IRGPT: Understanding Real-world Infrared Image with Bi-cross-modal Curriculum on Large-scale Benchmark

Zhe Cao
Beijing Institute of Technology
Beijing, China
zhe.cao@bit.edu.cn

Jin Zhang
Beijing Institute of Technology
Beijing, China
jinzhang@bit.edu.cn

Ruiheng Zhang*
Beijing Institute of Technology
Beijing, China
ruiheng.zhang@bit.edu.cn

Abstract

Real-world infrared imagery presents unique challenges for vision-language models due to the scarcity of aligned text data and domain-specific characteristics. Although existing methods have advanced the field, their reliance on synthetic infrared images generated through style transfer from visible images, which limits their ability to capture the unique characteristics of the infrared modality. To address this, we propose IRGPT, the first multi-modal large language model for real-world infrared images, built upon a large-scale InfraRed-Text Dataset (IR-TD) comprising over 260K authentic image-text pairs. The proposed IR-TD dataset contains real infrared images paired with meticulously hand-crafted texts, where the initial drafts originated from two complementary processes: (1) LLM-generated descriptions of visible images, and (2) rule-based descriptions of annotations. Furthermore, we introduce a bi-cross-modal curriculum transfer learning strategy that systematically transfers knowledge from visible to infrared domains by considering the difficulty scores of both infrared-visible and infrared-text. Evaluated on a benchmark of 9 tasks (e.g., recognition, grounding), IRGPT achieves state-of-the-art performance even compared with larger-scale models.

1. Introduction

With the advancement of large model technologies, Vision-Language Models (VLMs) designed for visible images have achieved significant breakthroughs [19, 48]. This success can be attributed to two main factors: the availability of large-scale aligned image-text datasets [13, 31] and the powerful visual perception capabilities provided by pre-trained vision encoders [8, 16] for Large Language Models (LLMs). VLMs such as LLaVA [24] and Qwen [45] have demonstrated exceptional vision capabilities, enabling them
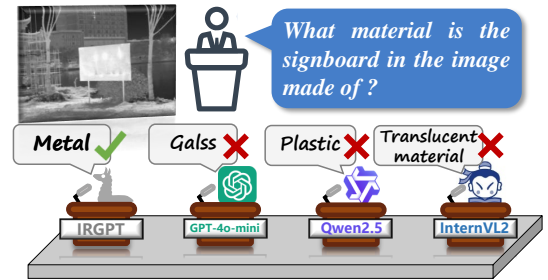

Figure 1. Exist MLLMs exhibit hallucinations when interpreting real-world infrared images, while IRGPT effectively suppresses this phenomenon.

to deeply understand real-world natural images through the medium of corresponding text. However, infrared images, a distinctive and widely utilized image type, have been overlooked by multi-modal large language models (MLLMs), leading to frequent hallucinations when these models process infrared imagery. As illustrated in Fig. 1, the salient metallic signboard in the image is erroneously identified as a non-metallic material object.

In the field of infrared images, the scarcity of large-scale image-text datasets poses challenges to the development of infrared MLLMs. Prior studies [11, 14] have explored methods to obtain infrared-text paired data. However, these methods using visible images to synthesize infrared counterparts create a modality gap, as the synthesised data quality directly depends on generative models. Moreover, such data strictly adheres to the training distribution of the generative model, neglecting the authenticity of infrared images.

To establish a realistic and effective infrared-text dataset (IR-TD) for advancing infrared MLLMs, we also leverage visible images as a medium to employ a new-path approach to obtain infrared-text pairs. On the one hand, we manually choose visible-infrared image pairs and utilize LLMs to generate draft descriptions for visible images, which are then carefully adapted to annotate corresponding infrared
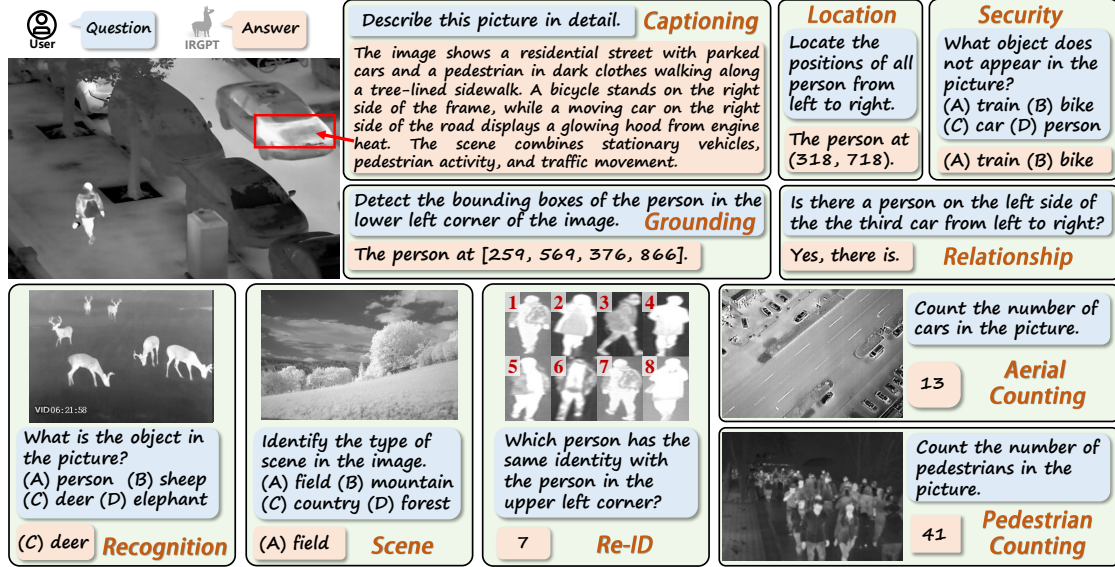
Figure 2. IRGPT effectively handles complex infrared image-text tasks (e.g., grounding, recognition), demonstrating successful knowledge transfer from visible to infrared domains and a deep semantic understanding of infrared data. We further showcase IRGPT's captioning capability in understanding that moving vehicles exhibit elevated engine temperatures.

