# UniBind: LLM-Augmented Unified and Balanced Representation Space to Bind Them All
# –Supplementary Material–

Yuanhuiyi Lyu[1] * Xu Zheng[1] * Jizhou Zhou[1] Lin Wang[1,2] †
[1]AI Thrust, HKUST(GZ) [2]Dept. of CSE, HKUST
{yuanhuiyilv, jiazhouzhou}@hkust-gz.edu.cn, zhengxu128@gmail.com, linwang@ust.hk
Project Page: https://vlislab22.github.io/UniBind/

## A. Construction of Knowledge Base

The knowledge base consists of two components: **1)** Category descriptions, generated by large language models (LLMs). **2)** Multi-modal data descriptions produced by multi-modal large language models (multi-modal LLMs).

Illustrating with ImageNet-1k [2] as an example, we initially generate descriptive texts for 1,000 categories using GPT-4 [11] and LLaMA [17]. For each category, we generate 1,000 descriptive texts, limiting the output sequence's maximum length to 77 tokens. Specific instances are detailed in Sec.A.1. Subsequently, we generate descriptions for the visual data. In the case of ImageNet-1k[2], we generate descriptions for each image in the dataset using BLIP-2 [9], ensuring the sequence length remains below 77 tokens. Concrete examples are provided in Sec.A.2. Finally, the organizational structure of our knowledge base is delineated in Sec.A.3.

### A.1. Cases of Generation via LLMs

We employ GPT-4 [11] and LLaMA [17] to generate category descriptions. Illustrated in Fig 1, we generate 1,000 descriptions for the category `[water snake]`. To achieve this, we utilize the prompt `["Please generate 1,000 sentences related to this sentence <A photo of a water snake>"]` as input, facilitating the generation of effective descriptions for the localization of the embedding center.

### A.2. Cases of Generation via Multi-modal LLMs

For the image, event, and thermal modalities, we produce multi-modal data descriptions using BLIP-2 [9]. Specifically, we employ paired RGB images from event and thermal data to generate these descriptions. The process of generating descriptions for image data is illustrated in Fig. 2.

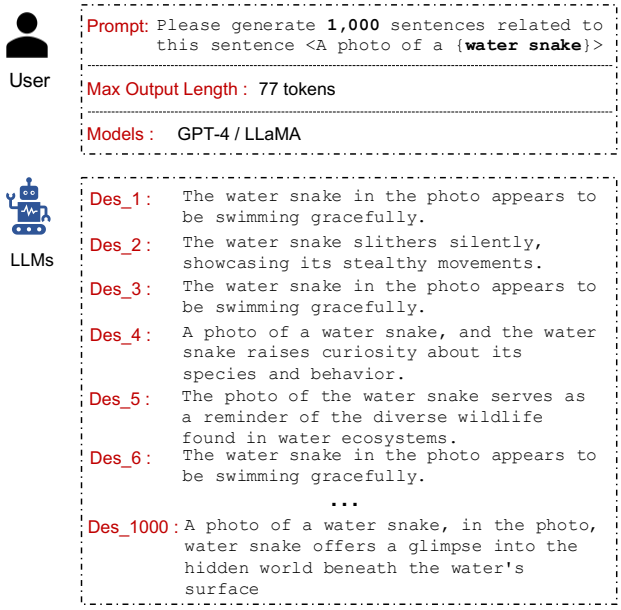


Figure 1. The case of description generation via LLMs. We show the generated descriptions for the category `[water snake]`.

We utilize the prompt `[Generate a detailed description of this photo]` as the text input, while the visual inputs consist of RGB images.

For the audio, video, and point cloud modalities, we employ the LLaMA-adapter [20] to generate descriptions for multi-modal data. Illustrated in Fig. 3, we present a case of generating descriptions for point cloud data. In this instance, we utilize the prompt `[Generate a detailed description of this 3D object]` as the text input, with the visual inputs consisting of point cloud data.

### A.3. Organizational Structure

Lastly, we present the organizational structure of our knowledge base in Fig. 4. The knowledge base arranges descrip-

*Equal Contribution.
†Corresponding author.

