

UNIALIGN: Scaling Multimodal Alignment within One Unified Model

Bo Zhou^{1*}, Liulei Li^{2*}, Yujia Wang³, Huafeng Liu¹, Yazhou Yao^{1†}, Wenguan Wang²

¹ Nanjing University of Science and Technology, ² Zhejiang University, ³ Zhejiang Sci-Tech University

<https://github.com/NUST-Machine-Intelligence-Laboratory/UNIALIGN>

Abstract

We present UNIALIGN, a unified model to align an arbitrary number of modalities (e.g., image, text, audio, 3D point cloud, etc.) through one encoder and a single training phase. Existing solutions typically employ distinct encoders for each modality, resulting in increased parameters as the number of modalities grows. In contrast, UNIALIGN proposes a modality-aware adaptation of the powerful mixture-of-experts (MoE) schema and further integrates it with Low-Rank Adaptation (LoRA), efficiently scaling the encoder to accommodate inputs in diverse modalities while maintaining a fixed computational overhead. Moreover, prior work often requires separate training for each extended modality. This leads to task-specific models and further hinders the communication between modalities. To address this, we propose a soft modality binding strategy that aligns all modalities using unpaired data samples across datasets. Two additional training objectives are introduced to distill knowledge from well-aligned anchor modalities and prior multimodal models, elevating UNIALIGN into a high performance multimodal foundation model. Experiments on 11 benchmarks across 6 different modalities demonstrate that UNIALIGN could achieve comparable performance to SOTA approaches, while using merely 7.8M trainable parameters and maintaining an identical model with the same weight across all tasks.

1. Introduction

Humans accomplish tasks by processing and combining signals from different modalities [37, 53]. This gives rise to multimodal learning, which aims to build AI models that can extract and relate information across multimodal inputs [97]. Recent efforts have been made to learn a unified representation space via the alignment between modalities [18, 22, 39]. Though enabling extension to novel modalities, existing solutions typically rely on a *one-versus-one* paradigm where merely one extended modality (e.g., point

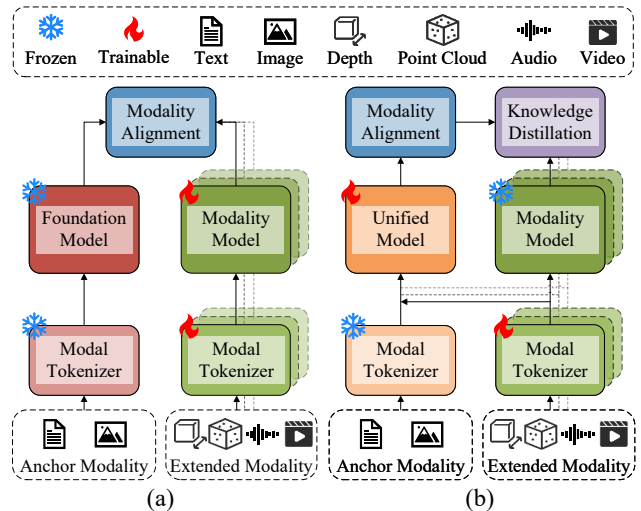


Figure 1. **Comparison of architectures** between standard multimodal alignment methods [18, 39, 109] and UNIALIGN: (a) Existing work adopts specialized encoders tailored to each modality; (b) UNIALIGN unifies the alignment across multiple modalities in a single model, with one training phase across all datasets.

cloud, audio) could be aligned with the anchor modalities (e.g., text, image) at each time (Fig. 1(a)). This presents several drawbacks: **i)** one dedicated encoder is required for each modality, leading to exponential increases in training cost and parameter count as the number of aligned modalities grows, severely limiting scalability; **ii)** task-specific training for each dataset, results in multiple independent models for different tasks; and **iii)** the separate alignment of each extended modality hinders the model to leverage the full breadth of data across all datasets to shape the shared embedding space from a global view. This further impedes cross-modal knowledge transfer where insights gained from one modality could benefit understanding in others.

These observations prompt us to rethink: *Is this one-versus-one alignment paradigm essential?* and *Can we achieve joint alignment of all modalities within one single model?* In response, we propose UNIALIGN (Fig. 1(b)), which embraces a *all-in-one* design to realize multimodal alignment by constructing a foundation model that includes: **① one** base encoder to UNify the encoding of *all* anchor and

*The first two authors contribute equally to this work.

†Corresponding author: Yazhou Yao.

extended modalities, and **②** *one* training phase to **ALIGN** *all* modalities with *all* samples across datasets collectively.

Concretely, to achieve **①**, recent work has consistently demonstrated that Transformers pre-trained on one modality possess exceptional generalization ability to novel modalities [13, 67]. Moreover, the mixture-of-experts (MoE) [72] schema presents a highly efficient strategy to scale model capacity under a fixed computational cost [108]. These insights motivate us to build a foundational model accommodating multiple modalities upon a single pre-trained Transformer (e.g., CLIP [67]). Here, MoE can dynamically scale and allocate model capacities by selectively activating experts to encode inputs across diverse modalities. We explore two strategies targeting at expert grouping and routing mechanism to enhance the awareness of different modalities within MoE. To preserve the knowledge learned from anchor modalities, Low-Rank Adaptation (LoRA) [27] is integrated with MoE, enabling a significant reduction in training overhead. To fulfill **②**, we propose a soft modality binding strategy, leveraging massive unpaired data samples across modalities and datasets. Specifically, given unpaired data from diverse datasets with different modalities, the class labels of samples are encoded into text embeddings to create a unified representation. The semantic similarity between text embeddings can inform the alignment between modalities (e.g., audio and video), by encouraging the similarity between modality embeddings and textual label embeddings (e.g., $\text{audio} \leftrightarrow \text{text}_{\text{video}}$ and $\text{video} \leftrightarrow \text{text}_{\text{audio}}$) to be identical as the semantic similarity (i.e., $\text{text}_{\text{video}} \leftrightarrow \text{text}_{\text{audio}}$). Moreover, rooted in the principles of knowledge distillation [2, 26, 110], the alignment between extended and anchor modalities is guided by the knowledge persisting between anchor modalities, which are well-aligned in semantics compared to extended modalities. To further boost the performance, we transfer knowledge from multiple well-trained foundation models which only contain partial modalities. Given the complexity of multimodal, multi-dataset joint training, a specialized batching strategy combined with modality-aware gradient accumulation is proposed, to mitigate the potential dispersion in optimization across different modalities.

UNIALIGN enjoys several advantages: **First**, it supports flexible alignment between arbitrary number of anchor and extended modalities, while ensuring negligible increase in parameter size. **Second**, the alignment for all modalities is accomplished in one training session, significantly improving training efficiency. **Third**, unpaired data from different modalities can be used collectively for optimization. Uncommon modalities with scarce data could take full advantages of the rich resource of common modalities. **Fourth**, benefited from soft modality binding, extended modalities are bridged for mutual boosting. **Fifth**, a compact and high-performance multimodal foundational

model generally applicable to all tasks is delivered at minimal cost. This contrasts with existing approaches, which typically provide multiple loosely integrated modality encoders tailored to specific tasks, thereby enabling broader transferability to other domains and downstream applications [3, 8, 16, 61, 62, 95, 96, 101, 102].