images and create Q&A pairs. On the other hand, we develop rule-based methods for caption generation and Q&A pair construction, specifically designed by various annotations. The resulting IR-TD, containing 260K+ high-quality infrared-text pairs, adheres to two fundamental principles: 1) authenticity of infrared imagery, and 2) semantic alignment between modalities.

While addressing the data scarcity issue, IR-TD introduces new challenges stemming from inherent characteristics of real-world infrared images, including sparse semantic information, imaging disparities across different spectral bands, and thermal information's dominance in target saliency. These factors render it a significant challenge for an infrared-specific MLLM developed from scratch to rival those MLLMs extensively trained on visible images.

To address this challenge, we introduce a bi-cross-modal curriculum that builds upon the robust comprehension capabilities of existing MLLMs. By leveraging the similarity between infrared and visible images, as well as the semantic alignment between infrared images and text, our approach transfers MLLM understanding from visible to real-world infrared image through incremental pre-training from simple to complex infrared data. With this approach, we propose **IRGPT**, the first infrared-specific MLLM grounded in real-world infrared image. As illustrated in Fig. 2, the developed model showcases strong cross-modal understanding capabilities, particularly in infrared question-answering tasks. To validate IRGPT's performance on common infrared tasks, we construct a benchmark with 9 different infrared Q&A tasks containing 37K+ test samples on the IR-

TD dataset. Through evaluations across various tasks, our model achieves the state-of-the-art performance, demonstrating significant potential for practical applications. The main contributions of this paper are as follows:

- We introduce a large-scale open-source infrared-text dataset, comprising pre-trained subset, fine-tuning instruction subset, and a multi-task benchmark for infrared-text data, which includes 9 types of tasks such as recognition and localization. The total number of samples exceeds 260K.
- To the best of our knowledge, we propose the first vision-language model based on real-world infrared images named IRGPT, for which both the incrementally pre-trained and fine-tuned infrared images are collected from the real world rather than being synthetically generated.
- We also propose a bi-cross-modal curriculum transfer learning strategy that balances the transferability between infrared and visible images and the semantic consistency between infrared images and text.
- In the 9 benchmark tasks, our IRGPT achieves state-of-the-art (SOTA) results. Under zero-shot settings, IRGPT demonstrates a 76.35 improvement in positive sum and a 31.56 reduction in negative sum compared to baseline.

2. Related Works

Multi-modal Large Language Model. In the real world, information exists in diverse forms, making it impossible for a simple LLM to perceive and understand the world fully. Significant efforts have been made to explore how

LLMs can adapt to multi-modal data [17, 53], with notable advancements such as the LLaVA framework [24], which has enabled LLMs to "see" and interpret visual information. From the perspective of current advancements in state-of-the-art MLLMs [1, 6, 22, 37, 42], supporting visual modality data has become an inevitable trend. However, most of these models overlook the infrared images with unique characteristics [4]. Treating infrared images as equivalent to visible images can not avoid leading to severe hallucinations, posing significant challenges for downstream applications and deep understanding of the real scenario.

Infrared Image in Large Models. With the development of multi-modal foundation models, infrared image processing has seen two primary research directions. (a) Unimodal approaches focus on intrinsic infrared characteristics: InfMAE [23] employs masked image modeling for foundational model training, while IRSAM [49] adapts SAM [16] for small target detection through edge enhancement. (b) Multi-modal methods leverage cross-modal synergies: TC-MoA [54] uses MoE frameworks for modality fusion, and works like LanguageBind [52], PandaGPT [34], and ImageBind [11] connect thermal infrared with text/visible modalities. The most relevant work is Infrared-LLaVA [14], which trains an infrared-text large model using the data generated by sRGB-TIR [18]. However, its reliance on video-generated images, which lack authenticity, negatively introduces modality-related hallucinations [25].

Multi-modal Curriculum Learning. Curriculum learning (CL), first proposed in [2], mimics human developmental learning by progressively training models from simple to complex samples through data sequencing strategies [32, 40]. This approach has proven effective for enhancing both training efficiency and model generalization in conventional deep learning [15, 20, 41]. Our analysis highlights two critical CL directions requiring deeper exploration: 1) its application in large-scale models, and 2) curriculum design for multi-modal learning [27, 38, 43, 44]. Recent work shows initial progress in addressing multi-modal training cases [3, 51], while existing methods still lack explorations for training multi-modal large-scale models effectively.

