

Decoupled and Reusable Adaptation for Efficient Cross-Modal Transfer

Yajing Liu^{1,2}, Yumeng Zhang³, Yue Si³, Baojie Fan⁴, Jiandong Tian^{1*}

¹State Key Laboratory of Robotics and Intelligent Systems,
 Shenyang Institute of Automation, Chinese Academy of Sciences

²University of Chinese Academy of Sciences ³ Shenyang Ligong University

⁴Nanjing University of Posts and Telecommunications

{liuyajing,tianjd}@sia.cn, {zhangyumeng,siyue}@stu.sylu.edu.cn, jobfbj@gmail.com

Abstract

Cross-modal transfer methods have achieved significant progress in extending RGB-based foundation models to non-RGB modalities. However, existing transfer paradigms are primarily task-oriented, meaning that changing tasks requires re-training and re-storing, leading to substantial redundancy in data, computation and storage. To address this limitation, we propose an efficient cross-modal transfer paradigm that decouples the process into a one-time general modality knowledge transfer and a flexible task knowledge transfer. In Stage 1, we propose a Progressive Self-Supervised Tuning strategy that integrates modality-aware structural reconstruction with semantic discriminative learning, which enables task-agnostic modality knowledge learning using only unlabeled data through a one-time training process, resulting in reusable target-modality LoRAs. In Stage 2, we incorporate the modality LoRAs and further propose a Task-Prompted Mixture-of-Modality Experts module. This design enables lightweight task knowledge injection while effectively balancing task-specific, modality-general and modality-specific knowledge in multimodal fusion process for diverse downstream tasks. Extensive experiments across six cross-modal transfer scenarios, along with analyses of data, computation, and storage efficiency, demonstrate the superiority of our method.

1. Introduction

Multimodal data such as thermal, depth, event, and LiDAR provide complementary information beyond RGB, making them increasingly valuable for various downstream vision applications [6, 20, 29, 44]. However, due to the scarcity of annotated data and the imbalance between these modalities and the easily accessible RGB data, large-scale foundation models in these domains remain underdeveloped. Conse-

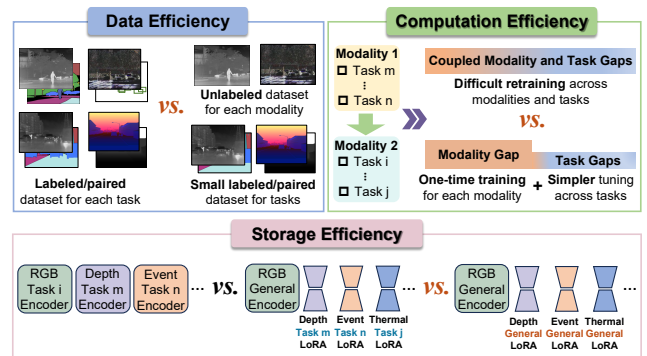


Figure 1. Comparison between existing cross-modal transfer strategies and ours. Strategies after “vs.” denote ours. Existing methods perform task-oriented transfer, requiring retraining with labeled task data whenever the task or modality changes. And each trained task model is stored separately. In contrast, our method decouples the cross-modal transfer process by leveraging large easily accessible unlabeled modality data to bridge the modality gap, resulting in a general modality LoRA for storage. When switching tasks, smaller labeled data is needed for simple tuning.

quently, existing approaches typically rely on training task-specific models for each modality and downstream task.

Given that these sensor modalities share certain visual priors with RGB, recent studies [21, 26, 48, 54, 59] have explored cross-modal transfer techniques to adapt RGB foundation models to non-RGB modalities. For example, Zhao et al. [59] adopted a parameter-efficient adapter strategy to transfer SAM 2 [37] into an RGB–Thermal semantic segmentation model under textual guidance. And Yu et al. [54] proposed the DSAM model to explore the interaction between RGB and depth modalities, enabling the SAM [19] model to be adapted for depth-based camouflaged object detection. While these efforts have achieved promising results, they still follow a task-oriented transfer paradigm, where task-specific labeled data are directly used for supervised training to adapt the RGB foundation model to a specific target-modality task. Consequently, switching to a new

*Corresponding author

task requires collecting additional labeled data, retraining the model from scratch and storing it separately, overlooking the shared modality knowledge across tasks and leading to substantial label, computation, and storage costs.

To overcome these limitations, we aim to achieve efficient cross-modal transfer that maintains high data, computation, and storage efficiency when adapting to downstream non-RGB domains. Building upon the above discussion, we make two key observations regarding the transfer from RGB to non-RGB modalities.

First, the task-oriented transfer paradigm in existing cross-modal transfer methods [3, 21, 26, 54, 59] couples the learning of modality and task gaps within a single adaptation process. Consequently, when adapting to a new downstream task, the model must be retrained from scratch, resulting in substantial computation and storage costs. However, this paradigm overlooks an important insight. Although downstream tasks may differ, the modality gap that dominates cross-modal discrepancies is largely shared across tasks, which can be addressed through a one-time fine-tuning process and the learned modality adaptation can be reused across tasks via a unified module, thereby avoiding redundant retraining. Therefore, developing more efficient training strategies and transfer paradigms is essential for improving computation and storage efficiency.

Second, cross-modal transfer from RGB foundation models involves both a modality gap and a task gap, with the modality gap being the dominant challenge [32, 46]. Therefore, relying solely on task-specific labeled data is insufficient to simultaneously bridge both gaps, especially when labeled samples for a specific task are scarce. Although target-modality data from different tasks have heterogeneous annotations, they collectively capture the modality-level intrinsic characteristics and underlying distribution of the target modality. Thus, aggregating such data from different tasks into a large unlabeled dataset is crucial for learning task-agnostic modality knowledge. Therefore, effective cross-modal transfer must leverage unlabeled data to enhance data efficiency.

Based on the above analysis, we propose a novel cross-modal transfer paradigm, which decouples the transfer process into a one-time general modality knowledge transfer (Stage 1) and a flexible task knowledge transfer (Stage 2). In Stage 1, the key challenge lies in designing an effective training strategy that can learn general target-modality knowledge without relying on task labels. To address this, inspired by the success of self-supervised learning (SSL) in visual representation learning, we introduce a Progressive Self-Supervised Tuning (PSST) strategy, which leverages a key property of non-RGB modalities: their structural and radiometric patterns inherently encode semantic cues. Specifically, we introduce MAE-based [14] structural reconstruction and DINOv2-based [33] semantic dis-

criminative learning, and integrate the two SSL techniques through the aforementioned property to progressively learn modality knowledge from low-level textures and structures to high-level semantics with only unlabeled data. Stages 1 is performed only one-time, and the trained target-modality LoRA is stored for reuse across different downstream tasks in Stage 2. In Stage 2, to enable flexible, rapid and robust task-level adaptation, we propose a Task-Prompted Mixture-of-Modality Experts (TP-MoME), which injects task-specific knowledge into the model via controllable and lightweight visual prompts and performs multimodal fusion using a MoME design, effectively balancing task-specific, modality-general, and modality-specific knowledge during the fusion process for diverse downstream tasks.

