UNIALIGN: Scaling Multimodal Alignment within One Unified Model

Supplementary Material

Appendix

The appendix is structured as follows:

- §1 discusses the limitation and social impact of our method.
- §2 provides additional details on the architecture of the proposed model.
- §3 presents more experimental details.
- §4 supplements failure cases analysis and presents more qualitative results.

1. Limitation and Social Impact

1.1. Limitation

Although UNIALIGN does not achieve SOTA performance across all benchmarks, it possesses a significant strength: the utilization of a single encoder for multimodal alignment within one training phase, which substantially reduces model complexity. This strategy not only alleviates the computational burden but also promotes cross-modal learning through joint training, making it more efficient than training separate models. While our dependence on semantic similarity between labels may appear restrictive, it effectively harnesses available data to align modalities without requiring extensive paired datasets, showcasing UNIALIGN's innovative use of soft bindings.

1.2. Social Impact

In an age where training a foundation model often demands dozens of GPUs, UNIALIGN provides a new paradigm that enables the expansion of a foundational model into multiple domains with high efficiency. By achieving 90% of the performance of current methods using just four GPUs, our model not only maintains a compact structure but also makes cutting-edge AI technology more accessible and sustainable. This advancement ensures that a wide range of fields can benefit from advanced multimodal capabilities, driving innovation while reducing the environmental impact of AI development.

2. Architecture

2.1. Modality Tokenizer

Image and Video. We employ the conventional input format established in the original Vision Transformer (ViT) architecture. In this framework, images are divided into non-overlapping patches of size $P \times P$. For video input, we follow [7, 24] to partition video input into two-frame clips, and construct spatio-temporal patches of size $T \times P \times P$. The image tokenizer's weights are inflated to three dimensions to handle video inputs, which are then projected into spatio-temporal embeddings. Temporal position embeddings are added to the original two-dimensional position embeddings.
3D Point Cloud. For 3D point clouds, we sample 8,192 points, grouping them into subclouds using Farthest Point Sampling (FPS) and k-Nearest Neighbors (KNN) for neighboring points. We construct local patches by sampling 512 sub-clouds, each comprising 32 points. Further, we employ a mini-PointNet [12] to project these sub-clouds into point embeddings. Additionally, we incorporate learnable positional embeddings on top of these embeddings, serving as inputs to unified models.
Depth. Single-view depth data is converted into disparity. The depth is treated as a one-channel image, with image and depth channels separately converted into patches. To capture positional information, we incorporate learnable positional embeddings.
Audio. Audio inputs are sampled at 16 kHz, and a Mel spectrogram with 128 frequency bins is extracted using a 25 ms Hamming window with a 10 ms hop length. The spectrogram, being a two-dimensional signal, is transformed into an audio embedding using one-dimensional convolution. Subsequently, we introduce learnable positional embeddings to capture the spatial structure of the spectrogram.

2.2. Feature Projector

Since the features extracted by the unified modality encoder cannot directly match the heterogeneous features of all partial-modal foundation models, we add a feature projector composed of multiple learnable linear layers after the final layer of the student model. This allows the unified features to effectively align with the features of different foundation models.

3. More Experimental Details and Results

3.1. Datasets

ImageNet. ImageNet1K [4] is one of the most widely used datasets for image classification tasks. It contains over 1.2 million images categorized into 1,000 distinct classes, covering a wide array of objects and scenes. The dataset is well-known for its role in benchmarking the performance of various deep learning models in computer vision. Each class includes hundreds to thousands of images, providing a rich resource for training and evaluating algorithms. In this study, we utilize ImageNet1K to explore transfer learning and evaluate model performance on image classification.
ULIP-ShapeNet Triplets. The ULIP-ShapeNet Triplets is

sourced from ShapeNet55 [2] as detailed in the work of [22]. Each instance of the 3D point cloud is constructed from CAD models. To create anchor images, virtual cameras are strategically placed around each object. Additionally, textual information is generated by integrating metadata into a specific prompt template. This dataset covers around 52.5k individual 3D point cloud instances.