3. IR-TD: Infrared-Text Dataset

3.1. Tri-modal Data Translation

In the field of infrared imaging, large-scale text annotation shows a distinct scarcity [23, 26]. On one hand, textual descriptions for infrared images cannot be effectively obtained through web scraping like the visible image. On the other hand, the difficulty of describing infrared images seriously affects the quality of annotation. As a result, conventional methods for image-text pair collection cannot be directly applied to the infrared domain. With recent advancements in generative models, some works (e.g., IID [14]) have at-
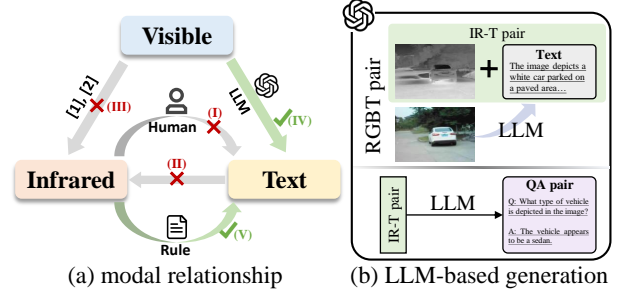

Figure 3. The translation of tri-modal data. (a) Tri-modal data transformation relationships and potential generative paths. (b) The LLM-based infrared-text pair generation method in IR-TD.

tempted to synthesize infrared-text pairs through style transfer approaches. Such methods adhere to fixed generative distributions, which severely degrade the quality of data reliant on synthetic models. To explore pathways for generating infrared-text pairs, we analyze the feasibility of transformations between infrared images, visible images, and textual descriptions, as illustrated in Fig. 3(a).

(I) Manual Annotation: This approach faces significant challenges, particularly the difficulty in achieving diversity, making it impractical for large-scale data pair collection.

(II) Text-to-Image Generation: Most models do not support the generation of infrared data.

(III) Synthetic Image Generation: Converting visible images into infrared ones fails to preserve the authenticity.

(IV) Generation with Aligned RGBT Pairs: Leveraging LLMs to describe visible images that correspond to infrared images and using these descriptions as proxies for infrared image annotations.

(V) Rule-Based Generation: According to the labels, it is possible to generate large-scale image-text pairs, but the drawback is the lack of diversity.

Our proposed solution integrates approaches (IV) and (V) to balance quality and diversity. Specifically, for infrared-visible paired images, we utilize LLMs to generate detailed descriptions of visible images, which are then adopted as textual annotations for the corresponding infrared images. For special scenarios where visible image is insufficient (e.g., nighttime conditions), we employ Retinexformer [5] to improve observability, which can be found in Appendix B. In other cases where visible images present challenges, such as small targets or camouflaged objects, we employ a rule-based approach to generate image-text pairs with available annotations, ensuring that the constructed IR-TD has correct and clear semantics. To effectively train infrared MLLMs, we partition IR-TD into three distinct segments: the pre-training subset for incremental pre-training, the instruction subset for fine-tuning, and the benchmark for validation. Tab. 1 provides some fundamental characteristics of our proposed IR-TD, in comparison

Dataset	IID[14]	IR-TD[1]
Pre-training image samples	118k	190k
Instruction samples	12k	33k
Benchmark samples	22k	37k
Original datasets	7	63
Benchmark tasks	6	9
Open-source	✗	✓
Image authenticity	✗	✓
Data Redundancy	✓	✗

Table 1. The fundamental characteristics of the IR-TD, along with a comparative analysis against the closed-source IID in [14].
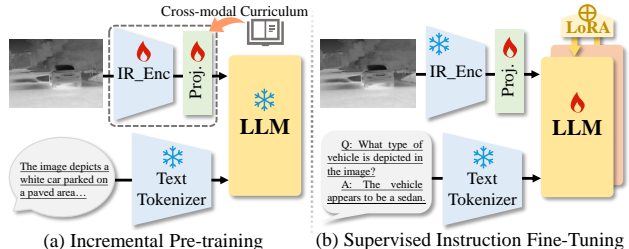

Figure 4. Training strategy of IRGPT. (a) Incremental pre-training stage trains the vision encoder and projector layer via curriculum learning for image-text alignment. (b) Supervised instruction fine-tuning adapts LLM with LoRA and also trains the projector.

with the closed-source dataset IID in [14].

3.2. Pre-training Subset Generation

We aggregated 63 publicly available datasets (details in Appendix H) to collect raw RGBT pairs. To address field-of-view discrepancies caused by inconsistent camera parameters (e.g., datasets like [35, 36]), we aligned semantic content by cropping images based on object locations derived from labels, with examples provided in Appendix J. This ensured textual consistency between infrared-visible pairs. We also rigorously resampled datasets to mitigate redundancy from video-extracted infrared frames (e.g., retaining only 1% of VTUAV's 1.7M images). The final curated dataset contains 84,284 RGBT pairs. As shown in Fig. 3(b), visible images were processed by LLMs to generate descriptive texts, while rule-based methods created 106k infrared-text pairs using annotations.