By embracing the *all-in-one* philosophy, UNIALIGN demonstrates remarkable efficiency in both training cost and parameter size. After a single pre-training step without further tuning, UNIALIGN achieves top-leading performance on 11 benchmarks with one identical model and **7.8M** trainable parameters **across 6 modalities** (previous SOTA [18, 39] needing **over 80M** trainable parameters for **each extended modality**). This evidences the great potential of our unified modality encoding and aligning strategies. We hope this work could stimulate the community to rethink multimodal alignment from a universal and compact perspective.

2. Related Work

Multimodal Learning. Multimodal learning has attracted significant attention given its ability to address a wide range of applications by integrating information from diverse modalities such as text, images, audio, and video [49, 51, 67, 103]. Current research in this field generally falls into three categories: **First**, methods involving joint training of multiple modalities aim to improve performance across different downstream tasks, by leveraging shared representations [17, 78, 107]. Though incorporating multiple modalities as inputs, these work lacks explicit alignment between modalities. **Second**, inspired by the successes of vision-language alignment [67], some approaches extend this alignment to additional modalities [18, 38, 40, 56, 59, 91, 98, 109]. These methods typically require an independent encoder for each modalities. Though [14, 18, 39, 90] maintain a consistent architecture across tasks, they conduct separate training on individual datasets, resulting in multiple task-specific models. Additionally, these solutions often rely on massive paired data to achieve robust zero-shot performance across various tasks, ignoring the usage of rich unpaired data. **Third**, the integration of multimodal inputs with large language models (LLMs) has opened new avenues for multimodal learning [23, 50, 52, 88, 100]. In these approaches, all modalities are projected into the same representation space as LLMs, without alignment among extended modalities and supporting merely one anchor modality (i.e., text). This leads to suboptimal performance on certain vision tasks and limits the applications primarily to language-related fields.

In this work, we propose UNIALIGN, supporting alignment between an arbitrary number of anchor and extended modalities within one model using one unified modality encoder. After a single training phase, UNIALIGN can handle a dozen of downstream tasks without any fine-tuning. UNIALIGN also achieves high training efficiency, utilizing

merely 7.8M training parameters for all tasks.

Multi-Task Learning. To maximize the utilization of resource and improve generalization across tasks, multi-task learning seeks to develop models that can handle multiple tasks simultaneously [4]. Traditional methods often require separate subnetworks for each task, which hinders scalability [83]. In recent years, the mixture-of-experts (MoE) architecture [15, 72] has gained prominence in multi-task learning [31]. MoE leverages multiple subnetworks, or “experts”, that are selectively activated to optimize task performance while reducing computational costs [35]. This approach proves beneficial across various domains, such as language processing [31, 35] and vision tasks [41, 43, 69]. With the success of LoRA [27], there has been a surge to integrate MoE with LoRA, which shows promise in enhancing both computational efficiency and task adaptability in large-scale models [11, 21, 28, 29]. However, these methods are mainly confined to text and image domains [7, 11], and lack alignment between modalities [46]. In contrast, UNIALIGN extends MoE to a multimodal setup, devising modality-aware expert selection and routing mechanism.

Knowledge Distillation. Knowledge distillation [26, 34, 55] is an essential technique for model compression, enabling the transfer of knowledge from large, computationally intensive models to smaller, more efficient ones. This is particularly valuable for multimodal foundation models, where the computational overload is heavy [32, 45, 79, 99]. Initially, knowledge distillation focused on extracting knowledge from individual large models, typically using the logits or intermediate feature representations to guide the training of student models [1, 25, 36, 63, 93]. Recent research has expanded this concept to integrate capabilities from multiple models, creating hybrid models that not only improve task-specific performance but also broaden the range of potential applications [68, 71, 84, 104, 105, 112].

Compared to existing work, UNIALIGN distills knowledge from both well-aligned anchor modalities, and multiple well-trained foundation models specialized in partial modalities. This allows our method to handle a broader range of tasks that require understanding and processing information with different composition of diverse modalities.

3. Method

In this section, we first introduce the unified encoding of all modalities within one single model (§3.1), then describe the universal modality alignment strategy across all datasets (§3.2), and finally present the implementation details of UNIALIGN (§3.3).

3.1. Unified Modality Encoding

Preliminary: Mixture-of-Experts (MoE). Let $\mathbf{X} \in \mathbb{R}^{L \times D}$ denote the tokenized input, where L is the sequence length and D is the token embedding size. In vanilla Transformer,

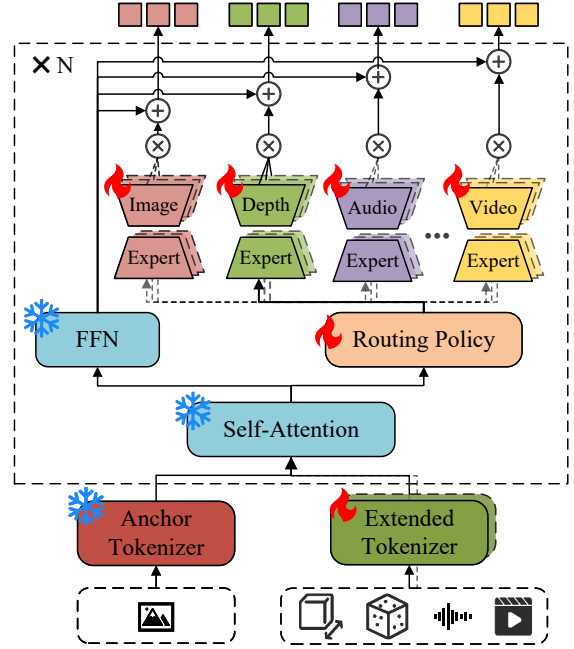


Figure 2. Architecture of the unified modality encoder.

each encoder block comprises a multi-head self-attention (MSA) layer and a feed-forward network (FFN), as follows:

$$\mathbf{X}' = \mathbf{X} + \mathcal{F}^{\text{MSA}}(\text{LN}(\mathbf{X})), \quad (1)$$

$$\mathbf{X}'' = \mathbf{X}' + \mathcal{F}^{\text{FFN}}(\text{LN}(\mathbf{X}')), \quad (2)$$

where \mathbf{X}'' is the outputs and LN represents layer normalization. In the mixture-of-experts (MoE) schema, an MoE function (*i.e.*, \mathcal{F}^{MoE}) is adopted to replace \mathcal{F}^{FFN} :

$$\mathbf{X}'' = \mathbf{X}' + \mathcal{F}^{\text{MoE}}(\text{LN}(\mathbf{X}')), \quad (3)$$

$$\mathcal{F}^{\text{MoE}}(\mathbf{X}') = \sum_{i=1}^n \mathcal{G}(\mathbf{X}')_i \mathcal{F}_i^{\text{Expert}}(\mathbf{X}'), \quad (4)$$

where \mathcal{F}^{MoE} jointly aggregates outputs from multiple expert networks (*i.e.*, $\mathcal{F}^{\text{Expert}}$) according to the router $\mathcal{G}(\mathbf{X}')_i = \text{softmax}(\mathcal{F}^{\text{Router}}(\mathbf{X}'))_i$ to assign weights for different experts and render the predictions.

