Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework

Jingjing Zhang * and Yangshu Lin

Posted Date: 13 November 2025

doi: 10.20944/preprints202511.0987.v1

Keywords: multimodal learning; fine-grained alignment; visual grounding; iterative rectification; cross-modal reasoning

 doi:10.20944/preprints202511.0987.v1

Article

Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework

Jingjing Zhang * and Yangshu Lin

Zhoukou Normal University
* Correspondence: 20190121@stu.sqxy.edu.cn

Abstract

Current multimodal models show strong general understanding across vision and language but often struggle with detailed visual grounding, complex reasoning, and spatial consistency. To address these challenges, we introduce a Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework (FGAM). The framework follows a two-stage paradigm. In the first stage, fine-grained cross-modal pre-training constructs region–text pairs and applies contrastive and spatial consistency objectives to enhance precise visual-semantic alignment. In the second stage, iterative reasoning and rectification fine-tuning introduces a self-evaluation loop where a rectification module reviews and refines model outputs based on visual evidence. Experiments on multiple multimodal backbones and benchmarks demonstrate that FGAM improves fine-grained reasoning, spatial understanding, and reduces hallucinations. Ablation and human evaluations confirm the effectiveness of each component and the overall reliability of the framework.

Keywords: multimodal learning; fine-grained alignment; visual grounding; iterative rectification; cross-modal reasoning

1. Introduction

Multimodal Large Language Models (MLLMs) have demonstrated remarkable progress in comprehending and generating text related to visual information, effectively bridging the gap between perception and language [1,2]. Their ability to process and fuse information from diverse modalities has opened new avenues for advanced AI applications, ranging from sophisticated image captioning to complex visual question answering (VQA) [3], and showcasing advanced generalization capabilities [4]. However, despite these advancements, existing MLLMs frequently encounter significant hurdles when tasked with processing *fine-grained visual information*, particularly in scenarios demanding precise localization of specific objects within an image and subsequent intricate reasoning based on these minute details.

Current MLLMs often struggle to accurately identify concrete, subtle regions or objects within an image, and subsequently, to perform refined logical inference on these subtle visual cues. This limitation frequently leads to undesirable phenomena such as *spatial hallucinations*, where the model describes an object as present in a specific location when it is not, or *reasoning biases*, where conclusions are drawn based on imprecise visual understanding [5]. Such inaccuracies severely restrict the applicability of MLLMs in high-precision visual understanding domains, including medical image analysis [6], industrial quality inspection, and detailed scene description, where even minor misinterpretations can have critical consequences. The challenge of fine-grained cross-modal alignment and retrieval is also a significant area of research [7–9]. Motivated by these challenges, we propose a novel learning framework designed to significantly enhance MLLMs' fine-grained visual localization and reasoning capabilities by bolstering their granular perception in image-text alignment and integrating a robust iterative correction mechanism.

 doi:10.20944/preprints202511.0987.v1

In this paper, we introduce the **Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework (FGAM)**, a two-stage training paradigm engineered to address the aforementioned shortcomings in MLLMs' fine-grained visual comprehension and reasoning. The first stage, *Fine-Grained Cross-Modal Pre-training*, focuses on building robust "region-text" alignment. This is achieved by constructing region-text paired samples, leveraging existing image annotations (e.g., bounding boxes, segmentation masks) or generating pseudo-labels via pre-trained visual detection models like Grounding DINO [10]. We then design a novel "region-text contrastive learning" loss that, in addition to global image-text alignment, performs contrastive learning between local image region representations and their corresponding textual descriptions, thereby fostering more granular visual-semantic correspondences. Furthermore, a "spatial consistency constraint" is introduced to ensure textual descriptions remain spatially coherent and conflict-free across different regions, mitigating potential spatial hallucinations. The second stage, *Iterative Reasoning and Rectification Fine-tuning*, builds upon this fine-grained pre-training. During this stage, the model not only generates an initial response but also undergoes a "self-evaluation and correction" loop. A dedicated "rectification module" assesses the initial response against the raw visual input for fine-grained localization errors or reasoning inconsistencies. If issues are detected, this module provides "correction feedback" to guide the main model in revising its answer, iteratively refining the output until acceptable fine-grained consistency is achieved. This iterative process is optimized through reinforcement learning or supervised learning on datasets augmented with correction trajectories.