3.3. Instruction Subset Generation

The fine-tuning instruction subset, similar to the pre-training subset, is bifurcated into LLM-based and rule-based components. The LLM-based portion employs the same prompts as LLaVA [24] and generates Q&A pairs based on the descriptive image-text pairs obtained in Sec. 3.2. This approach facilitates the effective transference of the LLM's question-answering capabilities from visible images to infrared images. Conversely, the rule-based segment is predominantly task-oriented, crafting Q&A pairs that align with the task's framework to bolster the LLM's relevance in addressing task-specific queries when confronted with infrared images.

3.4. Benchmark Defination

The benchmark primarily encompasses 9 cross-modal tasks, designed to comprehensively evaluate the various capabilities of the model. Details on the establishment of the

benchmark can be found in the Appendix A. Specifically, we provide the following definitions:

- **Recognition**: Construct four options, one of which represents the target category contained in the image, and require the model to identify the correct target option.
- **Grounding**: Provide a textual description specifying a target in the image and let model return its bounding box.
- **Location**: Use text to specify a target category, and require the model to provide the coordinate locations of all targets of that category in the image.
- **Relationship**: Determine whether the spatial relationship described in the text is correct.
- **Re-ID**: Construct an image using eight samples arranged in a 2x4 grid, numbered 1-8. Only one sample matches the identity of the top-left sample, testing the model's re-identification capability.
- **Security**: Construct four options, and require the model to identify all targets not present in the image.
- **Aerial Counting**: Vehicle counting via UAV-captured overhead imagery in roadway and parking lot scenarios.
- **Pedestrian Counting**: The image contains a large, crowded group of pedestrians, and the task requires estimating the number of pedestrians.
- **Scene**: Construct four options, one of which correctly describes the scene in the image, and require the model to select the correct option.

4. IRGPT: Infrared Vision-Language Model

4.1. Model Architecture and Training Strategy

Even with the foundation of IR-TD, it is challenging to train a large-scale model dedicated to infrared images from scratch. This difficulty arises primarily due to two factors: absence of a vision-language pre-trained infrared encoder and the semantic sparsity inherent in infrared imagery (e.g. low contrast leading to less informative images). Consequently, our IRGPT will leverage a pre-trained MLLM as a foundation for transfer learning, enhancing its image comprehension and textual expression capability. Our model

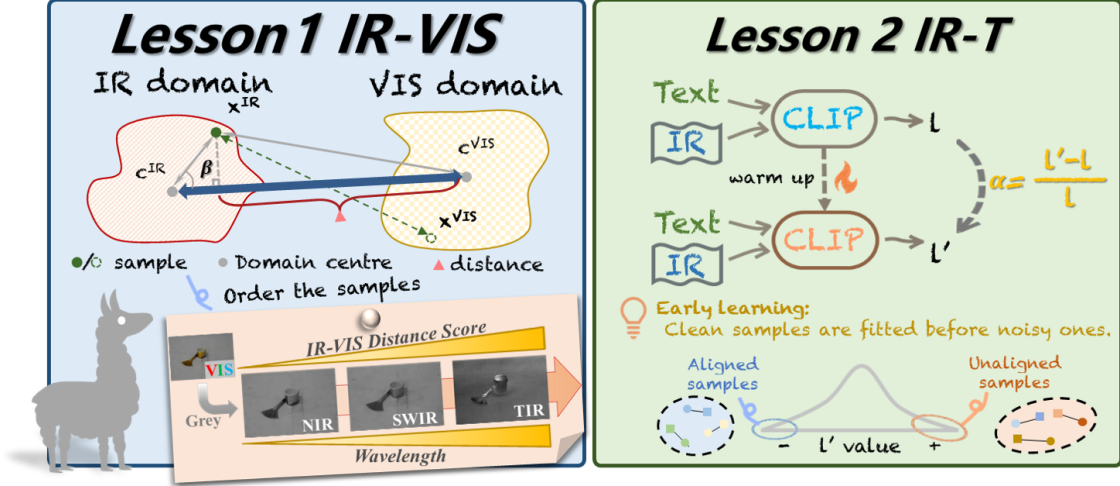---

[1]The dataset is available here.

Figure 5. Bi-cross-modal curriculum transfer strategy. **Lesson 1** focuses on quantifying the domain gap between infrared and visible images through geometric metrics, constructing a domain projection distance as a sample transfer difficulty score. **Lesson 2** emphasizes semantic alignment between infrared images and text pairs, utilizing a pre-warmed CLIP to compute loss, which serves as a sample alignment difficulty score. Notably, the loss variation rate is treated as an auxiliary weight in pre-train. NIR: Near infrared, SWIR: Short wavelength infrared, TIR: Thermal infrared.

consists primarily of a vision encoder, a base LLM, a text tokenizer, and a trainable vision projector. As illustrated in Fig. 4, our training strategy is divided into incremental pre-training and supervised instruction fine-tuning.

Incremental Pre-training: This phase focuses solely on training the vision encoder and its corresponding projection layer. The objective is to adapt the vision encoder to effectively process infrared image inputs while ensuring alignment with textual data.

Supervised Instruction Fine-tuning: During this phase, both the visual projector and the LLM undergo fine-tuning. Notably, the LLM is fine-tuned with the Low-Rank Adaptation strategy to prevent the loss of expressive power.