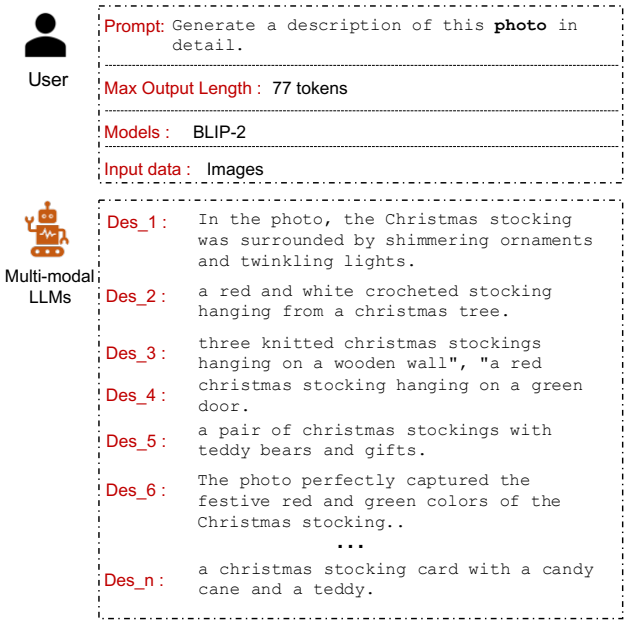


Figure 2. The case of generation via BLIP-2 [9]. We present the generated descriptions for ImageNet-1k [2].

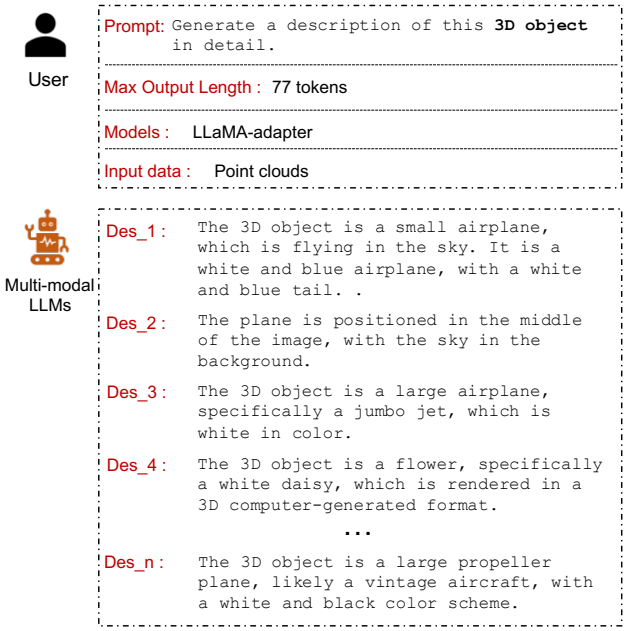


Figure 3. The case of generation via LLaMA-adapter [20]. We show the generation descriptions for ModelNet-40 [18]

tions generated from the same dataset in a table, with each table featuring four keys: *ID*, *Category*, *Description*, and *Source*. Descriptions with the same *Category* key value are selected to localize embedding centers for categories. During training, we retrieve paired descriptions of input visual data using the *ID* key.

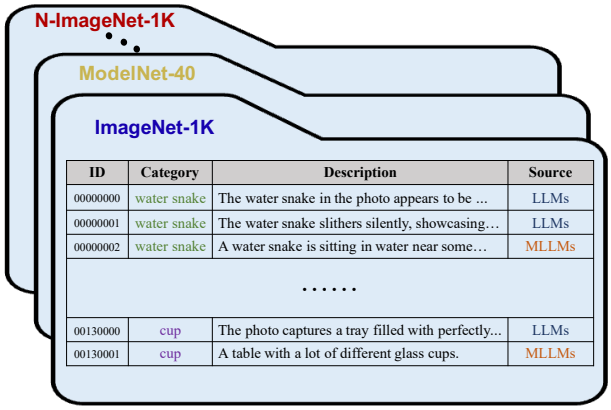


Figure 4. The organizational structure of our knowledge base.