Multimodal Adaptation with MoE. We aim to unify the encoding of multimodal inputs within one foundation model (*e.g.*, the image encoder of CLIP [67]). As illustrated in Fig. 2, inputs of various modalities are first converted into sequential tokens. We follow [39, 42, 107] to preprocess multimodal data and implement modality-specific tokenizers, with details summarized in *Appendix*. To adapt the encoding of samples from various modalities, UNIALIGN explores two modality-aware MoE strategies:

- **Modality-Aware Expert Grouping** divides all $\mathcal{F}^{\text{Expert}}$ in one MoE layer into multiple sets, where each set \mathcal{B} contains n experts to process inputs from a specific modality \mathcal{M} :

$$\mathcal{B}_{\mathcal{M}} = \{\mathcal{F}_{\mathcal{M},1}^{\text{Expert}}, \dots, \mathcal{F}_{\mathcal{M},n}^{\text{Expert}}\}. \quad (5)$$

Under this setup, following [46], we construct a hierarchical routing mechanism. In the first routing round, tokens are

directed to a specific expert set based on the modality of inputs. Next, within an expert set, a learnable router $\mathcal{F}_{\mathcal{M}}^{\text{Router}}$ for modality \mathcal{M} assigns weights to each individual expert conditioned on the input tokens. This separation lets each expert set to specialize in one modality, and allows for easy expansion to novel modalities by loading new expert sets.

• **Modality-Aware Global Routing** uses multiple modality-specific routers, to direct tokens across all experts in a MoE layer without grouping. Given the routing function $\mathcal{F}_{\mathcal{M}}^{\text{Router}}$ for modality \mathcal{M} , the global weight $\mathcal{G}_{\mathcal{M}}$ is computed as:

$$\mathcal{G}_{\mathcal{M}}(\mathbf{X}')_i = \text{softmax}(\mathcal{F}_{\mathcal{M}}^{\text{Router}}(\mathbf{X}')_i). \quad (6)$$

By allowing the experts to be accessible to each router, this strategy enables cross-modality knowledge sharing among experts. However, since the experts are not explicitly specialized for one modality, the generalization potentially at the expense of fine-grained modality specialization.

Modality-Aware LoRA Expert. Low-Rank Adaptation (LoRA) [28] has demonstrated both effectiveness and efficiency in fine-tuning large models [9, 81]. Therefore, to maintain the knowledge in pre-trained foundation models, the original feed-forward network (*i.e.*, \mathcal{F}^{FFN} in Eq. 2) are preserved in a fixed state instead of being replaced during the forward of MoE (*i.e.*, Eq. 3), resulting in:

$$\mathbf{X}'' = \mathbf{X}' + \mathcal{F}^{\text{FFN}}(\text{LN}(\mathbf{X}')) + \mathcal{F}^{\text{MoE}}(\text{LN}(\mathbf{X}')). \quad (7)$$

Here the experts in \mathcal{F}^{MoE} leverage \mathcal{F}^{FFN} as the base weight for LoRA adaption, with each of them employs a specific instance of LoRA (*i.e.*, $\mathcal{F}^{\text{LoRA}}$) parameterized as follows:

$$\mathcal{F}_{\text{MoE}}(\mathbf{X}') = \sum_{i=1}^n \mathcal{G}(\mathbf{X}')_i \mathcal{F}_i^{\text{LoRA}}(\mathbf{X}'), \quad (8)$$

$$\mathcal{F}_i^{\text{LoRA}}(\mathbf{X}') = \frac{\alpha}{r} \cdot \mathbf{B}_i \mathbf{A}_i \mathbf{X}', \quad (9)$$

where low-rank matrices $\mathbf{A} \in \mathbb{R}^{d_{in} \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times d_{out}}$ are scaled by a factor α normalized by the rank $r \ll \min(d_{in}, d_{out})$. In this way, we achieve seamless integration of LoRA with MoE, enabling efficient tuning and knowledge preservation of the foundation models, while supporting adaptation to inputs across various modalities.

3.2. Universal Modality Alignment

Soft Modality Binding across Datasets. Prior work [18] reveals that modalities expressed in a joint embedding space could be automatically aligned through an intermediate modality. For example, the alignment of $\langle \text{Audio}, \text{Image} \rangle$ and $\langle \text{Image}, \text{Text} \rangle$, could lead to the alignment between $\langle \text{Audio}, \text{Text} \rangle$ with Image as the bridge. However, this approach requires massive naturally paired samples between anchor and extended modalities, which are often challenging and time-consuming to obtain. We thus propose a soft modality binding strategy, which utilizes unpaired samples across datasets with different modalities

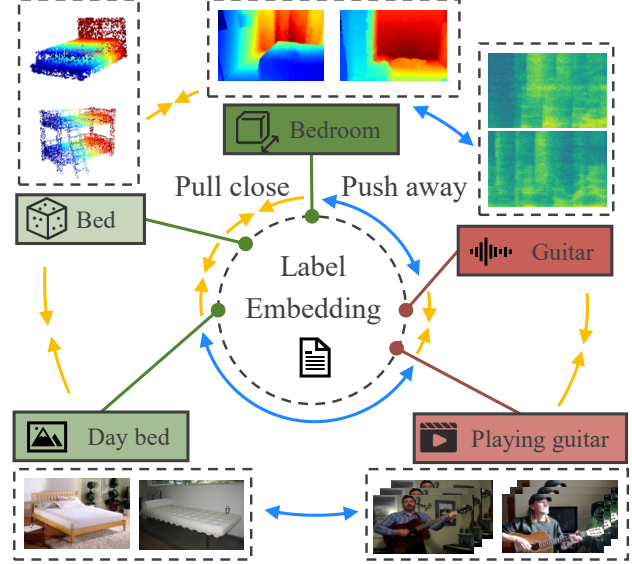


Figure 3. Our proposed soft modality binding strategy. Unpaired samples from diverse datasets with different modalities are bridged with the semantic similarity between class labels as soft guidance.

(Fig. 3). Specifically, let $\mathcal{Y}^{\mathcal{M}_1} = \{\mathbf{y}_1^{\mathcal{M}_1}, \dots, \mathbf{y}_{N_1}^{\mathcal{M}_1}\}$ and $\mathcal{Y}^{\mathcal{M}_2} = \{\mathbf{y}_1^{\mathcal{M}_2}, \dots, \mathbf{y}_{N_2}^{\mathcal{M}_2}\}$ represent text embeddings of class labels from two arbitrary datasets in modalities \mathcal{M}_1 and \mathcal{M}_2 , respectively. The semantic similarity of textual labels across datasets can be computed as:

$$\mathcal{S}(\mathbf{y}_i^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_2}) = \frac{\exp(\mathbf{y}_i^{\mathcal{M}_1} \cdot \mathbf{y}_j^{\mathcal{M}_2})}{\sum_{n=1}^{N_2} \exp(\mathbf{y}_i^{\mathcal{M}_1}, \mathbf{y}_n^{\mathcal{M}_2})}. \quad (10)$$