To validate the efficacy and generality of FGAM, we conducted extensive experiments on popular MLLM architectures, specifically LLaVA-1.5-7B [11] and MiniGPT-4-7B [12]. For the Stage-1 fine-grained pre-training, we utilized a large-scale subset of LAION-400M [13] (approximately 20 million samples) complemented by richly annotated datasets like COCO [14] and Visual Genome [15]. For Stage-2 iterative rectification fine-tuning, we extended existing multimodal instruction tuning datasets such as LLaVA-Instruct-150K [11] and ShareGPT4V [16] by incorporating problems requiring fine-grained visual localization and complex reasoning, along with semi-automatically generated error correction trajectories. Our evaluation encompassed a comprehensive suite of benchmarks, including specialized fine-grained localization and reasoning tasks (Grounded VQA, Factual-VQA, A-OKVQA, SpatialSense), hallucination detection benchmarks (MMHal-Bench, POPE), and general multimodal capabilities (MME, MMBench, MM-Vet, SEED-Bench, VQAv2, GQA, VizWizQA, TextVQA, ScienceQA). The results demonstrate that FGAM consistently yields significant improvements over strong baselines, particularly in fine-grained visual understanding and hallucination reduction tasks, confirming the effectiveness of our proposed fine-grained alignment and iterative correction mechanisms. For instance, FGAM improved LLaVA-1.5-7B's performance on SpatialSense by +2.1% and reduced its hallucination rate on MMHal-Bench by 0.04 points.

Our main contributions are summarized as follows:

- We propose FGAM, a novel two-stage learning framework that significantly enhances MLLMs' fine-grained visual localization and complex reasoning capabilities through dedicated fine-grained alignment and iterative rectification.
- We introduce a Fine-Grained Cross-Modal Pre-training stage featuring region-text contrastive learning and spatial consistency constraints, enabling MLLMs to learn more precise visual-semantic correspondences at a granular level.
- We develop an Iterative Reasoning and Rectification Fine-tuning stage, incorporating a self-evaluation and correction mechanism that allows MLLMs to iteratively refine their responses, effectively mitigating spatial hallucinations and reasoning biases.

2. Related Work

2.1. Multimodal Large Language Models (MLLMs)

Research in Multimodal Large Language Models (MLLMs) addresses foundational challenges such as visual in-context learning [2] and weak-to-strong generalization [4]. For multimodal sentiment

 doi:10.20944/preprints202511.0987.v1

analysis, methods have been developed to model unaligned sequences using graph-based networks [17] and improve robustness with Bayesian fusion [18]. Key advancements also include debiasing for multilingual text classification [19], creating efficient fine-tuning frameworks like LlamaFactory [20], and exploring unsupervised pre-training for domain adaptation [21]. To enhance contextual reasoning, researchers have injected commonsense knowledge via graphs [22] and proposed 'thread of thought' approaches to manage chaotic contexts [23]. The field is supported by crucial resources like the SIMMC 2.0 benchmark for task-oriented dialogue [24]. MLLM applications are expanding into specialized domains, including medical image analysis [6], audio content generation [1], controllable 3D urban design [25], personalized architecture generation [26], and aiding creative practices with digital tools [27]. Furthermore, related modeling techniques are applied in complex engineering tasks, such as online parameter estimation for motors under sensorless control [28–30].

2.2. Fine-Grained Visual Grounding and Hallucination Mitigation

Achieving fine-grained visual grounding is a key challenge, with recent efforts in 2D-3D cross-modal retrieval using decoupled discriminative learning [9], hierarchical perspectives [8], and prototypical voting [7]. Visual grounding techniques include hierarchical prefix fusion for multimodal extraction [31] and end-to-end transformers for video grounding [10]. These principles of spatial and semantic understanding are also critical in autonomous systems for enhancing SLAM with dense semantics [32], enabling collision-free driving [33], and developing safe automated planners [34]. Mitigating hallucinations is another critical research area, addressed by developing taxonomies and datasets of hallucination types [35], grounding dialogue responses to knowledge graphs [36], and establishing evaluation benchmarks [37]. Practical mitigation strategies include inference-time algorithms that guide generation away from undesirable content [38]. Relatedly, work on understanding nuanced language in emotional support conversations [39] and generating contrastive explanations for model interpretability [40] also informs progress in fine-grained multimodal understanding.