# B. Implementation Details

## B.1. Datasets and Metrics

We experiment with our method on 13 datasets. Next, we show the details and metrics of these benchmark datasets.

**ImageNet-1K (IN-1K)** [2]. It is a standard image dataset designed for recognition tasks encompassing 1,000 categories. It serves the dual purpose of both training and evaluation. In the zero-shot setting, we assess both baseline models and our proposed method on the test set without training. Accuracy is employed as the metric for evaluation.

**Places-Stanford-365 (P365)** [10]. The Stanford-365 dataset is designed for scene recognition, comprising 365 categories. The evaluation setting for this dataset mirrors that of ImageNet-1K [2].

**Caltech-101 (cal)** [3]. The Caltech-101 dataset is a well-established benchmark dataset within the domain of computer vision, tailored specifically for object recognition tasks. Comprising images from 101 distinct object categories, it captures diverse scenes from real-world settings. In this study, we employ the dataset for evaluating models in the context of object recognition. Additionally, we utilize both the Caltech-101 and N-Caltech-101 datasets for cross-modal retrieval tasks. Accuracy is employed as the metric for the recognition task, while recall serves as the metric for the cross-modal retrieval task.

**ModelNet-40 (ModelNet40)** [18]. ModelNet-40 serves as a widely adopted benchmark dataset within the realm of 3D object recognition. Encompassing 40 object categories, it includes items such as chairs, tables, airplanes, cars, and various household objects. Each category is adequately represented with a substantial number of instances, ensuring a comprehensive and representative sample. In the evaluation of recognition, accuracy is employed as the metric.

**ShapeNet-part (ShapeNet)** [1]. ShapeNet-part stands as a prominent benchmark dataset extensively employed for 3D

| Modalities | Dataset | batch size | lr | total epochs |
|---|---|---|---|---|
| Image | ImageNet-1K (IN-1K) [2] | 1,024 | 5e-4 | 15 |
| | Places-Stanford-365 (P365) [10] | 1,024 | 5e-5 | 20 |
| | Caltech-101 (cal) [3] | 128 | 1e-4 | 20 |
| PointCloud | ModelNet-40 (ModelNet40) [18] | 128 | 5e-5 | 10 |
| | ShapeNet-part (ShapeNet) [1] | 128 | 5e-5 | 10 |
| Audio | ESC 5-folds (ESC) [13] | 64 | 1e-4 | 10 |
| | Urban-Sound-8K (Urban-S) [15] | 64 | 1e-4 | 10 |
| Thermal | LLVIP (LLVIP) [7] | 64 | 1e-3 | 20 |
| | RGB-T Selected (RGB-T) [6] | 64 | 5e-3 | 20 |
| Video | MSR-VTT (MSR-VTT) [19] | 128 | 5e-4 | 20 |
| | UFC-101 (UFC-101) [16] | 128 | 5e-4 | 20 |
| Event | N-Caltech-101 (N-cal) [12] | 128 | 1e-4 | 20 |
| | N-ImageNet-1K (N-IN-1K) [8] | 1,024 | 5e-4 | 15 |

Table 1. The hyperparameters of experiments with PointBind [5].

segmentation tasks. Within the scope of this work, we delineate the evaluation task on ShapeNet-part as a recognition task. The dataset comprises 16 categories of 3D objects. The evaluation metric employed for this task is accuracy.

**ESC 5-folds (ESC)** [13]. The dataset comprises 2,000 audio clips of 5 seconds each, classified into 50 distinct classes. In the zero-shot setting, we employ the entire audio dataset to assess both baseline models and our proposed method. Conversely, in the fine-tuning setting, models are trained exclusively on the training set and subsequently evaluated on the test set. The metric employed for evaluation in both settings is accuracy.