ModelNet40. The ModelNet40 dataset [20] is a significant benchmark in the field of 3D object classification. It contains a total of 12,311 CAD models organized into 40 separate categories, with 9,843 samples allocated for training purposes and 2,468 for testing. This dataset encompasses a range of everyday items, including chairs, tables, desks, and various household objects. Each item is depicted as a 3D point cloud and includes manual annotations that identify its category. In this study, our primary focus is on leveraging the test samples for the purpose of zero-shot classification.

ScanObjectNN. The ScanObjectNN dataset [18] is a vital asset in the realm of 3D object recognition and segmentation. It comprises a wide range of 3D object instances captured using a standard RGB-D camera. This dataset features various household items, furniture, and typical indoor objects. Each object instance is equipped with comprehensive semantic and instance-level annotations, enhancing the dataset's utility. Overall, it contains 2,902 objects distributed among 15 unique categories. In this study, we leverage the variant outlined by [23] for zero-shot classification, following the methodology proposed in [10], which includes 581 test shapes across these 15 categories.

SUN-RGBD. We leverage paired RGB and depth images, along with the corresponding class labels, sourced from the SUN-RGBD dataset [15]. In our training process, we utilize the training set, which consists of approximately 5,000 samples. To evaluate the classification performance, we use the test set referred to as SUN Depth-only, which contains 4,660 samples. During testing, we focus exclusively on depth data as input and develop classification templates based on the 19 scene categories present in the dataset.

NYU-Depth v2. We utilize the depth maps from the NYU-Depth v2 Depth-only test set [14], which contains 654 samples for our evaluation. The dataset consists of 16 semantic classes, and we build on the approach detailed in previous studies [7, 9] to implement a 10-class classification scheme. Notably, the "others" category aggregates seven different semantic classes: ['computer room', 'study', 'playroom', 'office kitchen', 'reception room', 'lobby', 'study space']. For classification purposes, we calculate the similarity for the "others" class by identifying the maximum cosine similarity among these seven class names.

Audioset. For our study, we leverage the Audioset dataset [5] for both training and assessment. This dataset is composed of 10-second video clips collected from YouTube and is annotated with 527 unique classes. It is organized into three predefined splits: an unbalanced training set with around 2M videos, a balanced training set containing approximately 20k videos, and a test set with about 18k videos. Due to the unavailability of some videos for download, we finally have 0.5M/18k/17k for these three splits. We utilize the training splits to develop our model, while the test split is reserved for evaluation. During the evaluation process, as well as when using textual data as anchor input in training, we incorporate textual class labels along with associated templates.

ESC 5-folds. The ESC50 dataset [5] is a well-established benchmark in the area of environmental sound classification. It includes a collection of 2,000 sound recordings, organized into 50 distinct categories that encompass various types of sounds, such as animal calls, natural soundscapes, and sounds produced by humans. Each category contains 40 audio samples, each lasting five seconds. The dataset is designed with a predefined 5-fold evaluation scheme, where each fold consists of 400 test audio clips. In this study, we focus on evaluating zero-shot predictions across the five folds.

AudioCaps. This dataset [8] features audio-visual clips obtained from YouTube, paired with textual descriptions. The clips are sourced from the Audioset dataset. For this study, we followed the dataset splits described in prior research [11], specifically removing clips that overlap with the VGGSound dataset. Consequently, we have a total of 813 clips in the test split designated for zero-shot evaluation.

VGGSound. This audio-visual dataset [3] is compiled from YouTube and includes roughly 200,000 video clips, each with a duration of about 10 seconds. The clips are classified into 309 different categories, which range from human activities to sounds made by objects and interactions between people and objects. In this study, we use the audio and video from the training set for joint training.

MSRVTT. The MSR-VTT dataset [21] is a comprehensive resource designed for open-domain video captioning, comprising 10,000 video clips categorized into 20 different themes. Each clip is accompanied by 20 English sentences, resulting in approximately 29,000 distinct words across all captions. The dataset is typically divided into three parts: 6,513 clips are allocated for training, 497 for validation, and 2,990 for testing. In line with previous research [9], we present our results based on the 1K-A test set.