The incremental pre-training process faces challenges due to the inherent complexity of training data, particularly the varying learning difficulties of infrared-text pairs, which significantly constrain model quality. To address this, we propose Bi-cross-modal Curriculum Transfer Learning, a novel framework that facilitates progressive knowledge transfer from visible to infrared image understanding.

4.2. Bi-cross-modal Curriculum Transfer Learning

4.2.1. Lesson 1: IR-VIS

The modality gap between infrared and visible images represents a critical component of sample difficulty. From an intuitive perspective, we propose to rank infrared images based on geometric metrics since images more similar to visible counterparts would be easier to learn for the vision encoder already adapted to visible images [29, 55]. Notably, our observations reveal that infrared images become progressively dissimilar from visible images with increasing detection wavelengths. Near-infrared images resemble grayscale versions of visible images, while thermal infrared images exhibit fundamentally different characteristics compared to visible images.

Our methodology employs a feature extractor in [23] retrained with masking strategies on both infrared and visible grayscale images to extract discriminative features. To achieve sample-level distance scoring, individual distance computation is required. However, the absence of paired visible counterparts for many infrared images necessitates an innovative solution. As shown in Fig. 5, we therefore propose calculating cross-modal distances through intra-domain distance projection, where the distance from each sample to the opposite modality is estimated by projecting intra-domain distances onto the inter-domain direction.

We first compute the domain discrepancy between infrared and visible domains using the widely adopted Maximum Mean Discrepancy (MMD) [9, 10, 21, 39]:

$$\text{MMD}^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \phi\left(x_i^{ir}\right) - \frac{1}{m} \sum_{j=1}^{m} \phi\left(x_j^{vis}\right) \right\|_{\mathcal{H}}^2, \quad (1)$$

where $\|\cdot\|_{\mathcal{H}}^2$ represents the squared norm in the Reproducing Kernel Hilbert Space (RKHS). $\phi : \mathcal{X} \to \mathcal{H}$ denotes the feature map induced by Gaussian kernel. $x_i^{ir} \in \mathcal{X}^{ir}$ and $x_j^{vis} \in \mathcal{X}^{vis}$ denote the i-th infrared domain sample and j-th visible domain sample, respectively.

Subsequently, we calculate domain center vectors:

$$c^{ir} = \frac{1}{n} \sum_{i=1}^{n} \phi\left(x_i^{ir}\right), \quad c^{vis} = \frac{1}{m} \sum_{j=1}^{m} \phi\left(x_j^{vis}\right). \quad (2)$$

Finally, the sample distance score is obtained by projecting the infrared sample's distance vector to the infrared domain center onto the inter-domain direction and adding the domain discrepancy:

$$d_i = \left(\phi\left(x_i^{ir}\right) - c^{ir}\right) \cdot \frac{c^{vis} - c^{ir}}{\|c^{vis} - c^{ir}\|} + \text{MMD}. \quad (3)$$

This approach enables effective quantification of cross-modal sample difficulty while addressing the challenge of incomplete visible-infrared correspondence. It is worth noting that the distances obtained by this method are not precise. However, since our goal is to acquire a difficulty ranking of the samples, this approximation is acceptable.

4.2.2. Lesson 2: IR-T

Another critical factor constraining the visual comprehension capabilities of MLLMs lies in the alignment quality between visual and textual modalities. Empirical observations [7, 30] suggest that semantically well-aligned samples demonstrate superior learnability, whereas premature exposure to loosely-aligned samples during training may induce oscillatory optimization trajectories that impede effective model convergence. Consequently, the degree of cross-modal alignment serves as an essential indicator for assessing sample complexity. Prior research [12, 46] has investigated sample-level assessment of multi-modal alignment relationships, predominantly grounded in early learning principles. This theoretical framework posits that deep models inherently prefer assimilating clean samples over noisy counterparts during initial training phases.

Inspired by these insights, we implement a pre-warmed CLIP [30] to establish a discriminative criterion for infrared-text alignment through its loss values. However, the observed loss distribution displays pronounced Gaussian characteristics (as shown in Fig. 6) with significant concentration around the mean value. This phenomenon obscures the identification of two crucial sample groups. The first group includes hard samples that require intensive optimization efforts, and the second consists of misaligned samples that warrant weight attenuation to prevent erroneous gradient updates. These issues stem from the static loss measurement from the pre-warmed model, which merely reflects instantaneous model performance and fails to capture dynamic learning behaviors.

To address the limitations posed by static loss, we propose a novel dynamic metric called loss variation rate. This metric is designed to dynamically assess model learning progress and optimize training efficiency by calculating the rate of change in loss as follows:

$$\alpha = \frac{l' - l}{l}, \quad (4)$$

where l and l' denote pre-warm-up and post-warm-up loss values respectively. This adaptive ratio enables dual functionality: 1) suppressing the influence of misaligned samples through weight modulation, while 2) preserving optimization focus on valuable hard samples.