**Urban-Sound-8K (Urban-S)** [15]. The UrbanSound8K dataset is a widely used collection of audio data designed for research in the field of urban sound recognition. UrbanSound8K consists of 8,732 audio clips, each lasting 4 seconds. These clips are extracted from longer field recordings and are labeled with specific sound classes. The dataset is annotated with 10 sound classes and we evaluate models on the test set with accuracy metric.

**LLVIP (LLVIP)** [7]. The LLVIP dataset consists of RGB image and Thermal image pairs. We follow ImageBind [4] to process it for a binary classification task. We crop out pedestrian bounding boxes and random bounding boxes (same aspect ratio and size as a pedestrian) to create a balanced set of 15,809 total boxes (7,931 'person' boxes). The metric used is top 1 accuracy.

**RGB-T Selected (RGB-T)** [6]. We follow the processing methodology employed for LLVIP [7] in handling the RGB-T dataset. For a binary classification task, we specifically select 10,000 total bounding boxes, out of which 5,131 are labeled as 'person.' The top-1 accuracy is designated as the evaluation metric.

**MSR-VTT (MSR-VTT)** [19]. MSR-VTT contains a diverse set of videos covering a wide range of topics and scenarios. The dataset consists of 10,000 video clips from 20 categories. In this work, we evaluate multi-modal methods on the recognition task with these 20 categories. The metric used is accuracy.

**UFC-101 (UFC-101)** [16]. The UFC-101 dataset is a prevalent benchmark in the domain of action recognition. It encompasses a total of 13,320 video clips, portraying 101 distinct human action categories. For evaluation, accuracy is employed as the metric.

**N-Caltech-101 (N-cal)** [12]. N-Caltech-101 comprises paired event data associated with the Caltech-101 [3] dataset. This dataset is employed for tasks encompassing event recognition, event-to-image retrieval, and image-to-event retrieval. The evaluation metric for the recognition task is accuracy, while for retrieval tasks, we utilize recall.

**N-ImageNet-1K (N-IN-1K)** [8]. N-ImageNet-1K encompasses paired event data derived from the ImageNet-1K [2] dataset. The evaluation focuses on assessing the event recognition capabilities of models within this dataset. Accuracy is employed as the metric for this evaluation.

### B.2. Experiment Details

#### B.2.1 Zero-shot Recognition

In the zero-shot setting, we evaluate all baseline models and our UniBind without training. For all baseline models, we use the default templates from CLIP [14], and we use our localized embedding centers for our UniBind.

#### B.2.2 Fine-tuning Recognition

For the fine-tuning setting, our experiments were done on 80GB A800 GPUs, and we detail the hyperparameters used for training with PointBind [5] reported in Tab 1

## C. Additional Ablation Study

### C.1. LLM-augmented Contrastive Learning

We present additional visualization results are shown in Fig. 5. We show the comparison of the PointBind [5] representation space and our UniBind representation space. In the representation space built by PointBind, embeddings from different modalities tend to cluster around their respective modalities. Thereby, with LLM-augmented contrastive learning, multi-modal embeddings cluster around the same semantic label in our unified modality-agnostic representation space.

We additionally present additional results pertaining to the cross-modal retrieval task. Our experimentation involves E-CLIP [21] and PointBind [5], and we subsequently present the outcomes for both event-to-image retrieval and image-to-event retrieval in Table 2. The observed improvement in recall scores incrementally rises from the top 1 to the top 20, highlighting the effectiveness of our approach in aligning modalities with semantics.

