPS, Aspect ratio and Resolution The first aspect examined in the dataset is the Frames Per Second (FPS) of the videos. FPS refers to the number of frames or images displayed per second in a video. The aspect ratio of a video represents the proportional relationship between its width and height dimensions. It is a critical factor in determining the visual presentation and viewing experience of the videos. The distribution of FPS and aspect ratios in Figure 9 provides insights into the smoothness and fluidity of the recorded content and sheds light on the various formats and orientations used. Video resolution refers to the number of pixels in each dimension that a video contains. It directly affects the clarity, sharpness, and level of detail in the visual content. Examining the distribution of resolutions (Figure 10) in the dataset provides an understanding of the available video quality and the technological capabilities of the recorded material.

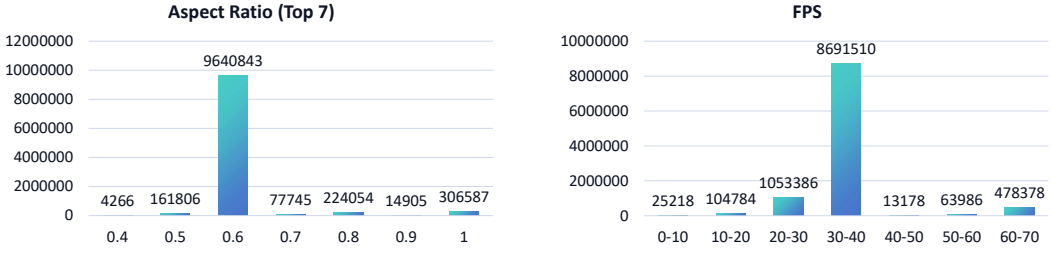

Figure 9: The distribution of FPS (Frames Per Second) and aspect ratio in the videos of the VIDAL-10M dataset.

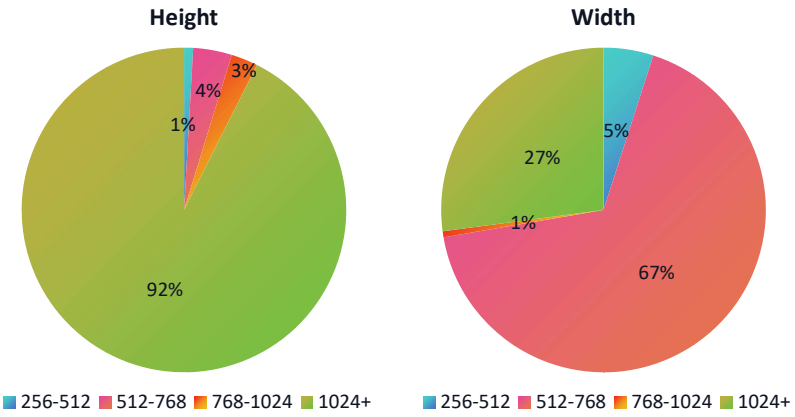

Figure 10: Height and width distribution of videos in VIDAL-10M dataset.

B PRETRAINING DETAILS

In this section, we introduce our training configuration.

Table 8: pretraining setting

Config	Video	Infrared	Depth	Audio
Vision encoder	ViT-Base/32	ViT-Huge/14	ViT-Huge/14	ViT-Huge/14
Optimizer	BertAdam	AdamW	AdamW	AdamW
Optimizer Momentum		$\beta_1, \beta_2 = 0.9, 0.98$		
Epochs	1	1	1	8
Learning rate	1e-4	1e-4	5e-4	5e-4
Coefficient learning rate		1e-3		
Weight decay		0.2		
Batch size	512	1024	1024	512
Warmup steps	-	200	200	200
Temperature		learnable		
Learning rate schedule		cosine decay		
Slice frame position	cut from head frames	-	-	-
Max words	32	77	77	77
Max frames	12	-	-	-
Cls header	mean-pooling	-	-	-
LoRA rank	-	2	2	8
LoRA alpha	-	16	16	16
LoRA dropout	-	0.1	0.1	0.1

Video-Language. For the video-text retrieval, we verify that the *VIDAL-10M* dataset is highly aligned. We adopted the training framework of CLIP4Clip, and the model is initialized from ViT-B/32, and the rest of the parameters are the same as the default settings, except for 1 epoch and batch size of 512.

Depth-Language. The model is initialized from OpenCLIP with a frozen language encoder. For each individual sample, we employ a random selection approach to extract either a depth image from the video sequence. Subsequently, we resize these frames to have a short edge length of 256 units, followed by a central cropping process to attain dimensions of 224×224. Additionally, we tripled the number of channels in both the depth image. The text templates employed for zero-shot classification are sourced from OpenCLIP, with a modification consisting of the substitution of ”photo” with ”depth photo” across all templates. This alteration yields an approximate performance gain of 1%.

Infrared-Language. Following depth-language, it is worth noting that the text templates corresponding to infrared images retain the "photo" designation, as no discernible performance improvement is observed from this particular modification.

Audio-Language. The data are preprocessed as in 3.1. Unlike depth and infrared, spectrograms differ much from the domain of conventional visual images. Therefore, it is not easy to overfit during training, so we increase the training epoch and the rank of LoRA. Additionally, we replace "the/a photo of" with "the/a sound of" across all templates for audio zero-shot classification.

C DOWNSTREAM DATASETS

Video-language. We perform video-text retrieval experiments on 2 datasets. **(a) MSR-VTT** (Xu et al., 2016) comprises 10K YouTube videos, each paired by 200K captions. In our analysis, we present results based on the 1K-A test subset. **(b) MSVD** (Chen & Dolan, 2011) consists of about 120K sentences and reports results on test data (670 samples).

Infrared-language. **(a) LLVIP** (Jia et al., 2021) constitutes a dataset for pedestrian object detection within the infrared spectrum. Following ImageBind, we extracted all people from the images, designating all other objects as background elements. This process resulted in a dataset comprising 7,622 'background' classes and 7,954 'person' classes, which was subsequently employed for binary classification testing. **(b) FLIR v1** (Teledyne FLIR, 2015a) offers comprehensive annotations for both thermal and visible spectrum frames. From the test data, we derived a dataset containing 11,696 images by extracting bounding boxes. This dataset encompasses 4 categories – ['bicycle', 'car', 'dog', 'person']. **(c) FLIR v2** (Teledyne FLIR, 2015b) includes 16,696 images after processing similarly, which were categorized into 12 classes – ['bike', 'bus', 'car', 'hydrant', 'light', 'motor', 'other vehicle', 'person', 'sign', 'skateboard', 'stroller', 'truck'].

Depth-language. We use **NYU-v2 Depth-only (NYU-D)** (Silberman et al., 2012) to validate by 654 test samples. Through preprocessing, we constrained the depth images to a maximum depth of 10 meters. Following ImageBind, we undertook a category reorganization process, resulting in a total of 10 scene categories.

D LICENSE

Unless explicitly noted otherwise, our released datasets are provided to users under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License ("CC BY-NC-SA 4.0"), in conjunction with the additional terms outlined herein. The CC BY-NC-SA 4.0 license can be accessed at `https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode`. By downloading or utilizing our datasets from our website or other sources, you agree to adhere to the terms of CC BY-NC-SA 4.0, as well as the terms outlined in our dataset Terms. In the event of any conflict between the terms of CC BY-NC-SA 4.0 and our dataset Terms, the latter shall prevail. We once again emphasize that this dataset is exclusively intended for non-commercial purposes, such as academic research, teaching, or scientific publications. We strictly prohibit any commercial use of the dataset or any derived works, including the sale of data or utilization of data for commercial gain.