In the experiments, our method achieves superior performance across six cross-modal transfer scenarios. Besides, compared with the SOTA cross-modal transfer method, our method improves 6.2% mIoU under 20% labeled data setting, reduces training by 9 epochs to reach the same performance threshold, and enhances 44.3% efficiency on trainable parameters for the same number of downstream tasks.

Our main contributions are summarized as follows:

- (1) We explore the efficiency problem of cross-modal transfer, and propose a novel paradigm that decouples the transfer process into a one-time general modality knowledge transfer and a flexible task knowledge transfer, achieving high data, computation, and storage efficiency.
- (2) We propose a Progressive Self-Supervised Tuning strategy that enables task-agnostic modality knowledge learning using only unlabeled data through one-time training, and acquiring reusable target-modality LoRAs.
- (3) We propose a Task-Prompted Mixture-of-Modality Experts module that achieves lightweight task knowledge learning while effectively balancing task-specific, modality-general, and modality-specific knowledge in multimodal fusion process for diverse downstream tasks.
- (4) We evaluate our method on six cross-modal transfer scenarios and analyze the data, computation, and storage efficiency to demonstrate its effectiveness and superiority.

2. Related Works

2.1. VFMs and Transfer Learning

Vision foundation models (VFMs) pre-trained on large-scale RGB data have demonstrated remarkable generalization capabilities across diverse visual tasks. SAM [19] and its video extension SAM 2 [37] demonstrate strong zero-shot segmentation capabilities through promptable architectures trained on billion-scale mask data. Beyond segmentation, models such as CLIP [36] and DINOv2 [33] provide powerful feature representations for cross-modal understanding and dense prediction. To effectively adapt these models to downstream tasks, parameter-efficient transfer

learning techniques have gained significant attention, such as LoRA [2, 24, 53], adapter [4, 10, 47, 58], and prompt tuning [18, 21, 55, 62], which allow for lightweight adaptation of frozen backbones with minimal additional parameters. Recent studies have successfully applied such techniques to VFMs, achieving strong performance on segmentation [21, 30, 47, 49], detection [12, 38, 60], and classification [1, 41, 58] tasks while avoiding full model retraining.

2.2. Cross-Modal Transfer and Adaptation

Cross-modal transfer extends RGB-based models to non-RGB modalities such as depth, thermal, event. Recent studies [3, 9, 15, 26, 28, 34, 35, 47, 54, 59, 63, 64] increasingly incorporate pretrained VFMs for cross-modal tasks. For example, Cai et al. [3] introduced modality-aware prompting to SAM [19] for RGB+X segmentation, while DSAM [54] adapted SAM [19] to depth for camouflaged object detection. Besides, SHIFNet [59] employed language guidance to adapt SAM 2 [37] for RGB-T segmentation. Beyond SAM, RGB-Th-Bench [31] explored RGB-T understanding in vision-language models (VLMs) and VS-TDX [5] analyzed the limited ability of VLMs to interpret non-RGB sensor data and proposed sensor-aware fine-tuning. Despite these advancement, existing methods are largely task-oriented, training models end-to-end for specific tasks using labeled data. This paradigm requires re-training and re-storing model parameters for each new task, and leads to significant redundancy in annotation, computation, and storage, limiting the scalability and practical deployment of cross-modal foundation models. In this work, we aim to develop an efficient paradigm for cross-modal transfer.

2.3. Mixture-of-Experts Methods

Mixture-of-Experts (MoE) architectures have recently gained increasing attention in computer vision for their ability to scale model capacity efficiently. Vision MoE [40] pioneered sparse expert layers in Vision Transformers, and subsequent studies [7, 50, 51] extended MoE to dense prediction tasks, where experts specialize in distinct visual patterns or semantic regions for multitask learning. Besides, MoE has been applied in text-related multimodal learning [13, 16, 23, 63]. For example, Uni-MoE [23] investigated the visual question answering (VQA) domain, modeling interactions among speech, audio, and text modalities, while FuseMoE [13] and CL-MoE [16] focused on VQA based on RGB image data. For non-RGB vision tasks, MLE-SAM [63] regarded the LoRAs for different modalities as expert models and employed a gating mechanism for fusion in segmentation. However, this is essentially the same as non-MoE multimodal fusion. In this work, we propose TP-MoME, a novel design tailored for cross-modal transfer, which leverages the unique advantages of MoE to achieve flexible multimodal fusion

3. Proposed Method

3.1. Overview

In this work, we propose a two-stage training pipeline that decouples the cross-modal transfer process into a one-time general knowledge transfer and a flexible task knowledge transfer to achieve efficient transfer learning. The overall framework is illustrated in Figure 2. Formally, let $\mathcal{X} = \{X_m | m = 1, \dots, M\}$ denote a large collection of unlabeled target-modality data gathered from the Internet, where m indexes the target modality. In Stage 1 (Sec. 3.2), the Progressive Self-Supervised Tuning (PSST) strategy jointly leverages MAE-based structural reconstruction and DINOv2-style semantic alignment on $\mathcal{X}$, progressively capturing low-level modality-aware structures and high-level semantic representations. For each target modality, Stage 1 is trained only once to obtain a reusable modality LoRA. In Stage 2 (Sec. 3.3), given labeled downstream task samples (X, Y) , modality-specific LoRAs are deployed according to the task requirements. Learnable visual prompts P are inserted into each block to acquire task knowledge, while the Mixture-of-Modality-Experts (MoME) module is incorporated into designated blocks to enable multimodal fusion. The fused embeddings or single-modal embeddings are fed into multimodal or single-modal task decoder.

3.2. Progressive Self-supervised Tuning Strategy

To enable label-free cross-modal knowledge transfer, we collect diverse target-modality data spanning both low- and high-level tasks, where inconsistent task semantics make unified supervised training infeasible. Motivated by the capability of self-supervised learning (SSL) to capture robust representations from diverse data, we propose a Progressive Self-supervised Tuning (PSST) strategy.

However, not all SSL tasks are suitable for non-RGB modality learning. The most widely used contrastive learning objectives are designed based on RGB priors, such as enforcing brightness and texture invariance across images. In contrast, changes in brightness or texture often correspond to changes in semantic content for non-RGB modalities such as thermal, depth, or event. Based on this observation, we argue that MAE-style single-image reconstruction and DINOv2-style self-teaching are more appropriate for non-RGB modality learning and incorporate them to gradually learn modality-aware knowledge from low-level textures and structures to high-level semantics.