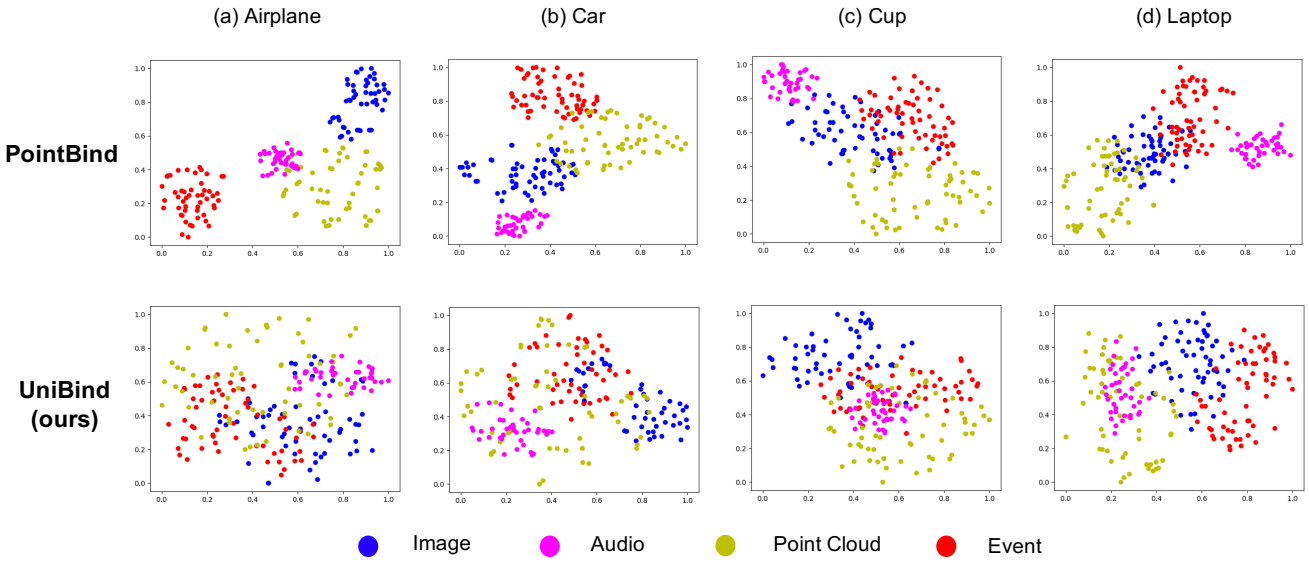


Figure 5. Representation space visualization of PointBind [5] and our UniBind.

| Model | Image-to-Event | | | | Event-to-Image | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@20 | R@1 | R@5 | R@10 | R@20 |
| E-CLIP [21] | 79.52 | 90.11 | 93.08 | 95.51 | 76.29 | 89.97 | 91.80 | 94.61 |
| **E-CLIP w LCL** | 78.95 | 89.79 | 94.32 | 97.06 | 77.04 | 91.51 | 93.62 | 96.70 |
| Δ | -0.57 | -0.32 | **+1.24** | **+1.55** | **+0.75** | **+1.54** | **+1.82** | **+2.09** |
| PointBind (**+Event**) [5] | 14.07 | 31.40 | 40.79 | 49.46 | 9.00 | 22.23 | 29.32 | 37.70 |
| **PointBind w LCL** | 14.12 | 31.91 | 41.25 | 50.98 | 14.29 | 33.65 | 44.34 | 55.66 |
| Δ | **+0.05** | **+0.51** | **+0.46** | **+1.52** | **+5.29** | **+11.42** | **+15.02** | **+17.96** |

Table 2. Multi-modal retrieval result **with/without LLM-augmented contrastive Learning**. We evaluate E-CLIP and PointBind in Image-to-Event and Event-to-Image tasks.

## C.2. Embedding Center Localization

We show more visualization results from image, point cloud, event, audio, video, and thermal modalities in Fig. 6. Our embedding centers result in more distinct semantic boundaries between different categories, effectively enhancing recognition accuracy and reducing interference from other categories.