Similarly, given samples $(\mathbf{x}^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_1})$ from the first dataset and $(\mathbf{x}^{\mathcal{M}_2}, \mathbf{y}_j^{\mathcal{M}_2})$ from the second dataset, the semantic similarity between $\mathbf{x}^{\mathcal{M}_1}$ and text embedding of labels from the second dataset (*i.e.*, $\mathcal{S}(\mathbf{x}^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_2})$), as well as $\mathbf{x}^{\mathcal{M}_2}$ and text embeddings of labels from the first dataset (*i.e.*, $\mathcal{S}(\mathbf{x}^{\mathcal{M}_2}, \mathbf{y}_i^{\mathcal{M}_1})$) could be computed. Taking the similarity between class labels of two datasets (*i.e.*, $\mathcal{S}(\mathbf{y}_i^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_2})$) as a soft prior for alignment, we bind modalities across datasets by optimizing the following objective:

$$\begin{aligned} \mathcal{L}_{\text{soft_bind}} = & \text{KL}(\mathcal{S}(\mathbf{x}^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_2}) | \mathcal{S}(\mathbf{y}_i^{\mathcal{M}_1}, \mathbf{y}_j^{\mathcal{M}_2})) \\ & + \text{KL}(\mathcal{S}(\mathbf{x}^{\mathcal{M}_2}, \mathbf{y}_i^{\mathcal{M}_1}) | \mathcal{S}(\mathbf{y}_j^{\mathcal{M}_2}, \mathbf{y}_i^{\mathcal{M}_1})), \end{aligned} \quad (11)$$

where KL refers to the Kullback-Leibler divergence. Our soft modality binding strategy is different from [18], which requires instance-level textual annotations for each sample. Instead, our approach utilizes merely class labels which are instance-agnostic and challenging for alignment.

Knowledge Distillation from Anchor Modalities. Given multiple well-aligned anchor modalities (*e.g.*, \mathcal{M}_1^a and \mathcal{M}_2^a) and one novel modality to be extended (*e.g.*, \mathcal{M}_k^e), the alignment performance between anchor and extended modalities (*e.g.*, \mathcal{M}_k^e and \mathcal{M}_1^a) is typically less effective than that between anchor modalities (*e.g.*, \mathcal{M}_1^a and \mathcal{M}_2^a). Therefore,

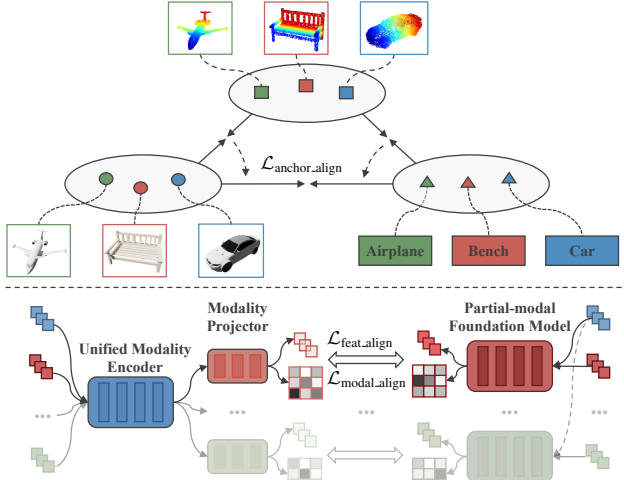


Figure 4. Top: Knowledge distillation from anchor modalities to instruct the alignment of novel modalities. Bottom: knowledge distillation from multiple partial-modal foundation models.

knowledge within the alignment between anchor modalities (Fig. 4 (Top)) could be transferred to improve the alignment learning of extended modalities via knowledge distillation.

For instance, given a 3D point cloud sample x_i , we could obtain its tokenized input $x_i^{\mathcal{P}}$ in the extended modalities, along with the text embedding of class label $x_i^{\mathcal{T}}$ and the visual embedding of one rendered image from a specific viewpoint $x_i^{\mathcal{I}}$, both in the anchor modalities. By leveraging the similarity between anchor modalities (*i.e.*, $\mathcal{S}(x_i^{\mathcal{T}}, x_i^{\mathcal{I}})$) as the guidance, we could optimize the alignment between the extended modality and anchor modalities through:

$$\begin{aligned} \mathcal{L}_{\text{anch_distil}} = & \text{KL}(\mathcal{S}(x_i^{\mathcal{P}}, x_i^{\mathcal{T}}) | \mathcal{S}(x_i^{\mathcal{I}}, x_i^{\mathcal{T}})) \\ & + \text{KL}(\mathcal{S}(x_i^{\mathcal{P}}, x_i^{\mathcal{I}}) | \mathcal{S}(x_i^{\mathcal{T}}, x_i^{\mathcal{I}})). \end{aligned} \quad (12)$$

This knowledge distillation strategy is generally applicable to most modalities, where 3D point cloud, video, and depth data can be converted into images with class labels.

Knowledge Distillation from Partial-Modal Foundation Models. To further boost the alignment across modalities, we employ foundation models \mathcal{E}_f skilled in partial modalities (*e.g.*, audio \leftrightarrow text [22], point cloud \leftrightarrow image [98]) as teacher models to guide the learning of our unified modality encoder \mathcal{E}_u (*i.e.*, Fig. 4 (Bottom)) from two perspectives:

- **Feature Alignment** aims to ensure features from our unified modality encoder to be identical as those from partial-modal foundation models. Specifically, denoting $\hat{F}_i^{\mathcal{M}_k}$ and $F_i^{\mathcal{M}_k}$ as features in modality \mathcal{M}_k output by \mathcal{E}_f and \mathcal{E}_u , we use the mean squared error (MSE) to align them as:

$$\mathcal{L}_{\text{feat_align}} = \|\hat{F}_i^{\mathcal{M}_k} - F_i^{\mathcal{M}_k}\|_2^2. \quad (13)$$

- **Cross-Modality Knowledge Transfer** aims to benefit the learning of \mathcal{E}_u via the well-aligned representation space across modalities shaped by large-scale paired data. Specifically, given a paired sample $(x_i^{\mathcal{M}^a}, x_i^{\mathcal{M}^e})$ in anchor modality \mathcal{M}^a and extended modality \mathcal{M}^e respectively, the align-

ment results of our unified modality encoder are encouraged to approximate those of partial-modal foundation models:

$$\begin{aligned} \mathcal{L}_{\text{modal_align}} = & \text{KL}(\mathcal{S}(x_i^{\mathcal{M}^e}, x_i^{\mathcal{M}^a}) | \mathcal{S}(\hat{x}_i^{\mathcal{M}^e}, \hat{x}_i^{\mathcal{M}^a})) \\ & + \text{KL}(\mathcal{S}(x_i^{\mathcal{M}^a}, x_i^{\mathcal{M}^e}) | \mathcal{S}(\hat{x}_i^{\mathcal{M}^a}, \hat{x}_i^{\mathcal{M}^e})), \end{aligned} \quad (14)$$

where \hat{x}_i and x_i are outputs of \mathcal{E}_f and \mathcal{E}_u , respectively.

In summary, the loss for knowledge distillation from partial-modal foundation models can be formulated as:

$$\mathcal{L}_{\text{model_distil}} = \lambda \mathcal{L}_{\text{feat_align}} + \mathcal{L}_{\text{modal_align}}. \quad (15)$$

Following [99], we set $\lambda = 2000$ to scale the magnitude.