MAE warm-up training. In warm-up stage, we perform a modality-aware masked autoencoding pretraining to capture the intrinsic structure of the target modality. Given an unlabeled non-RGB image $x_m \in X_m$, a mask generator $G(\cdot)$ selects a subset of patches to be visible. The masked input $\tilde{x}_m = x_m \odot G(x_m)$ is encoded through the frozen

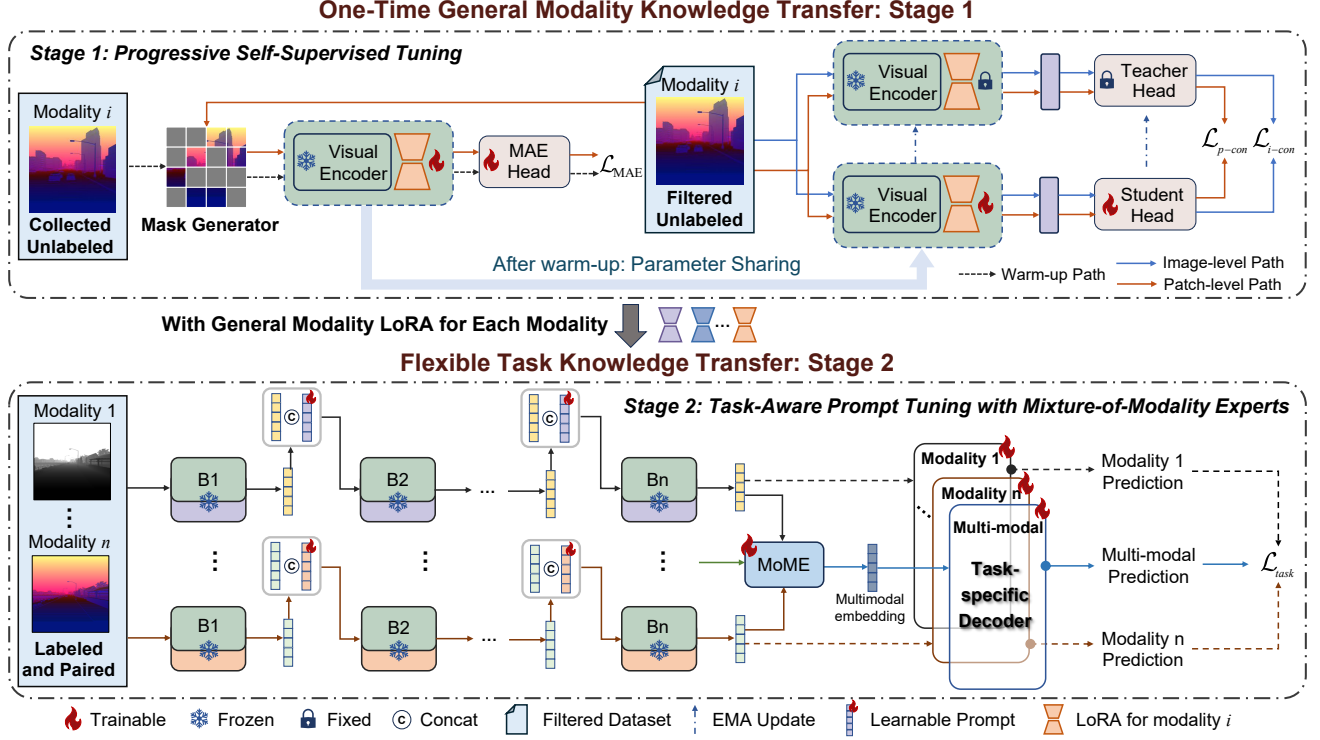


Figure 2. Overview of the proposed framework. In Stage 1, given the collected unlabeled data for Modality i , Progressive Self-Supervised Tuning (PSST) leverages MAE-based structure reconstruction to warm-up the modality LoRA with $\mathcal{L}_{MAE}$. Subsequently, with the filtered unlabeled data, PSST integrated MAE-based reconstruction with $\mathcal{L}_{MAE}$ and DINOv2-style semantic learning with $\mathcal{L}_{i-con}$ and $\mathcal{L}_{p-con}$ for unified training with sharing LoRA parameters to progressively learn low-level and high-level modality knowledge. With the frozen modality LoRAs, in Stage 2, Task-aware Prompt Tuning introduces learnable prompts in each block for each target modality to capture task knowledge and the proposed MoME module is deployed after specific blocks to integrate task-prompted multimodal embeddings for fusion. The hierarchical outputs of MoME or single-modality embeddings from different blocks are fed into the task decoder.

backbone E_{θ_0} augmented by the modality LoRA $\Delta\theta$:

$$z_m = E_{\theta_0 + \Delta\theta}(\tilde{x}_m). \quad (1)$$

The MAE head D_{MAE} reconstructs the original image, and the reconstruction objective is:

$$\mathcal{L}_{MAE} = \frac{1}{|G|} \sum_{i \in G} \|D_{MAE}(z_m)_i - x_{m,i}\|_1. \quad (2)$$

This training is conducted for several epochs to encourage $\Delta\theta$ to internalize modality-specific local structure.

Following the warm-up phase, we select a filtered subset of unlabeled data $\mathcal{D}_{high,m}$ with high structural complexity or informative semantics (reflected by high MAE reconstruction variance). On this subset, we jointly train the MAE and DINOv2 branches, where the former provides structural supervision and the latter focuses on semantic alignment. And the MAE branch and the student network in DINOv2 branch share the same frozen backbone and trainable LoRA parameters during training, which are updated by gradient back propagation. The teacher network in DINOv2 branch

is fixed and its weights are updated by Exponential Moving Average (EMA) of the student weights.

Image-level consistency learning. For each sample in $\mathcal{D}_{high,m}$, two augmented views are generated and passed into the teacher (T) and student (S) networks. We encourage the student to align with the semantic distribution of the teacher formed by the class-token representation. The student and teacher heads [33] map their class tokens into normalized teacher and student semantic distributions p_t and p_s , respectively, which are also called prototype scores. The image-level consistency objective is defined as:

$$\mathcal{L}_{i-con} = - \sum p_t \log p_s. \quad (3)$$

This image-level constraint enforces global semantic coherence within the target modality.