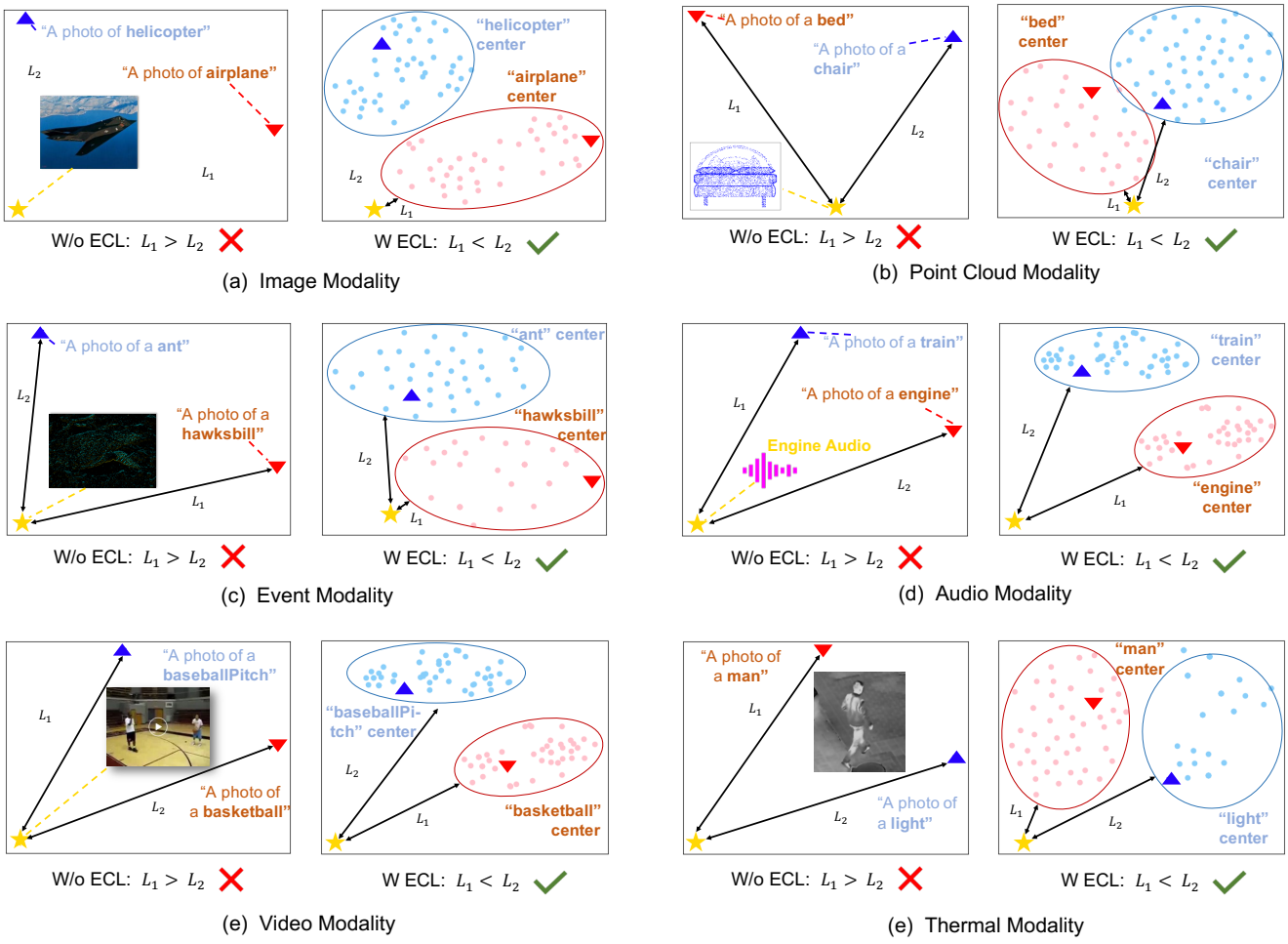


Figure 6. Embedding centers.

## References


[1] 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

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 3

[3] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 2, 3

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[5] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 3, 4

[6] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 3

[7] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021. 3

[8] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2146–2156, 2021. 3

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[10] Alejandro López-Cifuentes, Marcos Escudero-Vinolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020. 2, 3

[11] OpenAI. Gpt-4 technical report, 2023. 1

[12] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015. 3

[13] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[15] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014. 3

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[18] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2, 3

[19] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296, 2016. 3

[20] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 1, 2

[21] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-clip: Towards label-efficient event-based open-world understanding by clip. *arXiv preprint arXiv:2308.03135*, 2023. 3, 4