UCF101. The UCF101 dataset [16] is a widely recognized benchmark for action recognition tasks. It is an expanded version of the UCF50 dataset, comprising 13,320 video clips categorized into 101 distinct classes. These classes are organized into five main groups: body movements, interactions between individuals, interactions involving objects, musical instrument performances, and various sports. All video frames are captured at a resolution of 320×240 pixels and a frame rate of 25 frames per second, with clips sourced from YouTube. In this study, we evaluate zero-shot video classi-

	Image	*Video*	*3D Point Cloud*	*Depth*	*Audio*
Optimizer			AdamW		
Optimizer momentum			$\beta_1 = 0.9, \beta_2 = 0.98$		
Peak LR			2e-4		
Weight decay			0.05$^\diamond$		
Batch size			2048*		
Warmup steps			10,000		
Total epochs			100		
Modality augmentation	RandResizeCrop(size=224) RandHorizontalFlip(p=0.5) RandAugment(m=9,n=2) ColorJitter(0.4) RandErasing(p=0.25)	RandShortSideScale(min=256, max=340) RandCrop(size=224) RandHorizontalFlip(p=0.5) RandAugment(m=9,n=2,p=0.3) RandErasing(p=0.25)	RandDropout RandScale RandShift RandPerturb RandRotate	RandResizeCrop(size=224) RandHorizontalFlip(p=0.5) RandAugment(m=9, n=2) RandErasing(p=0.25)	Frequency masking(12) Time masking(48) NoiseAug
Image augmentation	-	-	RandResizeCrop(size=224)	RandResizeCrop(size=224) RandHorizontalFlip(p=0.5) RandAugment(m=9,n=2) ColorJitter(0.4) RandErasing(p=0.25)	RandShortSideScale(min=256, max=340) RandCrop(size=224) RandHorizontalFlip(p=0.5) RandAugment(m=9, n=2, p=0.3)

Table 1. Training hyper-parameters for each modality. * The total batch size is the sum of the mini-batches of all modalities involved in the weighted sampling. $^\diamond$ Weight decay excludes parameters for BatchNorm, LayerNorm, bias terms, and logit scale.

fication on the validation split. Subsequently, we fine-tune the pretrained model on the training split and evaluate the supervised results.

3.2. Data Input and Augmentation

Image and Video. Building on previous research [6, 17], we utilize a resolution of 224×224 and apply standard augmentations to both images and videos. For video input, we sample 8 frames at stride 8. To ensure consistency in modal knowledge distillation from various partial modal foundation models, we use the same random seed for data augmentation as that of the foundation model. Additionally, we disable Mixup and Cutmix, following the guidelines in [19].

3D Point Cloud. As discussed in 2.1, we uniformly sample 8,192 points from the 3D shape input and construct local patches using the Farthest Point Sampling (FPS) and k-Nearest Neighbors (kNN) algorithms. During training, we apply standard augmentation to the point clouds, as described in [22], ensuring that the random seed is consistent with that used in the partial modal foundation model.

Depth. As mentioned in 2.1, the depth maps are converted into disparity maps. Following the methodology outlined by [9], we apply strong augmentation to the depth data, maintaining the same random seed as used in the partial modal foundation model.

Audio. In accordance with [9], we sample a 5-second audio clip and apply spectrogram masking during training. The maximum time mask length is set to 48 frames, and the maximum frequency mask length is set to 12 bins. Unlike [9], we maintain the same random seed as the partial modal foundation model and disable the Mixup technique.

3.3. Training Setup

In Table 1, we list the hyperparameters used in joint training. Our experiments were done on 40 GB Tesla A100 GPU clusters. Our base network is initialized based on the CLIP model.

3.4. Additional Ablations

Feature Projector. We conducted ablation studies to evaluate the effectiveness of different projection methods used after the unified modality encoder. As shown in Table 2, the linear projection achieved an accuracy of 45.2% on the SUN-D dataset and 55.3% on ModelNet40, while the MLP projection slightly outperformed it with 45.8% and 55.9%, respectively. These results indicate that the feature projector plays a crucial role in aligning unified features with the heterogeneous features of various foundation models, enhancing overall model performance.

#	Proj Head	SUN-D	ModelNet40
1	Linear	45.2	55.3
2	**MLP**	**45.8**	**55.9**

Table 2. Ablation studies on feature projector.