MAE-guided patch-level consistency. To enhance local and fine-grained discriminability, we introduce a patch-level consistency term guided by the MAE’s reconstruction errors $\{e_i = \|D_{MAE}(z_m)_i - x_{m,i}\|_1\}$, which are retained as patch-level structural uncertainty indicators, serving as

guidance weights for the semantic learning. For each image, a subset of patches is masked in the student input but visible to the teacher. The student is encouraged to infer the semantic embedding of masked regions by aligning with the teacher’s visible patch descriptors, with larger MAE reconstruction errors receiving higher supervision weights:

$$\mathcal{L}_{p\text{-con}} = - \sum_i w_i p_{t,i} \log p_{s,i}, \quad (4)$$

where $w_i = \frac{\exp(\alpha e_i)}{\sum_j \exp(\alpha e_j)}$ is a normalized weighting function reflecting structural uncertainty derived from MAE reconstruction error e_i . This adaptive weighting ensures that DINOv2 emphasizes patches where the MAE struggled most, effectively steering the model toward modality-specific semantic cues that deviate from RGB priors.

The training objective of DINOv2 branch is:

$$\mathcal{L}_{\text{DINO}} = \mathcal{L}_{i\text{-con}} + \mathcal{L}_{p\text{-con}} + \mathcal{L}_{\text{koleo}}, \quad (5)$$

where $\mathcal{L}_{\text{koleo}}$ is the regularizer used in original DINOv2 [33] to encourage a uniform span of the features within a batch.

Joint reconstruction and discriminative learning. The final PSST loss integrates structural reconstruction and semantic alignment:

$$\mathcal{L}_{\text{PSST}} = \lambda \mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{DINO}}. \quad (6)$$

where λ balances reconstruction and semantic alignment.

This progressive design first learns modality-sensitive structural priors via MAE, then aligns high-level semantics through DINOv2-style learning, which enables the LoRA-augmented backbone to evolve from structure perception to semantic understanding, achieving robust cross-modal adaptation without label supervision.

3.3. Task-Prompted Mixture-of-Modality Experts

To leverage the above modality LoRA for subsequent downstream tasks while enabling rapid and robust adaptation with minimal parameter overhead, we propose a Task-Prompted Mixture-of-Modality Experts (TP-MoME) module, which achieves lightweight task knowledge injection via prompting and guides a flexible Mixture-of-Modality Experts (MoME) to effectively balance modality generality and specificity during multimodal fusion.

Specifically, we introduce a set of trainable task prompts $\{P_i\}_{i \in \mathcal{I}}$ that are inserted into each Transformer block ($i \in \mathcal{I}$) to capture task-specific priors. Given the frozen encoder with the Stage 1 trained frozen LoRA $E_{\theta_0 + \Delta\theta}$, the patch token embedding of modality m at block i is denoted as $R^{(m,i)} = E_{\theta_0^{i-1} + \Delta\theta^{i-1}}(R^{(m,i-1)})$, $R^{(m,i)} \in \mathbb{R}^{T \times d}$. The trainable task prompts $P_i^m \in \mathbb{R}^{L_p \times d}$ are concatenated with the patch tokens to form the input sequence to block i :

$$\tilde{R}^{(m,i)} = \text{concat}(X_0, P_i^m, R^{(m,i)}), \quad (7)$$

where $X_0 \in \{[\text{CLS}], \emptyset\}$ is the encoder-dependent class token. $\tilde{R}^{(m,i)}$ is then forwarded through the frozen transformer block i to output the updated features $R_{\text{prompt}}^{(m,i)}$.

To achieve flexible and robust cross-modal integration, the proposed MoME module introduces a set of learnable expert networks. MoME consists of one generalizable expert E_0 and N modality-specific experts, each implemented as an independent FFN with layer norm and focus on distinct aspects of multimodal features. Suppose that block i is one of the fusion layers, the input features $\hat{R}_{\text{cat}}^{(i)}$ formed from projecting the concatenated prompt-updated features of each modality:

$$R_{\text{cat}}^{(i)} = \text{Concat}(R_{\text{prompt}}^{(1,i)}, R_{\text{prompt}}^{(2,i)}, \dots, R_{\text{prompt}}^{(M,i)}), \quad (8)$$

$$\hat{R}_{\text{cat}}^{(i)} = \text{Proj}(R_{\text{cat}}^{(i)}) \quad (9)$$

where $\text{Proj}(\cdot)$ is a one-layer MLP, and $\hat{R}_{\text{cat}}^{(i)}$ is a lower-dimension representation of multimodal input.

During training, each modality-specific expert receives a randomly masked version of this projected multimodal feature to encourage specialization for different modality configurations:

$$\tilde{R}_{\text{cat},n}^{(i)} = \text{Mask}_n(\hat{R}_{\text{cat}}^{(i)}), \quad (10)$$

where $\text{Mask}_n(\cdot)$ randomly drops one or more segments for the n -th expert during each iteration, forcing it to focus on different subsets of multimodal inputs and learn complementary representations. The generalizable expert, by contrast, observes the full projected input without masking.

Each modality-specific expert $E_{1,\dots,N}(\cdot)$ maps the masked input to an expert-specific output:

$$Z_n^{(i)} = E_n(\tilde{R}_{\text{cat},n}^{(i)}), n = 1, \dots, N. \quad (11)$$

A gating network g_{θ_g} dynamically predicts the mixture weights $\beta^{(i)} = [\beta_0^{(i)}, \dots, \beta_N^{(i)}]$ from the pooled concatenated multimodal features before projection,

$$\beta^{(i)} = \text{softmax}(W_g(\text{Pool}(R_{\text{cat}}^{(i)})) + b_g). \quad (12)$$

The fused feature of MoME at block i is then computed as a weighted aggregation:

$$R_{\text{MoME}}^{(i)} = \hat{R}_{\text{cat}}^{(i)} + \beta_0^{(i)} E_0(\hat{R}_{\text{cat}}^{(i)}) + \sum_{n=1}^N \beta_n^{(i)} Z_n^{(i)}. \quad (13)$$

The base projected feature is assigned a fixed weight of 1, ensuring a stable anchor during training and inference.

The fused features of MoME at specific blocks B_{fuse} form a hierarchical embedding set $R_{\text{MoME}} = \{R_{\text{MoME}}^{(i)} | i \in B_{\text{fuse}}\}$ and are fed into task-specific decoder D_{task} :

$$\hat{Y} = D_{\text{task}}(R_{\text{MoME}}). \quad (14)$$

The model is trained with $\mathcal{L}_{\text{TP-MoME}} = \mathcal{L}_{\text{task}}(Y, \hat{Y})$.