4.2.3. Curriculum Schedule

The distance scores and alignment scores derived from the two lessons establish distinct sample rankings from their respective perspectives. We constructed a comprehensive ranking of all samples through systematic integration of these dual sequence rankings. However, the static nature of sample difficulty in this configuration might compromise generalization performance. Empirical insights [47, 50] suggest that introducing stochasticity to difficulty ordering can enhance model generalization bounds. Following the classical baby-step [33], we partition the training data into M tiers based on difficulty levels. During pre-training, tiers are sequentially ordered by difficulty while implementing random sampling within each tier. This method achieves dual optimization objectives: 1) It preserves structured learning progression while introducing controlled variability in difficulty exposure, enhancing the model's adaptability to diverse challenge levels. 2) The stochastic tiered sampling provides inherent error tolerance for potential scoring discrepancies, creating more resilient learning dynamics compared to deterministic sequencing approaches.

Furthermore, we incorporate the variation rate derived from Eq. (4) as an adaptive weighting factor to simultaneously mitigate misaligned sample impacts and intensify focus on challenging instances. The final formulation integrates these components through:

$$w_i = \begin{cases} 1 - \sigma\left(\frac{\alpha_i}{\text{Median}(\{\alpha_j | \alpha_j > 0\})}\right) & \alpha_i > 0, \\ 1 + \sigma\left(\frac{-\alpha_i}{\text{Median}(\{-\alpha_k | \alpha_k \leq 0\})}\right) & \alpha_i \leq 0, \end{cases} \quad (5)$$

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} w_i \cdot \left[-\sum_{c=1}^{C} y_{i,c} \log p_{i,c} \right], \quad (6)$$

where $\text{Median}(\cdot)$ denotes the median operator and $\sigma(\cdot)$ represents the sigmoid function.

5. Experiment

5.1. Experiment Setting

Implementation Details. We utilized InternVL2-8B as the base model and adopted the LoRA strategy for LLM fine-tuning. During the incremental pre-training phase, the learning rate was set to 1e-3 for 1 epoch, while the supervised instruction fine-tuning phase employed a learning rate of 2e-5 for 5 epochs, with a consistent batch size of 32

Model	Scene	Rec.	Gro.	Rel.	ReID	Sec.	psum	Loc.	A.C.	P.C.	nsum
(a) Zero-shot											
Pandagpt-7B	28.49	45.10	23.22	47.75	23.69	22.10	190.35	39.54	25.19	78.29	143.02
LLaVA1.5-7B	35.89	23.11	11.17	23.68	2.61	18.63	115.09	83.12	42.39	117.22	242.73
Qwen2-VL-7B	49.80	59.42	25.74	_56.19_	4.62	30.08	225.85	50.41	18.05	89.12	157.58
InternVL2-8B	46.02	77.69	30.05	53.02	7.33	38.19	252.30	43.29	21.75	60.39	125.43
InternVL2-26B	62.24	72.88	**39.48**	55.42	8.16	43.72	281.90	38.25	21.68	64.65	124.58
IRGPT (anti-CL)	61.29	80.25	27.89	52.66	19.44	_45.31_	286.84	35.22	20.29	53.93	109.44
IRGPT (Random)	_63.93_	_83.38_	33.91	53.67	_28.18_	44.11	_307.18_	_34.99_	_15.33_	**47.17**	_97.49_
IRGPT (CL)	**65.12**	**86.28**	_36.68_	**58.33**	**33.55**	**48.69**	**328.65**	**33.32**	**13.25**	_47.30_	**93.87**
(b) Fine-tune											
Pandagpt-7B	82.93	99.12	49.04	97.22	47.51	99.32	475.14	12.29	1.31	6.13	19.73
LLaVA1.5-7B	80.04	96.32	44.48	95.11	42.89	97.56	456.40	20.69	1.71	11.66	34.06
Qwen2-VL-7B	83.21	98.47	47.93	95.99	44.83	98.84	469.27	16.44	1.48	7.39	25.31
InternVL2-8B	82.14	97.78	47.59	93.25	44.97	98.76	464.49	18.70	1.52	6.64	26.85
InternVL2-26B	_84.03_	98.94	48.77	95.89	47.09	99.06	473.78	12.50	1.38	5.68	19.57
IRGPT (anti-CL)	79.89	97.65	42.01	95.23	23.09	95.37	433.24	19.02	1.61	5.87	26.50
IRGPT (Random)	83.89	_99.07_	_50.32_	_97.13_	_49.68_	_99.37_	_479.46_	_9.74_	_1.12_	_2.09_	_12.94_
IRGPT (CL)	**85.12**	**99.79**	**51.58**	**98.69**	**50.79**	**99.82**	**485.79**	**3.32**	**0.25**	**0.82**	**4.39**

Table 2. Comparison with other models. Among the 9 task comparisons, there are 6 positive metrics and 3 negative metrics, represented by positive sum (psum) and negative sum (nsum) respectively. The tasks are denoted by abbreviations with their corresponding full names in Sec. 3.4. We compared zero-shot and fine-tuned results. The difference is whether models were fine-tuned on our instruction dataset. The best results are highlighted in **bold**, and the sub-optimal results are underlined.

across both stages. We train our model with an AdamW [28] optimizer and a cosine learning rate scheduler.