3. Method

In this section, we elaborate on our proposed **Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework (FGAM)**, a novel two-stage training paradigm designed to significantly enhance Multimodal Large Language Models (MLLMs) in fine-grained visual localization and complex reasoning tasks. FGAM addresses the limitations of existing MLLMs by fostering more precise visual-semantic correspondences at a granular level and integrating a robust iterative correction mechanism.

3.1. Overall Framework

The FGAM framework comprises two distinct yet complementary stages: **Stage-1: Fine-Grained Cross-Modal Pre-training** and **Stage-2: Iterative Reasoning and Rectification Fine-tuning**. In Stage-1, we focus on establishing strong, fine-grained alignments between visual regions and their corresponding textual descriptions. This is achieved through novel data construction and a region-text contrastive learning objective, complemented by a spatial consistency constraint. Subsequently, Stage-2 builds upon these granular representations by introducing an iterative self-evaluation and rectification loop during instruction tuning, enabling the MLLM to refine its responses based on visual evidence and mitigate hallucinations.

3.2. Stage-1: Fine-Grained Cross-Modal Pre-Training

The primary objective of Stage-1 is to imbue MLLMs with an enhanced capability to perceive and understand fine-grained visual details and their semantic correlates. This stage involves constructing specialized region-text paired data and training the MLLM using a combination of contrastive learning and spatial consistency regularization. The MLLM's visual encoder f_V and language encoder f_L are jointly trained to produce semantically rich and spatially aware embeddings.

 doi:10.20944/preprints202511.0987.v1

3.2.1. Region-Text Pair Construction

To facilitate fine-grained cross-modal alignment, we first enrich standard image-text pairs by decomposing images into multiple visual regions and associating each region with a precise textual description. For datasets with existing rich annotations, such as COCO and Visual Genome, we directly extract bounding boxes or segmentation masks and their corresponding captions. These annotations provide high-quality ground-truth region-text pairs. For large-scale unannotated datasets like LAION-400M, we employ a two-step pseudo-labeling process to generate these pairs:

1. **Region Proposal Generation**: We utilize a pre-trained open-vocabulary object detection model, such as Grounding DINO, to generate a diverse set of bounding box proposals for salient objects and regions within each image. Grounding DINO is particularly effective due to its ability to detect arbitrary objects specified by text prompts, allowing for flexible and comprehensive region extraction. Each proposal $\mathbf{b}_i$ is associated with a confidence score indicating the likelihood of it containing a meaningful object or region.
2. **Region Description Generation**: For each detected region $\mathbf{r}_i$ (defined by its bounding box $\mathbf{b}_i$), we extract its visual features. We then generate a concise, descriptive text $t_{\mathbf{r}_i}$ using one of two strategies: (a) employing a smaller, specialized image description model fine-tuned for regional captioning, or (b) leveraging advanced Large Language Models (LLMs) like GPT-4. When using an LLM, the original global image caption T and the visual features of $\mathbf{r}_i$ are provided as context, prompting the LLM to refine T into a region-specific description $t_{\mathbf{r}_i}$. This process aims to create high-quality, fine-grained "region-text" pairs $(\mathbf{r}_i, t_{\mathbf{r}_i})$, where $\mathbf{r}_i$ denotes the i-th visual region and $t_{\mathbf{r}_i}$ is its textual description.

This careful construction of region-text pairs is fundamental for training the MLLM to understand granular visual semantics.

3.2.2. Region-Text Contrastive Learning

Building upon the constructed region-text pairs, we introduce a novel **Region-Text Contrastive Learning (RTCL)** loss. This loss extends the conventional global image-text contrastive learning objective by enforcing alignment at a more granular level. Given an image I with its global caption T, and a set of N extracted regions $\{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ with their corresponding descriptions $\{t_{\mathbf{r}_1}, \ldots, t_{\mathbf{r}_N}\}$, we aim to learn representations such that the embeddings of matched region-text pairs are closer, while unmatched pairs are pushed apart.