4. Experiments

4.1. Experimental Setup

Datasets. For downstream performance evaluation, we conduct experiments on the following datasets: **(1) RGB-Thermal:** MFNet [11], PST900 [42]; **(2) RGB-Depth:** SUN-RGBD [45], NYU Depth v2 [43], DELIVER [57]; **(3) RGB-Polarization:** MCubeS [25]; **(4) RGB-Event:** DELIVER [57]; **(5) RGB-Depth-Event:** DELIVER [57]. The details of the sources and sizes of unlabeled datasets for Stage 1 are provided in the *Supplementary Material*.

Implementation Details. We employ Hiera-B+/L encoders as the pretrained RGB-based vision foundation models and the weights are initialized from SAM 2. In Stage 1 training, we inject LoRA modules into each self-attention layers of the encoder and the rank is set to 32. The batch size is 12, optimizer is AdamW with an initial learning rate of $3e-4$, Warmup Polynomial scheduler, and weight decay of 0.01. In Stage 2 training, the prompts are injected in each block and are the same dimension as the embedding dimension of each block of the encoder. The MoME module are inserted in blocks [2, 5, 21, 24] and [2, 8, 44, 48] for Hiera-B+ and Hiera-L respectively. More detailed training settings are presented in the *Supplementary Material*.

4.2. Comparison with State-of-the-arts

We conduct extensive Stage 2 tuning with the trained LoRAs for diverse downstream scenarios and compare the performance of our method with that of other methods, including **modality-specific full fine-tuning (FFT) models** such as CMX [56], CMNeXt [57], MaskMentor [61], DFormer [52], SegMiF [27], MMSFormer [39], StitchFusion [22], GeminiFusion [17] and **task-oriented parameter-efficient fine-tuning (PEFT) methods**, including DPLNet [8], GoPT [15], SPEFT [3], CWSAM [34], SAM-LoRA [63], MLE-SAM [63], MemorySAM [26].

SUN-RGBD dataset [45] contains 10,335 RGB-D images with 37 classes. As shown in Table 1, our model with Hiera-L achieves the best 56.3% mIoU, surpassing the FFT method GeminiFusion [17] by 1.7% mIoU and the PEFT method SPEFT [3] (Swin-L) by 1.3% mIoU. Our model with Hiera-B+ also achieves comparable result with 54.8% mIoU, surpassing SPEFT [3] with Swin-B by 1.1% mIoU.

NYU Depth V2 dataset [43] is an indoor scene understanding dataset consisting of 1,449 RGB-D samples with 40 categories. As shown in Table 1, our model with Hiera-L encoder achieves the best performance of 61.4% mIoU, surpassing the SOTA FFT method GeminiFusion [17] by 1.2% with more efficient tuning. Compared to the SOTA PEFT methods, our model with Hiera-B+ surpasses SPEFT [3] with Swin-B by 1.2% mIoU and our Hiera-L version surpasses its Swin-L version by 1.5% mIoU.

MFNet dataset [11] is an RGB-T urban street dataset

Table 1. Performance comparison of RGB-Depth downstream tasks on SUN-RGBD and NYU Depth v2 in terms of mIoU (%).

Method	Backbone	SUN	NYUv2
<i>Full Fine-Tuning Methods</i>			
CMXTITS23 [56]	MiT-B5	52.4	56.9
CMNeXtCVPR23 [57]	MiT-B4	50.4	56.9
DFormerICLR24 [52]	DFormer-L	52.5	57.2
GeminiFusionICML24 [17]	Swin-L	54.6	60.2
MaskMentorACMMM25 [61]	ViT-B	53.0	57.9
<i>Parameter-Efficient Tuning Methods</i>			
DPLNetIROS24 [8]	MiT-B5	52.8	59.3
GoPTAAAI24 [15]	ViT-B	52.1	54.1
SPEFTCVPR25 [3]	Swin-B	53.7	59.0
Ours	Hiera-B+	54.8	60.2
GoPTAAAI24 [15]	ViT-L	52.3	54.3
SPEFTCVPR25 [3]	Swin-L	55.0	59.9
Ours	Hiera-L	56.3	61.4

Table 2. Performance comparison of RGB-Thermal downstream tasks on MFNet and PST900 in terms of mIoU (%).

Method	Backbone	MFNet	PST900
<i>Full Fine-Tuning Methods</i>			
SegMiFICCV23 [27]	MiT-B3	56.1	-
CMXTITS23 [56]	MiT-B4	59.7	-
CMNeXtCVPR23 [57]	MiT-B4	59.9	86.9
MMSFormerOJSP24 [39]	MiT-B4	-	87.5
StitchFusionACMMM25 [22]	MiT-B4	57.8	84.7
<i>Parameter-Efficient Tuning Methods</i>			
DPLNetIROS24 [8]	MiT-B5	59.3	86.7
GoPTAAAI24 [15]	ViT-B	57.4	81.3
GoPTAAAI24 [15]	ViT-L	57.7	81.5
SPEFTCVPR25 [3]	Swin-B	59.9	87.6
Ours	Hiera-B+	60.5	88.7

for segmentation, consisting of 1,569 image pairs with 8 classes. As shown in Table 2, compared to the PEFT method SPEFT [3] with Swin-B, our model with Hiera-B+ achieves the best 60.5% mIoU. However, SPEFT [3] is limited to RGB+X bimodal tasks and exhibits more bias to specific tasks. Compared to FFT methods CMX [56] and CMNeXt [57], our performance remains promising.

PST900 dataset [42] contains 894 RGB-T image pairs with 5 classes. As shown in Table 2, compared to PEFT methods, our model with Hiera-B+ achieves the best 88.7% mIoU, surpassing DPLNet [8] by 2.0% mIoU and SPEFT [3] by 1.1% mIoU. Compared to FFT methods, our method achieves 1.2% and 4.0% mIoU improvements over MMSFormer [39] and StitchFusion [22] respectively.

MCubeS dataset [25] contains RGB-NIR-DoLP-AoLP image pairs for segmentation with 20 classes. We use RGB, DoLP and AoLP for evaluation. As shown in Table 3, for RGB-A segmentation, our method achieves the best 52.8%

Table 3. Performance (%) comparison of RGB-Polarization downstream tasks on MCubeS dataset.