3.3. Implementation Details

Network Configuration. We adopt CLIP [67] pre-trained vision encoder as the base network for unified modality encoding. The sequence length L and channel dimension D for tokenized inputs \mathbf{X} are set to 196 and 768, respectively. The LoRA configuration is defined with a scaling factor α of 32 and a rank r of 4. For knowledge distillation from partial-modal foundation models, we use ULIP [98] as the teacher model for 3D point cloud inputs, VIT-LENS [38] for depth inputs, ONE-PEACE [86] for audio and video inputs.

Batching Strategy. The batching strategies proposed by [17, 42] sample inputs from each dataset (modality) separately, which may lead to catastrophic forgetting [42]. To address this, we develop a new batching strategy, which mixes data from all modalities within each mini-batch. Furthermore, a modality-aware gradient accumulation is proposed to perform forward pass, loss computation, and gradient calculation individually for each modality. Backpropagation is conducted only after processing all modalities.

Training Objective. Following [68], we assign an equal weight to all training objectives:

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \alpha_1 \mathcal{L}_{\text{soft_bind}} + \alpha_2 \mathcal{L}_{\text{anch_distil}} + \alpha_3 \mathcal{L}_{\text{model_distil}}, \quad (16)$$

where $\alpha_{\{1,2,3\}} = 1$ and \mathcal{L}_{NCE} is the modality aligning loss for paired data, as utilized in CLIP [67].

4. Experiments

4.1. Experimental Setup

Pre-training Datasets. We perform a joint pre-training across multiple datasets with different modalities, including images from ImageNet-1K [8], 3D point clouds from ShapeNet [5], depth data from SUN RGB-D [74], audio and video from AudioSet [16] and VGGSound [6]. To address the differences in dataset size across modalities, we apply repeated augmentation [17] to small datasets (such as SUN-D) and category-balanced downsampling for large datasets (such as AudioSet) to ensure balanced performance during joint training. Once finished pre-training, we directly evaluate our model on downstream tasks without any fine-tuning.

Training. For image and video data, we use a resolution of 224×224 and apply standard augmentations. For single-view depth data, we employ the same augmentation as in

Model	Backbone	Trainable Params	Top-1 Accuracy	Top-5 Accuracy
Single Task				
Swin Transformer [47]	Swin-B	87.7M	83.5	96.5
ConvNeXt [48]	ConvNeXt-B	88.7M	83.8	96.7
DeiT III [80]	ViT-B	86.9M	83.8	96.5
Conformer [64]	Conformer-B	83.3M	83.8	96.7
Hiera [70]	Hiera-B+	70.0M	85.2	97.3
Multiple Tasks				
UniRepLKNet [10]	UniRepLKNet-S	55.6M	83.9	-
OMNIVORE [17]	Swin-B	88.2M	85.3	97.5
MetaFormer [107]	ViT-B	86.6M	85.4	97.4
UNIALIGN[†]	ViT-B	7.8M	85.3	97.2

[†]:An identical model with the same weight across all tasks.

Table 1. Results for image classification on ImageNet-1K (§4.2).

Model	Backbone	Trainable Params	ModelNet40 Top-1	ModelNet40 Top-5	ScanObjectNN Top-1	ScanObjectNN Top-5
Single Task						
PointCLIP [106]	ResNet-50	-	19.3	34.8	10.5	30.6
PointCLIPv2 [111]	ViT-B	-	63.6	85.0	42.2	74.5
CLIP2Point [30]	ViT-B	88.3M	49.5	81.2	25.5	59.4
RECON [66]	ViT-B	43.6M	61.2	78.1	42.3	75.6
ULIP [98]	ViT-B	43.6M	60.4	84.0	51.5	80.2
Multiple Tasks						
VIT-LENS [38]	ViT-B	34.1M	65.4	92.7	-	-
UNIALIGN[†]	ViT-B	7.8M	55.9	84.3	<u>42.5</u>	<u>76.3</u>

[†]:An identical model with the same weight across all tasks.

Table 2. Results for zero-shot 3d shape classification on ModelNet40 test [95] and ScanObjectNN test [82] (§4.3).

[38]. The augmentation of audio data also follows [38] except Mix-up [94]. We train the model with a batch size of 2048 (after gradient accumulation) for 100 epochs. An AdamW optimizer with a initial learning rate $2e^{-4}$ and a weight decay of 0.05 is employed. For knowledge distillation from partial-modal foundation models, we cache offline features to avoid repeatedly forwarding. The base modality encoder remains frozen during training, with only LoRA enhanced MoE modules and the projection head for knowledge distillation remaining trainable. Flash-attention is applied to reduce GPU memory cost.

Reproducibility. UNIALIGN is implemented in PyTorch and trained on four NVIDIA Tesla A100 GPUs.

4.2. Results for Image Classification

Dataset. We evaluate UNIALIGN for image classification on ImageNet-1K val [8] which contains 50,000 samples across 1,000 classes, spanning objects, animals, and more.

Metric. Following previous works [47, 48, 87], we report the count of trainable parameters, top-1 and top-5 accuracy.

Performance. As shown in Table 1, UNIALIGN demonstrates impressive performance on image classification, achieving a comparable performance to modality-specific approaches (*i.e.*, 85.3% vs. 85.2% of Hiera [70]). It is noteworthy that our method utilizes merely **9%** trainable parameters and a unified modality encoder architecture which preserves the capability to process other modalities.

Model	Backbone	Trainable Params	Anchor Modality	NYU-D	SUN-D
Single Task					
G-L-SOOR [76]	D-CNN [75]	-	-	62.3	44.1
RecgNet [12]	ResNet18 [24]	11.2M	-	57.7	47.9
Depth Swin [18]	Swin-B	88.2M	-	76.4	69.1
Multiple Tasks					
Text Paired [18]	ViT-H	88.1M	T	41.9	25.4
ImageBind [18]	ViT-H	88.3M	I	54.0	35.1
VIT-LENS [38]	ViT-L	50.9M	I+T	68.5	52.2
UNIALIGN[†]	ViT-B	7.8M	T	46.3	34.3
UNIALIGN[†]	ViT-B	7.8M	I+T	59.4	45.8
UNIALIGN[†]	ViT-B	7.8M	I+T	71.4*	58.2*

[†]:An identical model with the same weight across all tasks.

Table 3. Results for depth-only scene classification on NYU-D test [73] and SUN-D test [74] (§4.4). *: The depth maps are tokenized by the image tokenizer additionally.

4.3. Results for Zero-Shot 3D Shape Classification

Dataset. We evaluate zero-shot 3D shape classification on two datasets. ModelNet40 test [95] contains 2,468 samples from 40 categories, while ScanObjectNN test [82] consists of 581 samples distributed across 15 categories.

Metric. The top-1 and top-5 accuracy are reported.

Performance. As shown in Table 2, UNIALIGN achieves a top-1 accuracy of 55.9% and 42.5% on ModelNet40 val, and ScanObjectNN test, respectively. Though SOTA approaches such as ULIP [98] and ViT-LENS [38] yield higher scores, they require a modality-specific point-cloud encoder and an additional image encoder for alignment.