Evaluation Metrics. Given the diverse output requirements across the 9 infrared-text tasks, we implemented different evaluation metrics for each task's specific characteristics. For *Location*, *Aerial Counting*, and *Pedestrian Counting* tasks, we calculated prediction errors using MAE. The *Grounding* task involving bounding boxes was evaluated by mAP@0.5, while other tasks were measured using accuracy (Acc). For clarity, we aggregate the positive (psum) and negative (nsum) score metrics separately, where higher psum and lower nsum indicate better performance.

Compared Methods. We selected five MLLMs for comparison: PandaGPT-7B [34] (a pioneer in adapting LLMs to infrared images), LLaVA1.5-7B [24], Qwen2-VL-7B [37], InternVL2-8B [6] (baseline), and InternVL2-26B (larger-scale variant). IRGPT comprises three versions with distinct training strategies: anti-CL (hard to easy), Random (full randomization), and CL (strategy defined in Sec. 4).

5.2. Main Results

Observed from Tab. 2, our IRGPT model pre-trained with bi-cross-modal curriculum transfer learning achieves state-of-the-art zero-shot performance on 7 tasks, outperforming the baseline InternVL2-8B by 76.35 psum, with a notable 26.22 improvement on ReID tasks. The 31.56 reduction in nsum suggests enhanced capability in object localization and counting. After fine-tuning, our method surpasses all models including the larger InternVL2-26B, achieving the

highest psum of 485.79 (12.01 over InternVL2-26B) with a nsum of merely 4.39 (15.18 lower than InternVL2-26B).

Regarding training strategies, our bi-cross-modal curriculum transfer learning demonstrates clear superiority over random training, achieving a notable improvement of 14.47 in psum under zero-shot settings. However, we observe that the nsum metric exhibits a marginal decrease of merely 2.12, failing to demonstrate comparable improvement to psum. Notably, the Random variant achieves superior performance in personnel counting tasks. We attribute this discrepancy to the pretraining phase's primary emphasis on image-text alignment, which facilitates object category and spatial relation identification, rather than precise image comprehension and task-specific responses (e.g., localization and object counting). The anti-CL approach with inverse difficulty ordering proves particularly challenging, leading to severe performance degradation. For instance, in ReID tasks, the fine-tuned model frequently collapses to the same outputs, completely failing to capture pedestrian identity characteristics. More experiments and details are provided in the Appendix E&I.

5.3. Ablation Study

Ablation studies cover curriculum components and sampling strategies. The curriculum ablation results in Tab. 3 demonstrate that removing curriculum and using anti-curriculum learning both degrade infrared image comprehension. Notably, Lesson2 (L2) shows greater sensitivity to psum than Lesson1 (L1), as L2 emphasizes image-text sim-

L1	anti-L1	L2	anti-L2	α	psum	nsum
					464.49	26.85
	✓		✓	✓	433.24	26.50
			✓	✓	454.73	24.36
		✓		✓	470.23	17.29
	✓			✓	456.30	24.59
✓				✓	466.81	19.33
✓		✓			481.36	7.08
✓		✓		✓	485.79	4.39

Table 3. The ablation study on the pre-training effects of two cross-modal lessons. All comparisons here are based on results after instruction fine-tuning.

Schedule	zero-shot		fine-tune	
	psum	nsum	psum	nsum
Difficulty-Ascending	309.63	101.90	474.29	19.24
Difficulty-Descending	286.84	109.44	433.24	26.50
Bidirectional Ordering	308.29	105.08	469.84	19.53
Random	307.18	97.49	479.46	12.94
Descending-Stratified Random	302.85	103.62	470.22	18.92
Ascending-Stratified Random	328.65	93.87	485.79	4.39

Table 4. The ablation study of sampling schedule for batch selection. Difficulty-Descending/Ascending indicate difficulty-ordered selection. Bidirectional Ordering balances batch difficulty via alternating directions. '-Stratified' implies difficulty-based grouping. Ascending-Stratified Random is our method in Sec. 4.

ilarity whereas L1 prioritizes image quality. Image quality difficulty affects perceptual accuracy, while alignment complexity impacts semantic correctness.

As shown in Tab. 4, we compare different curriculum schedules. Our findings reveal that all reverse difficulty-based strategies adversely affect performance, while approaches incorporating randomness consistently enhance model effectiveness. Notably, even the descending-stratified random strategy achieves significantly superior results compared to the difficulty-descending approach.

5.4. Visualization

Sample Difficulty Distribution. We analyze normalized evaluation metric distributions from two experimental cases. As shown in Fig. 6, the sample distance distribution exhibits a bimodal pattern, arising from infrared cameras' predominant use of two spectral bands (thermal imaging and near-infrared) coupled with significant clarity variations across datasets. CLIP's normally distributed losses stem from the Central Limit Theorem compliance, and infrared-text alignment noise follows Gaussian distributions. These statistical properties do not affect difficulty ranking validity, as our metric relies on relative score rankings instead of absolute values.