Method	Backbone	Modality	mIoU
CMNeXtCVPR23 [57]	MiT-B2	RGB-A	48.4
MMSFormerOJSP24 [39]	MiT-B4	RGB-A	51.3
StitchFusionACMMM25 [22]	MiT-B4	RGB-A	52.1
CWSAMJSTARS25 [34]	ViT-B	RGB-A	49.8
SAM-LoRA _{arXiv24} [63]	ViT-B	RGB-A	48.7
MLE-SAM _{arXiv24} [63]	Hiera-B+	RGB-A	50.6
MemorySAM _{arXiv25} [26]	Hiera-B+	RGB-A	51.2
Ours	Hiera-B+	RGB-A	52.8
CMNeXtCVPR23 [57]	MiT-B2	RGB-A-D	49.5
MMSFormerOJSP24 [39]	MiT-B4	RGB-A-D	52.0
StitchFusionACMMM25 [22]	MiT-B4	RGB-A-D	52.5
CWSAMJSTARS25 [34]	ViT-B	RGB-A-D	48.4
SAM-LoRA _{arXiv24} [63]	ViT-B	RGB-A-D	49.4
MLE-SAM _{arXiv24} [63]	Hiera-B+	RGB-A-D	50.9
MemorySAM _{arXiv25} [26]	Hiera-B+	RGB-A-D	52.9
Ours	Hiera-B+	RGB-A-D	54.4

mIoU, surpassing MemorySAM [26] by 1.6% mIoU and StitchFusion [22] by 0.7% mIoU. For RGB-A-D segmentation, our method achieves the best 54.4% mIoU, with an improvement of 1.5% mIoU over MemorySAM [26].

DELIVER dataset [57] contains large-scale Depth-LiDAR-Event-RGB data pairs with 25 classes for segmentation. As shown in Table 4, for RGB-Event segmentation, our method achieves the best 59.5% mIoU. For RGB-Depth segmentation, we achieves the best 64.9% mIoU, surpassing FFT method CMNeXt [57] by 1.3% mIoU and also surpassing the SOTA PEFT method MLE-SAM [63] by 1.3% mIoU. For RGB-Depth-Event segmentation, our model achieves the superior 65.1% mIoU, surpassing CMNeXt [57] by 0.7% mIoU and the SOTA PEFT method MLE-SAM [63] by 2.4% mIoU.

4.3. Ablation Studies

In this section, we conduct several experiments on MCubeS [25] and DELIVER [57] datasets to analyze the effectiveness of each component in our framework.

Analysis of PSST strategy. As shown in Table 5, we first remove the general modality knowledge transfer process and train the model using only Stage 2 with the proposed TP-MoME, and the performance drops from 59.3% to 54.9% mIoU, with an average drop of 4.4% mIoU, highlighting the crucial role of our PSST strategy in bridging the modality gap. Moreover, as shown in Figure 4, solely the TP-MoME training (Stage 2) results in slower performance improvement compared to the full model, indicating that the prior general modality knowledge transfer accelerates downstream task adaptation and improves computation efficiency. On this basis, when MAE training is added, the model’s average performance improves by 1.3%

Table 4. Performance (%) comparison of on DELIVER dataset.

Method	Backbone	Modality	mIoU
CMXTTS23 [56]	MiT-B2	RGB-Event	56.5
CMNeXtCVPR23 [57]	MiT-B2	RGB-Event	57.5
SPEFTCVPR25 [3]	Swin-B	RGB-Event	58.4
Ours	Hiera-B+	RGB-Event	59.5
CMXTTS23 [56]	MiT-B2	RGB-Depth	62.7
CMNeXtCVPR23 [57]	MiT-B2	RGB-Depth	63.6
CWSAMJSTARS25 [34]	ViT-B	RGB-Depth	58.6
SAM-LoRA _{arXiv24} [63]	ViT-B	RGB-Depth	60.3
MLE-SAM _{arXiv24} [63]	Hiera-B+	RGB-Depth	63.6
MemorySAM _{arXiv25} [26]	Hiera-B+	RGB-Depth	63.5
Ours	Hiera-B+	RGB-Depth	64.9
CMNeXtCVPR23 [57]	MiT-B2	RGB-D-E	64.4
CWSAMJSTARS25 [34]	ViT-B	RGB-D-E	56.2
SAM-LoRA _{arXiv24} [63]	ViT-B	RGB-D-E	60.1
MLE-SAM _{arXiv24} [63]	Hiera-B+	RGB-D-E	62.7
MemorySAM _{arXiv25} [26]	Hiera-B+	RGB-D-E	62.4
Ours	Hiera-B+	RGB-D-E	65.1

Table 5. Ablation studies of each components. W-G denotes the proposed MAE warm-up followed by MAE-guided consistency learning. All scores are given in mIoU (%).

PSST			TP-MoME	MCubeS		DELIVER		Avg
MAE	DINO	W-G	MoME	R-A	R-A-D	R-D	R-D-E	
			✓	49.6	50.1	60.3	59.6	54.9
✓			✓	50.8	51.0	61.4	61.5	56.2
	✓		✓	51.4	51.9	62.4	62.5	57.1
✓	✓		✓	52.3	53.2	63.5	63.6	58.2
✓	✓	✓		52.4	52.7	63.6	63.6	58.1
✓	✓	✓	✓	52.8	54.4	64.9	65.1	59.3

mIoU, whereas adding DINOv2 step results in an average improvement of 2.2% mIoU. When both steps are included but trained independently (first MAE for cross-modal structural reconstruction followed by DINOv2 for semantic feature learning), the performance improves by 3.3% mIoU, which demonstrates the complementary contributions of both. Furthermore, when adding the proposed MAE warm-up followed by MAE-guided consistency learning strategy, the average performance further improves by 1.1% mIoU compared to independent training, which demonstrates the effectiveness of our progressive and interactive design.

Analysis of MoME module. We ablate the proposed MoME multimodal fusion by replacing it with a simple gated summation of modality features. As shown in Table 5, the average performance drops by 1.2% mIoU. Moreover, in scenarios with more modalities, such as R-A-D and R-D-E fusion, the model’s performance drops significantly by 1.7% and 1.5% mIoU respectively, indicating that the benefits of MoME in balancing different knowledge become in-

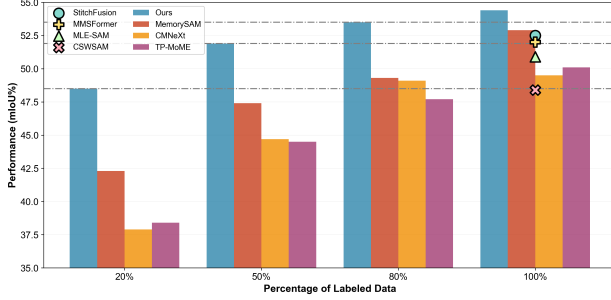


Figure 3. Label Efficiency Analysis on MCubeS. The four points at the 100% mark indicate the performance of the four methods when trained with the full set of labeled data.