4.4. Results for Depth-only Scene Classification

Dataset. Experiments for depth-only scene classification are conducted on SUN RGB-D test [74] which contains 4,660 samples across 19 classes, and NYU-D test [73] which contains 653 samples across 10 semantic classes.

Metric. The top-1 accuracy is adopted for evaluation.

Performance. A detailed comparison of UNIALIGN against several top-leading approaches for depth-only scene classification is provided in Table 3. Building on a ViT-B backbone and aligning extended modalities with both image and text as anchor modalities, UNIALIGN achieves a top-1 accuracy of 71.4% on NYU-D test and 58.2% on SUN RGB-D test, surpassing all competitors. Notably, aligning merely to the text modality leads to a marked performance decline (*i.e.*, 71.4%→46.3% on NYU-D). This suggests that **i)** the depth modality exhibits a closer alignment with image data, and **ii)** UNIALIGN can effectively leverage the additional anchor modality to enhance alignment performance.

4.5. Results for Audio Classification and Retrieval

Datasets. We evaluate audio classification on AudioSet val [16] containing 2M samples across 527 classes, and ESC [65] including 2K data across 50 classes. Audio-text retrieval is evaluated on AudioCaps test [33] which includes 813 audio-text pairs. Moreover, following [107], we

Method	Backbone	Anchor Modality	AudioSet mAP	ESC Top1	AudioCaps R@1	AudioCaps R@10
Single Task						
AVFIC [57]	ViT-B	-	-	-	8.7	37.7
AudioCLIP [22]	ViT-B	I+T	25.9	69.4	-	-
Multiple Tasks						
ImageBind-H [18]	ViT-H	I	17.6	66.9	9.3	42.3
VIT-LENS [38]	ViT-L	I+T	27.2	80.9	14.9	55.2
LanguageBind [109]	ViT-L	I+T	27.7	91.8	12.2	53.2
UNIALIGN [†]	ViT-B	I+T	23.8	70.5	11.7	49.3

[†]: An identical model with the same weight across all tasks.

Table 4. Results for audio classification on AudioSet *val* [16], ESC [65], and audio-text retrieval on AudioCaps *test* [33] (§4.5). Due to expiration of links, UNIALIGN is trained with only 0.5M videos from AudioSet, while methods such as ViT-Lens use 1.6M.

Method	Backbone	Pre-train Datasets	Trainable Params	Top-1
Single Task				
AST [19]	ViT-B	N/A	86.9M	92.6
AST [19]	ViT-B	AudioSet-20K	86.9M	96.2
SSAST [20]	ViT-B	AudioSet-2M	89.3M	97.8
Multiple Tasks				
MetaFormer [107]	ViT-B	LAION2B	1.1M	78.3
MetaFormer [107]	ViT-B	LAION2B	86.3M	97.0
UniRepLNet [10]	UniRepLNet-S	-	55.5M	98.5
UNIALIGN	ViT-B	Joint	1.8M	97.4

Table 5. Results for audio classification on Speech Commands V2 *val* [92]. All models are fine-tuned on the training set (§4.5).

fine-tune UNIALIGN on Speech Commands V2 *train* [92] for speech recognition.

Metric. Following [18, 22, 38], we measure mAP for AudioSet, top-1 accuracy for ESC and Speech Commands V2, and Recall@ k ($k = \{1, 5\}$) for AudioCaps.

Performance for Audio classification. UNIALIGN achieves promising results on AudioSet and ESC (Table 4). Though inferior to top-leading methods, it is important to note that, due to the expiration of links, a significant number of videos from AudioSet can not be downloaded from YouTube (*i.e.*, **1.6M** used in ViT-Lens [38] *vs.* **0.5M** used in UNIALIGN).

Performance for Audio retrieval. Similarly, despite the absence of 1.1M training data, UNIALIGN can still achieve comparable performance, achieving 11.7 R@1 accuracy.

Performance for Speech Recognition. As shown in Table 5, UNIALIGN demonstrates remarkable accuracy compared to existing work (*e.g.*, 97.4% *vs.* 97.0% for Meta-Transformer [107]), while utilizing significantly fewer trainable parameters (*e.g.*, **1.8M** *vs.* 86.3M). When tuning a similar number of parameters (*i.e.*, 1.1M for Meta-Transformer, and 1.8M for UNIALIGN), our method outperforms Meta-Transformer by **19.1%** (*i.e.*, 97.4% *vs.* 78.3%).

4.6. Results for Video Classification and Retrieval

Datasets. We utilize MSR-VTT *val* [96] which contains around 1,000 video-text pairs for zero-shot video-text retrieval, and UCF101 *val* [77] which contains 3,738 samples across 101 semantic classes for video classification.

Model	Backbone	Retrieval Modality	Trainable Params	R@1	R@5
Single Task					
MIL-NCE [54]	ResNeXt-101	V	88.8M	8.6	25.8
SupportSet [60]	R3D-34	V	93.5M	10.4	30.0
AVFIC [57]	ViT-B	A+V	86.6M	19.4	50.3
InternVideo [89]	ViT-L	V	305M	37.5	71.3
Multiple Tasks					
ImageBind [18]	ViT-H	A+V	89.3M	36.8	70.0
VIT-LENS [38]	ViT-L	A+V	127.6M	37.6	72.6
UNIALIGN [†]	ViT-B	A+V	7.8M	<u>37.0</u>	<u>70.7</u>

[†]: An identical model with the same weight across all tasks.

Table 6. Results for video retrieval on MSRVT *val* [96] (§4.6).

Model	Backbone	Zero-Shot	Fine-Tuned
Single Task			
ResT [44]	ResNet-101	58.7	-
ActionCLIP [85]	ViT-B	58.3	-
X-CLIP [58]	ViT-B	72.0	-
Multiple Tasks			
MetaFormer [107]	ViT-B	-	46.6
ImageBind [18]	ViT-H	64.8	98.1
UniBind [50]	ViT-H	73.7	93.3
UNIALIGN [†]	ViT-B	<u>68.5</u>	<u>94.4</u>

[†]: An identical model with the same weight across all tasks.

Table 7. Results for video classification on UCF101 *val* [77] (§4.6).

Metric. Recall@ k and top-1 accuracy are adopted as evaluation metrics for video-text retrieval and classification.

Performance for Video-Text Retrieval. As shown in Table 6, UNIALIGN achieves a R@1 score of 37.0% and R@5 score of 70.7%, which are comparable to advanced methods while using merely 7.8M trainable parameters.

Performance for Video Classification. We summarize the results for video classification in Table 7. As shown, UNIALIGN achieves an impressive accuracy of 68.5% on UCF101 *val* [77] without any fine-tuning, and outperforms ImageBind [18]. After fine-tuning all models on the training set of UCF101, our method achieves the second-highest performance. This may be attributed to the differences in backbone, as ImageBind utilizes ViT-H which could provide enhanced knowledge for downstream fine-tuning.

4.7. Qualitative Results

As shown in Fig. 5, UNIALIGN facilitates accurate cross-modality retrieval. Moreover, modalities can be combined to constrain attributes of the target in another modality.

4.8. Diagnostic Experiment

To evaluate the core designs of UNIALIGN and gain further insights, we conduct a series of ablation studies. Results are reported on SUN-D *test* [74] and ModelNet40 *test* [5].