Method	Teacher Modality	Teacher Model	SUN-D Top-1	ModelNet40 Top-1
Baseline	-	-	42.3	40.9
+ depth teacher	depth	MultiMAE*	43.3	41.1
+ depth teacher	depth	ViT-LENS$_L$	45.1	41.3
+ 3d teacher	3D	RECON	42.6	52.7
+ 3d teacher	3D	ULIP	42.9	55.1
+ depth teacher & + 3d teacher	depth + 3D	ViT-LENS$_L$ + ULIP	45.4	55.6

Table 3. Ablation studies about knowledge distillation from partial modal foundation models. *denotes our implementation of training MultiMAE on SUN RGB-D.

Knowledge Distillation from Partial-Modal Foundation Models. In our ablation study, detailed in Table 3, we explored the effects of knowledge distillation using different partial modal foundation models. Our baseline model, which includes depth experts and 3D experts, was trained without knowledge distillation from partial modal foundation models. Based on this baseline, we developed five variants: two single-teacher distilled models and one multi-teacher distilled model. Using the MultiMAE [1] depth teacher showed improvements, with further gains from ViT-LENS$_L$ [9]. For

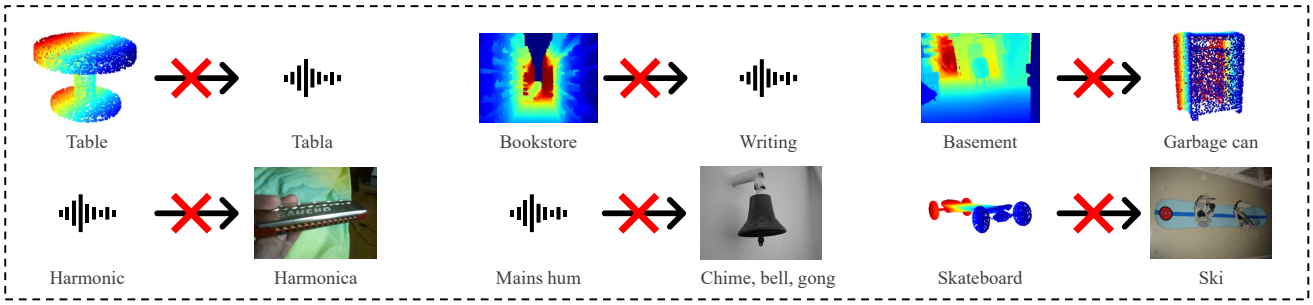

Figure 1. **Failure cases for cross-modal retrieval** (§4) **.**

3D teachers, RECON [13] and ULIP [22] both enhanced performance. The combination of ViT-LENS$_L$ and ULIP achieved the best overall results. The performance difference between different teacher models is mainly due to their different training methods. MultiMAE and RECON involve masked self-supervised learning, resulting in features that cannot be effectively processed by a simple feature projector. These findings highlight the benefits of using multiple partial modal foundation models in multimodal knowledge distillation.

4. Failure Cases and More Qualitative Results

As illustrated in Fig. 1 , several scenarios can lead to failure cases in cross-modal retrieval. First, when the category labels of different modalities are overly similar yet semantically distinct, confusion can arise during the retrieval process. Additionally, significant discrepancies between the category labels of different modalities may result in the inability to find semantically similar matching labels, leading to errors. Finally, if the retrieval task requires fine-grained recognition capabilities, candidates with similar appearances but different semantics can mislead the process. Nevertheless, our method can still produce accurate predictions in most scenarios, as shown in Fig. 2 .