Example of Difficulty Ranking. Fig. 7 presents three illustrative samples that demonstrate discernible trends in image quality and text-image alignment. The learning dif-
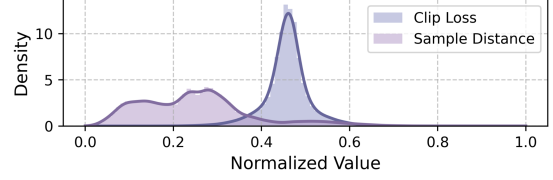

Figure 6. The distribution of difficulty score. The CLIP loss exhibits a normal distribution, while the sample distances show a bimodal distribution.
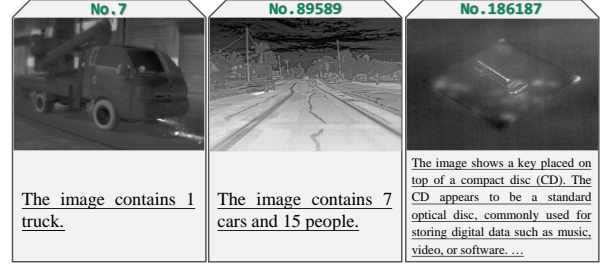

Figure 7. Sample pairs with ranking results. Demonstrating variations in semantic complexity and image quality.

ficulty of the samples exhibits a progressive escalation corresponding to ascending rankings. This empirical evidence substantiates the efficacy of the difficulty metric employed in our bi-cross-modal curriculum transfer learning framework, where sample complexity shows positive correlation with hierarchical ranking positions.

6. Conclusion

We introduce IRGPT, the first vision-language model specifically designed for real-world infrared image understanding. Through the development of the large-scale IR-TD dataset and a bi-cross-modal curriculum transfer learning framework, our approach successfully addresses the modality gap between visible and infrared domains while maintaining authentic semantic representation. The proposed model achieves state-of-the-art performance across multiple benchmarks, demonstrating superior efficiency compared to larger-scale counterparts.

Acknowledge

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 3

[3] Denis Blessing, Onur Celik, Xiaogang Jia, Moritz Reuss, Maximilian Li, Rudolf Lioutikov, and Gerhard Neumann. Information maximizing curriculum: A curriculum-based approach for learning versatile skills. In *Advances in Neural Information Processing Systems*, pages 51536–51561. Curran Associates, Inc., 2023. 3

[4] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*, 2024. 3

[5] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12504–12513, 2023. 3

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 7

[7] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019. 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2021. 5

[10] Pengfei Ge, Chuan-Xian Ren, Xiao-Lin Xu, and Hong Yan. Unsupervised domain adaptation via deep conditional adaptation network. *Pattern Recognition*, 134:109088, 2023. 5

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 3

[12] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 6

[13] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023. 1

[14] Shixin Jiang, Zerui Chen, Jiafeng Liang, Yanyan Zhao, Ming Liu, and Bing Qin. Infrared-llava: Enhancing understanding of infrared images in multi-modal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8573–8591, 2024. 1, 3, 4

[15] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24120–24131, 2023. 3

[16] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 3

[17] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2025. 3

[18] Dong–Guw Lee, Myung–Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8291–8298, 2023. 3

[19] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 1

[20] Conglong Li, Minjia Zhang, and Yuxiong He. The stability-efficiency dilemma: Investigating sequence length warmup for training gpt models. *Advances in Neural Information Processing Systems*, 35:26736–26750, 2022. 3

[21] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 5

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[23] Fangcen Liu, Chenqiang Gao, Yaming Zhang, Junjie Guo, Jinghao Wang, and Deyu Meng. Infmae: A foundation

model in the infrared modality. In *European Conference on Computer Vision*, pages 420–437. Springer, 2025. 3, 5

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1, 3, 4, 7

[25] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 3

[26] Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, et al. Lsotb-tir: A large-scale high-diversity thermal infrared object tracking benchmark. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3847–3856, 2020. 3

[27] Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*, 2024. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[29] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, pages 512–523. Curran Associates, Inc., 2020. 5

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 1

[32] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022. 3

[33] Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, 2010. 6

[34] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 3, 7

[35] Kunpeng Wang, Keke Chen, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network. *arXiv preprint arXiv:2412.14576*, 2024. 4

[36] Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network

and a unified benchmark. *IEEE Transactions on Multimedia*, 2024. 4

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 7

[38] Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723, 2020. 3

[39] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (1):264–277, 2023. 5

[40] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021. 3

[41] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5852–5864, 2023. 3

[42] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 3

[43] Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, et al. Training large language models for reasoning through reverse curriculum reinforcement learning. In *International Conference on Machine Learning*, pages 54030–54048. PMLR, 2024. 3

[44] Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online, 2020. Association for Computational Linguistics. 3

[45] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[46] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023. 6

[47] Shuo Yang, Zhe Cao, Sheng Guo, Ruiheng Zhang, Ping Luo, Shengping Zhang, and Liqiang Nie. Mind the boundary: Coreset selection via reconstructing the decision boundary. In *Forty-first International Conference on Machine Learning*, 2024. 6

[48] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[49] Mingjin Zhang, Yuchun Wang, Jie Guo, Yunsong Li, Xinbo Gao, and Jing Zhang. Irsam: Advancing segment anything model for infrared small target detection. In *European Conference on Computer Vision*, pages 233–249. Springer, 2025. 3

[50] Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. Coverage-centric coreset selection for high pruning rates. In *11th International Conference on Learning Representations, ICLR 2023*, 2023. 6

[51] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. Intra-and inter-modal curriculum for multi-modal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3724–3735, 2023. 3

[52] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[53] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. 3

[54] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024. 3

[55] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. 5