Modality-Aware MoE Strategy. As illustrated in §3.1, we explore two modality-aware MoE strategies: expert grouping and global routing. Results in Table 8 indicate that dividing experts into modality-specific groups yields better performance. Thus, we adopt it as the default MoE strategy.

Expert Network. Table 9 presents the ablation study on

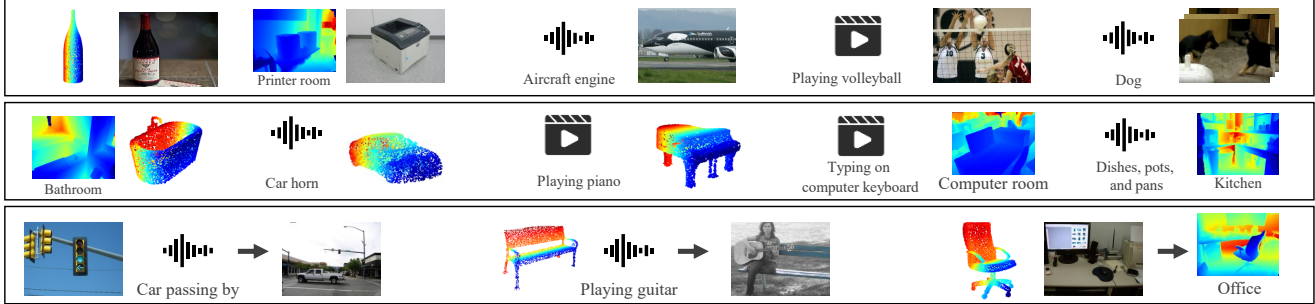


Figure 5. **Qualitative results for cross-modal retrieval.** *Top and middle:* retrieval using one single modality. *Bottom:* multiple modalities can be composed to retrieve targets across different modalities (§4.7).

#	Strategies	SUN-D	ModelNet40
1	Expert Grouping	45.8	55.9
2	Global Routing	45.0	55.3

Table 8. Ablation studies on modality-aware MoE strategies.

Top- k	#Experts	Init.	GPU Mem	SUN-D	ModelNet40
1	1	rand	27.2G	45.0	54.8
1	1	CLIP	27.2G	45.2	55.1
1	2	rand	27.4G	45.1	55.2
1	2	CLIP	27.4G	45.2	55.4
2	2	rand	35.9G	45.7	55.7
2	2	CLIP	35.9G	45.8	55.9

Table 9. Ablation studies on configuration of expert network. GPU memory is reported with a batch size of 448 on one Tesla A100.

expert network configurations. We investigate the impact of top- k routing, the number of experts per modality, and weight initialization. Starting with a baseline configuration of one expert per modality and random initialization, we achieve 45.0 on SUN-D test [74] and 54.8 on ModelNet40 test [5], with 27.2G GPU memory consumption. Increasing the expert number and initializing weights from the visual encoder of CLIP both lead to improved performance. Due to the constraints of GPU memory, we use two experts per modality and top-2 routing for expert selection. **Training Objectives.** Next we probe the impact of training objectives *i.e.*, soft modality binding with $\mathcal{L}_{\text{soft_bind}}$ (*cf.*, Eq. 11), as well as knowledge distillation through $\mathcal{L}_{\text{anch_distil}}$ (*cf.*, Eq. 12) and $\mathcal{L}_{\text{model_distil}}$ (*cf.*, Eq. 15). Results are summarized in Table 10, where the first row refers to the baseline that utilizes only paired data from single datasets and is trained with the standard InfoNCE loss. It can be observed that: **First**, our $\mathcal{L}_{\text{soft_bind}}$ successfully bridges the learning of multiple modalities for mutual boosting, leads to improved performance on both datasets. **Second**, $\mathcal{L}_{\text{anch_distil}}$ and $\mathcal{L}_{\text{model_distil}}$ could further boost the performance with compelling gains. This verifies the effectiveness of our knowledge distillation strategies, and leads UNIALIGN to a high performance model with optimized efficiency.

Modality Scaling. Finally, we examine the robustness of UNIALIGN to scale different numbers of modalities into a unified model. As shown in Table 11, modality-independent

\mathcal{L}_{NCE}	$\mathcal{L}_{\text{soft_bind}}$	$\mathcal{L}_{\text{anch_distil}}$	$\mathcal{L}_{\text{model_distil}}$	SUN-D	ModelNet40
✓	×	×	×	40.4	38.6
✓	✓	×	×	41.8	40.1
✓	✓	✓	×	42.3	40.9
✓	✓	✓	✓	45.8	55.9

Table 10. Ablation studies on training objectives.

#	Pre-Training Modality	ImageNet-1K	SUN-D	ModelNet40
1	I + D + P	84.7	34.4	31.6
2	I + T + D + P	84.9	39.9	38.1
3	I + T + D + P + V	85.1	40.3	38.4
4	I + T + D + P + V + A	85.1	40.4	38.6
5	I + T + D + P + V + A *	85.3	45.8	55.9

Table 11. Ablation studies on modality scaling. Results are reported using only \mathcal{L}_{NCE} as the default training loss. * denotes incorporate $\mathcal{L}_{\text{soft_bind}}$, $\mathcal{L}_{\text{anch_distil}}$, and $\mathcal{L}_{\text{model_distil}}$ into training. I: Image, T: Text, D: Depth, V: Video, A: Audio, P: 3D Point Cloud.

training with merely \mathcal{L}_{NCE} leads to a performance degeneration as more modalities integrated (*i.e.*, row #1-4). This suggests that a straightforward unification of multiple modalities may yield sub-optimal performance for individual ones. Nevertheless, soft modality binding and knowledge distillation (*i.e.*, row #4 vs. #5) could mitigate this negative impact, and boost performance to a level surpassing that achieved with fewer modalities (*i.e.*, row #1 vs. #5).

5. Conclusion

In this work, we focus on the all-in-one paradigm for multi-modal alignment, which yields a highly efficient and effective model, UNIALIGN, to unify the alignment of all anticipated modalities within one network. Inspired by MoE and knowledge distillation, we propose a modality-aware MoE to tackle multimodal inputs, and generally applicable alignment strategies to utilize unpaired data across datasets. Empirical results suggest that UNIALIGN achieves remarkable performance while being efficient in both training (7.8M trainable parameters to handle 6 modalities within a single training phase) and inference (one identical model with consistent weights across tasks). This work may potentially advance the development of compact, multimodal models.

Acknowledgement. This work was supported by the National Science and Technology Major Project (No. 2023ZD0121300), the National Natural Science Foundation of China (No. 62472222, 62302217), Natural Science Foundation of Jiangsu Province (No. BK20240080, BK20220936), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161), the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No. HMHAI-202403), and Bytedance Doubao Fund.