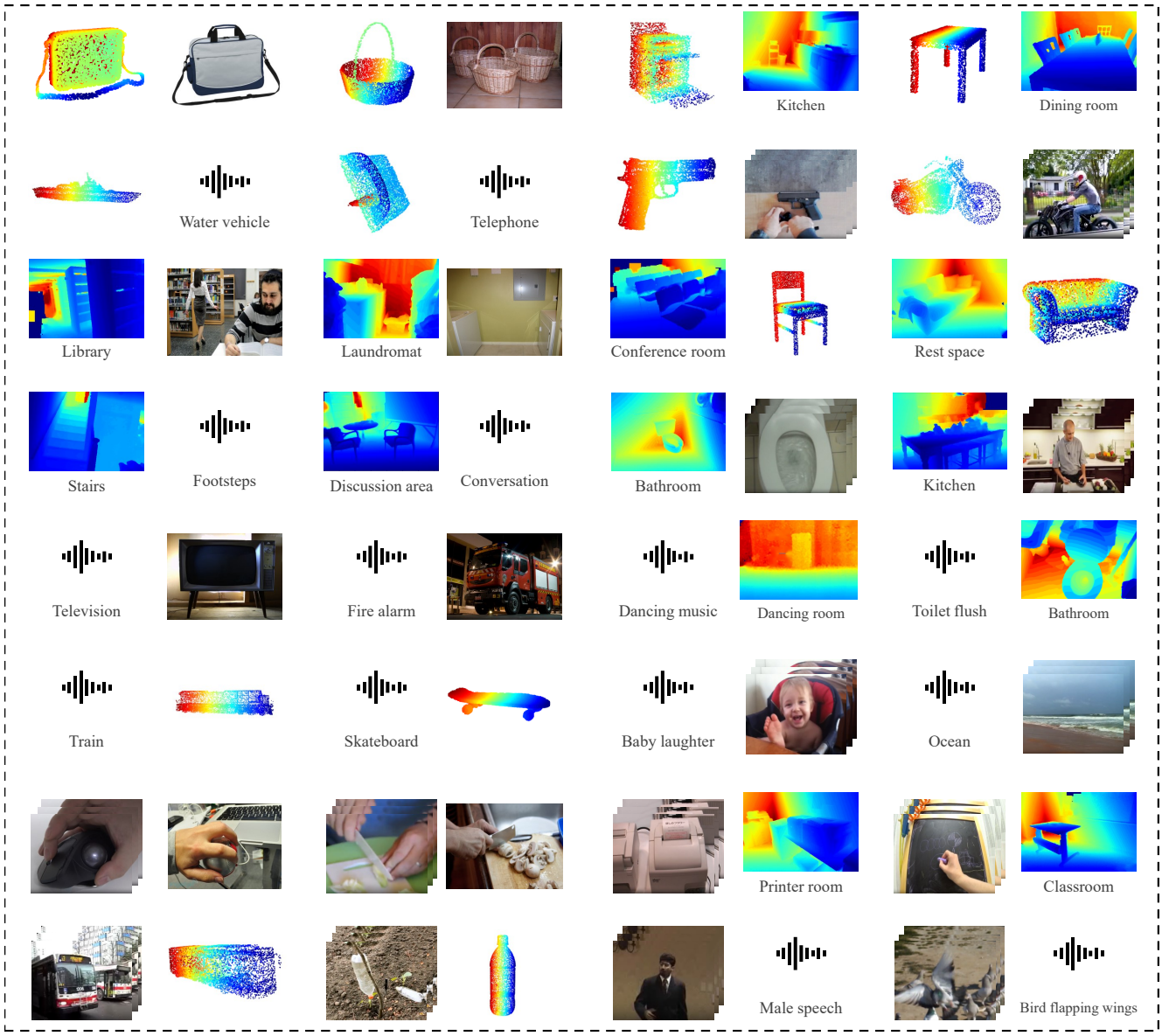

Figure 2. **More Qualitative results for cross-modal retrieval** (§4) **.**

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal Multi-task Masked Autoencoders. In *ECCV*, 2022. 3

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 2

[6] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A Single Model for Many Visual Modalities. In *CVPR*, 2022. 3

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind One Embedding Space to Bind Them All. In *CVPR*, 2023. 1, 2

[8] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 2

[9] Stan Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-Lens: Towards Omni-modal Representations. In *CVPR*, 2024. 2, 3

[10] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling Up 3d Shape Representation Towards Open-World Understanding. In *NeurIPS*, 2023. 2

[11] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 2

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 1

[13] Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with Reconstruct: Contrastive 3d Representation Learning Guided by Generative Pretraining. In *ICML*, 2023. 4

[14] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2

[15] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 2

[16] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[17] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 3

[18] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 2

[19] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast Pretraining Distillation for Small Vision Transformers. In *ECCV*, 2022. 3

[20] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2

[21] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2

[22] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a Unified Representation of Language, Images, and Point Clouds for 3d Understanding. In *CVPR*, 2023. 2, 3, 4

[23] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 2

[24] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 1