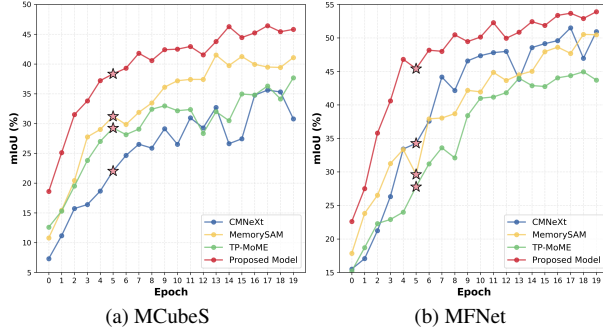


Figure 4. Computation efficiency analysis on MCubeS (RGB-AoLP-DoLP) and MFNet (RGB-Thermal) datasets.

creasingly pronounced as the number of modalities grows.

4.4. Efficiency Analysis

Data Efficiency. As shown in Figure 3, with only 20% labeled data, our method surpasses the performance of CWSAM [34] trained with 100% labeled data, also surpasses MemorySAM [26], CMNeXt [57], and TP-MoME trained with 50% labeled data, and achieves performance comparable to these models trained with 80% labeled data. Moreover, our model surpasses MLE-SAM [63] with just 50% labeled data, and even outperforms the SOTA FFT methods StitchFusion [22] and MMSFormer[39] using 80% labeled data. These results suggest that our Stage 1 training leads to competitive results with less labeled data, alleviating the reliance on large amounts of labeled data.

Computation Efficiency. As shown in Figure 4, on MCubeS, our method at epoch 5 surpasses MemorySAM [26] at epoch 12 and TP-MoME at epoch 19, and outperforms the fully fine-tuned CMNeXt [57] within the first 20 epochs. On MFNet, it also exceeds MemorySAM at epoch 14 and consistently outperforms TP-MoME during the first 20 epochs, demonstrating that our method can achieve superior performance with significantly fewer training epochs.

Storage Efficiency. Our method demonstrates remarkable storage and parameter efficiency when facing multi-task

Table 6. Storage efficiency analysis. #Inc. and #Accu. denote the number of parameter increment when a new task is added and the number of accumulated parameters respectively. * denotes that the parameters of which are frozen.

Method	Modal.	Task	#Inc. (M)	#Accu. (M)
CMNeXt	R-T	MFNet	116.6	116.6
	R-T	PST900	116.6	233.2
	R-D	SUN	116.6	349.8
	R-D	NYUv2	116.6	466.4
	R-D	DELIVER	58.7	525.1
MemorySAM	R	*VFM	80.8	80.8
	R-T	MFNet	12.3	93.2
	R-T	PST900	11.7	104.8
	R-D	SUN	17.3	122.1
	R-D	NYUv2	16.6	138.7
	R-D	DELIVER	15.0	153.7
Ours	R	*VFM	80.8	80.8
	T	LoRA	1.3	82.1
	D	LoRA	1.3	83.4
	R-T	MFNet	7.6	91.0
	R-T	PST900	7.6	98.6
	R-D	SUN	7.6	106.2
	R-D	NYUv2	7.6	113.8
	R-D	DELIVER	7.6	121.4

scenarios. As shown in Table 6, with five downstream tasks, our accumulated parameters are 121.4M with 40.6M trainable, which is significantly lower than 153.7M with 72.9M trainable for MemorySAM [26] and 525.1M trainable for CMNeXt [57], resulting in 44.3% efficiency improvement on the number of trainable parameters. More importantly, based on the above analysis, the modality-specific LoRA, with only 1.3M stored parameters, improves both the label and training efficiency for downstream tasks of that modality, which also reflects the storage efficiency.

5. Conclusions

In this paper, we address the efficiency challenge of cross-modal transfer by proposing a novel transfer paradigm that decouples the transfer process into a general modality knowledge transfer and a flexible task knowledge transfer. To learn task-agnostic target-modality knowledge, we introduce a Progressive Self-Supervised Tuning strategy in Stage 1, which obtains reusable target-modality LoRA using only unlabeled data through a one-time training process. In Stage 2, leveraging the learned LoRAs, we further propose the Task-Prompted Mixture-of-Modality Experts module that achieves fast and lightweight task knowledge learning as well as flexible multimodal fusion for diverse downstream tasks. Extensive experiments demonstrate that our method achieves superior cross-modal transfer performance, while significantly improving data efficiency, computation efficiency, and storage efficiency.

Acknowledgements

This work was supported by National Key Research and Development Program of China under Grant 2024YFE0115500.

References

- [1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 23922–23932, 2024. 3
- [2] Ahmed Agiza, Marina Neseem, and Sherief Reda. Mtlora: Low-rank adaptation approach for efficient multi-task learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 16196–16205, 2024. 3
- [3] Jiaxin Cai, Jingze Su, Qi Li, Wenjie Yang, Shu Wang, Tiesong Zhao, Shengfeng He, and Wenxi Liu. Keep the balance: A parameter-efficient symmetrical framework for rgb+x semantic segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 10587–10598, 2025. 2, 3, 6, 7
- [4] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3367–3375, 2023. 3
- [5] Sangyun Chung, Youngjoon Yu, Youngchae Chee, Se Yeon Kim, Byung-Kwan Lee, and Yong Man Ro. Are vision-language models truly understanding multi-vision sensor? *arXiv preprint arXiv:2412.20750*, 2024. 3
- [6] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 17234–17243, 2024. 1
- [7] Rizhi Ding, Hui Lu, and Manhua Liu. Denseformer-moe: A dense transformer foundation model with mixture of experts for multi-task brain image analysis. *IEEE Trans. Med. Imag.*, 2025. 3
- [8] Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, and Heng Fan. Efficient multimodal semantic segmentation via dual-prompt learning. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 14196–14203, 2024. 6
- [9] Yaqun Fang, Yi Shi, Jia Bei, and Tongwei Ren. Semantic-guided rgb-thermal crowd counting with segment anything model. In *ACM Int. Conf. Multimedia*, pages 570–578, 2024. 3
- [10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024. 3
- [11] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 5108–5115, 2017. 6
- [12] Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 28608–28618, 2024. 3
- [13] Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Proc. Adv. Neural Inf. Process. Syst.*, 37: 67850–67900, 2024. 3
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 16000–16009, 2022. 2
- [15] Qibin He. Prompting multi-modal image segmentation with semantic grouping. In *Proc. AAAI Conf. Artif. Intell.*, pages 2094–2102, 2024. 3, 6
- [16] Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19608–19617, 2025. 3
- [17] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. In *Proc. Int. Conf. Mach. Learn.*, pages 21753–21767, 2024. 6
- [18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proc. Eur. Conf. Comput. Vis.*, pages 709–727, 2022. 3
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 1, 2, 3
- [20] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multimodal data-efficient 3d scene understanding for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025. 1
- [21] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Praateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 20990–21000, 2025. 1, 2, 3
- [22] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. *arXiv preprint arXiv:2408.01343*, 2024. 6, 7, 8
- [23] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhao Luo, Lin Ma, and Min Zhang. Unimoe: Scaling unified multimodal llms with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025. 3
- [24] Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 26170–26180, 2025. 3
- [25] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proc.*

- IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19800–19808, 2022. 6, 7
- [26] Chenfei Liao, Xu Zheng, Yuanhuiyi Lyu, Haiwei Xue, Yihong Cao, Jiawen Wang, Kailun Yang, and Xuming Hu. Memorysam: Memorize modalities and semantics with segment anything model 2 for multi-modal semantic segmentation. *arXiv preprint arXiv:2503.06700*, 2025. 1, 2, 3, 6, 7, 8
- [27] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 8115–8124, 2023. 6
- [28] Peng Liu, Jinhong Deng, Lixin Duan, Wen Li, and Fengmao Lv. Segmenting anything in the dark via depth perception. *IEEE Trans. Multimedia*, 2025. 3
- [29] Xuewen Luo, Fengze Yang, Fan Ding, Xiangbo Gao, Shuo Xing, Yang Zhou, Zhengzhong Tu, and Chenxi Liu. V2x-unipool: Unifying multimodal perception and knowledge reasoning for autonomous driving. *arXiv preprint arXiv:2506.02580*, 2025. 1
- [30] Xianping Ma, Qianqian Wu, Xingyu Zhao, Xiaokang Zhang, Man-On Pun, and Bo Huang. Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Trans. Geosci. Remote Sens.*, 2024. 3
- [31] Mehdi Moshtaghi, Siavash H Khajavi, and Joni Pajarinen. Rgb-th-bench: A dense benchmark for visual-thermal understanding of vision language models. *arXiv preprint arXiv:2503.19654*, 2025. 3
- [32] Xin Niu, Enyi Li, Jinchao Liu, Yan Wang, Margarita Osadchy, and Yongchun Fang. Mind the gap: Learning modality-agnostic representations with a cross-modality unet. *IEEE Trans. Image Process.*, 33:655–670, 2024. 2
- [33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 5
- [34] Xinyang Pu, Hecheng Jia, Linghao Zheng, Feng Wang, and Feng Xu. Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2025. 3, 6, 7, 8
- [35] Abdul Qayyum, Moona Mazher, Devran Ugurlu, Jose Alonso Solis Lemus, Cristobal Roderio, and Steven A Niederer. Foundation model for whole-heart segmentation: Leveraging student-teacher learning in multi-modal medical imaging. *arXiv preprint arXiv:2503.19005*, 2025. 3
- [36] 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 *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763, 2021. 2
- [37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2, 3
- [38] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 3
- [39] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Mmsformer: Multimodal transformer for material and semantic segmentation. *IEEE Open J. Signal Process.*, 5: 599–610, 2024. 6, 7, 8
- [40] 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. *Proc. Adv. Neural Inf. Process. Syst.*, 34: 8583–8595, 2021. 3
- [41] Oindrila Saha, Grant Van Horn, and Subhansu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 17542–17552, 2024. 3
- [42] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *IEEE Int. Conf. on Robotics and Automation*, pages 9441–9447, 2020. 6
- [43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *Proc. Eur. Conf. Comput. Vis.*, pages 746–760, 2012. 6
- [44] Daniel Sliwowski, Shail Jadav, Sergej Stanovcic, Jędrzej Orbik, Johannes Heidersberger, and Dongheui Lee. Reassemble: A multimodal dataset for contact-rich robotic assembly and disassembly. *arXiv preprint arXiv:2502.05086*, 2025. 1
- [45] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 567–576, 2015. 6
- [46] Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Trans. Knowl. Data Eng.*, 2025. 2
- [47] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Med. Image Anal.*, 102:103547, 2025. 3
- [48] Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Naoto Yokoya, and Shijian Lu. Segment anything with multiple modalities. *arXiv preprint arXiv:2408.09085*, 2024. 1
- [49] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: side adapter network for open-vocabulary semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15546–15561, 2023. 3
- [50] Yangyang Xu, Xi Ye, and Duo Su. Multi-task dense prediction fine-tuning with mixture of fine-grained experts. *arXiv preprint arXiv:2507.19077*, 2025. 3
- [51] Yuqi Yang, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, and Bo Li. Multi-task dense prediction via mixture

- of low-rank experts. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 27927–27937, 2024. [3](#)
- [52] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. DFormer: Rethinking RGBD representation learning for semantic segmentation. In *Proc. Int. Conf. Learn. Represent.*, 2024. [6](#)
- [53] Dongshuo Yin, Yiran Yang, Zhechao Wang, Hongfeng Yu, Kaiwen Wei, and Xian Sun. 1% vs 100%: Parameter-efficient low rank adapter for dense predictions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 20116–20126, 2023. [3](#)
- [54] Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. Exploring deeper! segment anything model with depth perception for camouflaged object detection. In *ACM Int. Conf. Multimedia*, pages 4322–4330, 2024. [1](#), [2](#), [3](#)
- [55] Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, and Dongfang Liu. Visual fourier prompt tuning. *Proc. Adv. Neural Inf. Process. Syst.*, 37:5552–5585, 2024. [3](#)
- [56] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Trans. Intell. Transp. Syst.*, 24(12):14679–14694, 2023. [6](#), [7](#)
- [57] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 1136–1147, 2023. [6](#), [7](#), [8](#)
- [58] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. [3](#)
- [59] Jiayi Zhao, Fei Teng, Kai Luo, Guoqiang Zhao, Zhiyong Li, Xu Zheng, and Kailun Yang. Unveiling the potential of segment anything model 2 for rgb-thermal semantic segmentation with language guidance. *arXiv preprint arXiv:2503.02581*, 2025. [1](#), [2](#), [3](#)
- [60] Xiaowei Zhao, Xianglong Liu, Duorui Wang, Yajun Gao, and Zhide Liu. Scene-adaptive and region-aware multi-modal prompt for open vocabulary object detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 16741–16750, 2024. [3](#)
- [61] Zhida Zhao, Jia Li, Lijun Wang, Yifan Wang, and Huchuan Lu. Maskmentor: Unlocking the potential of masked self-teaching for missing modality rgb-d semantic segmentation. In *ACM Int. Conf. Multimedia*, pages 1915–1923, 2024. [6](#)
- [62] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 15659–15669, 2023. [3](#)
- [63] Chenyang Zhu, Bin Xiao, Lin Shi, Shoukun Xu, and Xu Zheng. Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts. *arXiv preprint arXiv:2412.04220*, 2024. [3](#), [6](#), [7](#), [8](#)
- [64] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 9516–9526, 2023. [3](#)