References

- [1] Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 3
- [2] Lucas Beyler, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *CVPR*, 2022. 2
- [3] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *CVPR*, 2024. 2
- [4] Rich Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997. 3
- [5] 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. 5, 7, 8
- [6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 5
- [7] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-MoLE: Sparse Mixture of LoRA Experts for Mitigating Data Conflicts in Instruction Finetuning MLLMs. *arXiv preprint arXiv:2401.16160*, 2024. 3
- [8] 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. 2, 5, 6
- [9] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *NeurIPS*, 2024. 4
- [10] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In *CVPR*, 2024. 6, 7
- [11] Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*, 2023. 3
- [12] Dapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-recognize networks for rgb-d scene recognition. In *CVPR*, 2019. 6
- [13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2
- [14] Bilal Faye, Hanane Azzag, and Mustapha Lebbah. Oneencoder: A Lightweight Framework for Progressive Alignment of Modalities. *arXiv preprint arXiv:2409.11059*, 2024. 2
- [15] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3
- [16] 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, 5, 6, 7
- [17] 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. 2, 5, 6
- [18] Rohit Girdhar, Alaeldin 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, 4, 6, 7
- [19] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 7
- [20] Yuan Gong, Cheng-I Lai, Yu-An Chung, and James R. Glass. Ssast: Self-Supervised Audio Spectrogram Transformer. In *AAAI*, 2022. 7
- [21] Yunhao Gou, Zhili Liu, Kai Chen, Lanqing Hong, Hang Xu, Aoxue Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Mixture of cluster-conditional lora experts for vision-language instruction tuning. *arXiv preprint arXiv:2312.12379*, 2023. 3
- [22] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022. 1, 5, 7
- [23] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One Framework to Align All Modalities with Language. In *CVPR*, 2024. 2
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [25] Byeongho Heo, Jeessoo Kim, Sangdoon Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 3
- [26] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3
- [27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 2, 3
- [28] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 4

- [29] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023. 3
- [30] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *ICCV*, 2023. 6
- [31] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. 3
- [32] Byoungjip Kim, Sungik Choi, Dasol Hwang, Moontae Lee, and Honglak Lee. Transferring pre-trained multimodal representations with cross-modal similarity matching. In *NeurIPS*, 2022. 3
- [33] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 6, 7
- [34] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, 2018. 3
- [35] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *ICLR*, 2023. 3
- [36] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 3
- [37] Konrad P Kording, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007. 1
- [38] 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, 5, 6, 7
- [39] 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. 1, 2, 3
- [40] Po-han Li, Sandeep P Chinchali, and Ufuk Topcu. Csa: Data-efficient mapping of unimodal features to multimodal features. *arXiv preprint arXiv:2410.07610*, 2024. 2
- [41] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *arXiv*, abs/2405.11273, 2024. 3
- [42] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021. 3, 5
- [43] Bin Lin, Zhenyu Tang, Yang Ye, Jiayi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3
- [44] Chung-Ching Lin, Kevin Lin, Lijuan Wang, Zicheng Liu, and Linjie Li. Cross-modal representation learning for zero-shot action recognition. In *CVPR*, 2022. 7
- [45] Haokun Lin, Haoli Bai, Zhili Liu, Lu Hou, Muye Sun, Linqi Song, Ying Wei, and Zhenan Sun. Mope-CLIP: Structured Pruning for Efficient Vision-Language Models with Module-wise Pruning Error Metric. In *CVPR*, 2024. 3
- [46] Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Sriniwasan Iyer, Mike Lewis, Gargi Gosh, Luke Zettlemoyer, and Armen Aghajanyan. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*, 2024. 3
- [47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 6
- [48] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 6
- [49] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2
- [50] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-Augmented Unified and Balanced Representation Space to Bind Them All. In *CVPR*, 2024. 2, 7
- [51] Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. Ms2sl: Multimodal spoken data-driven continuous sign language production. In *ACL Findings*, 2024. 2
- [52] Wenxuan Ma, Shuang Li, Jinming Zhang, Chi Harold Liu, Jingxuan Kang, Yulin Wang, and Gao Huang. Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm. In *ICCV*, 2023. 2
- [53] M Alex Meredith and Barry E Stein. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, 56(3):640–662, 1986. 1
- [54] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 7
- [55] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020. 3
- [56] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *ICLR*, 2023. 2
- [57] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 7

- [58] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 7
- [59] Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. In *NeurIPS*, 2023. 2
- [60] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 7
- [61] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *ECCV*, 2022. 2
- [62] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *CVPR*, 2024. 2
- [63] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, 2023. 3
- [64] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, 2021. 6
- [65] Karol J Piczak. Esc: Dataset for environmental sound classification. In *ACM MM*, 2015. 6, 7
- [66] 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. 6
- [67] 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 *ICML*, 2021. 2, 3, 5
- [68] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One. In *CVPR*, 2024. 3, 5
- [69] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, 2021. 3
- [70] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, 2023. 6
- [71] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling Diverse Vision Foundation Models for Robot Learning. In *CoRL*, 2024. 3
- [72] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 2, 3
- [73] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 6
- [74] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5, 6, 7, 8
- [75] Xinhang Song, Luis Herranz, and Shuqiang Jiang. Depth cnns for rgb-d scene recognition: Learning from scratch better than transferring from rgb-cnns. In *AAAI*, 2017. 6
- [76] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen. Image representations with spatial object-to-object relations for rgb-d scene recognition. *IEEE TIP*, 29:525–537, 2019. 6
- [77] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7
- [78] Siddharth Srivastava and Gaurav Sharma. Omnivec2 - A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning. In *CVPR*, 2024. 2
- [79] Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm : Distilling multimodal and efficient foundation models. In *ICCV*, 2023. 3
- [80] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 6
- [81] 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. 4
- [82] 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. 6
- [83] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE TPAMI*, 44(7):3614–3633, 2021. 3
- [84] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *CVPR*, 2024. 3
- [85] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Action-clip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 7
- [86] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 5
- [87] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu,

- Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, 2023. 6
- [88] Wenguan Wang, Yi Yang, and Yunhe Pan. Visual knowledge in the big model era: Retrospect and prospect. *Frontiers of Information Technology & Electronic Engineering*, 26(1):1–19, 2025. 2
- [89] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 7
- [90] Zehan Wang, Ziang Zhang, Luping Liu, Yang Zhao, Haifeng Huang, Tao Jin, and Zhou Zhao. Extending multi-modal contrastive representations. *arXiv preprint arXiv:2310.08884*, 2023. 2
- [91] Zehan Wang, Yang Zhao, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. In *NeurIPS*, 2023. 2
- [92] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. 7
- [93] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *Tech Report*, 2022. 3
- [94] 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. 6
- [95] 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, 6
- [96] 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, 7
- [97] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE TPAMI*, 45(10):12113–12132, 2023. 1
- [98] 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, 5, 6
- [99] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-KD: An Empirical Study of CLIP Model Distillation. In *CVPR*, 2024. 3, 5
- [100] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024. 2
- [101] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *CVPR*, 2021. 2
- [102] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, 2021. 2
- [103] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. Is-fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *ECCV*, 2024. 2
- [104] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *KDD*, 2017. 3
- [105] Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *AAAI*, 2021. 3
- [106] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, 2022. 6
- [107] 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. 2, 3, 6, 7
- [108] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. In *NeurIPS*, 2022. 2
- [109] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiayi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *ICLR*, 2024. 1, 2, 7
- [110] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018. 2
- [111] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, 2023. 6
- [112] Konrad Zuchniak. Multi-teacher knowledge distillation as an effective method for compressing ensembles of neural networks. *arXiv preprint arXiv:2302.07215*, 2023. 3