Let $f_V(\cdot)$ be the visual encoder that extracts region features $\mathbf{v}_{\mathbf{r}_i} = f_V(\mathbf{r}_i)$, and $f_L(\cdot)$ be the language encoder that extracts text features $\mathbf{e}_{t_{\mathbf{r}_i}} = f_L(t_{\mathbf{r}_i})$. The features $\mathbf{v}_{\mathbf{r}_i}$ and $\mathbf{e}_{t_{\mathbf{r}_i}}$ are projected into a shared embedding space. The region-text contrastive loss for a batch of B region-text pairs is formulated as a symmetric InfoNCE loss:

$$\mathcal{L}_{\text{RTCL}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathbf{v}_{\mathbf{r}_i}^{\top}\mathbf{e}_{t_{\mathbf{r}_i}}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{v}_{\mathbf{r}_i}^{\top}\mathbf{e}_{t_{\mathbf{r}_j}}/\tau)} + \log\frac{\exp(\mathbf{e}_{t_{\mathbf{r}_i}}^{\top}\mathbf{v}_{\mathbf{r}_i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{e}_{t_{\mathbf{r}_i}}^{\top}\mathbf{v}_{\mathbf{r}_j}/\tau)}\right] \tag{1}$$

where τ is a learnable temperature parameter that scales the similarity scores. The term $\mathbf{v}_{\mathbf{r}_i}^{\top}\mathbf{e}_{t_{\mathbf{r}_i}}$ represents the cosine similarity between the visual embedding of region $\mathbf{r}_i$ and the text embedding of its description $t_{\mathbf{r}_i}$. This loss ensures that the visual representation of a region is highly similar to its correct textual description, and dissimilar to other region descriptions in the batch (acting as negative samples), and vice versa. This objective is combined with a global image-text contrastive loss $\mathcal{L}_{\text{ITCL}}$ (e.g., a CLIP-style loss applied to the full image and global caption) to maintain overall multimodal understanding and prevent overfitting to region-level details.

 doi:10.20944/preprints202511.0987.v1

3.2.3. Spatial Consistency Constraint

To prevent potential spatial hallucinations and ensure the coherence of region-level understanding, we introduce a **Spatial Consistency Constraint (SCC)**. This constraint encourages the model's internal representations to reflect the geometric arrangement of objects, thereby avoiding contradictions or illogical spatial relationships between described regions. For instance, if a region is described as "a red car on the left," and an overlapping region is described as "a blue car on the right," the constraint would penalize such inconsistencies by aligning the predicted spatial relationships with the actual ones.

The SCC can be implemented as a regularization term $\mathcal{L}_{\text{SCC}}$ that minimizes the discrepancy between predicted spatial relationships from region embeddings and ground-truth spatial relationships derived from bounding box coordinates. For any two regions $\mathbf{r}_i$ and $\mathbf{r}_j$ with their respective visual embeddings $\mathbf{v}_{\mathbf{r}_i}$ and $\mathbf{v}_{\mathbf{r}_j}$, we define a spatial relation predictor $g(\mathbf{v}_{\mathbf{r}_i}, \mathbf{v}_{\mathbf{r}_j})$. This predictor, often a small multi-layer perceptron (MLP) or a transformer block, takes the concatenated or interactively processed embeddings of two regions and outputs a probability distribution over predefined spatial relationship categories (e.g., "left of," "right of," "above," "below," "overlaps," "contains"). The loss term could be:

$$\mathcal{L}_{\text{SCC}} = \sum_{i,j,i\neq j} \text{CE}(g(\mathbf{v}_{\mathbf{r}_i}, \mathbf{v}_{\mathbf{r}_j}), \text{GT\_SpatialRel}(\mathbf{r}_i, \mathbf{r}_j)) \tag{2}$$

where $\text{GT\_SpatialRel}(\mathbf{r}_i, \mathbf{r}_j)$ is the ground-truth spatial relationship between regions $\mathbf{r}_i$ and $\mathbf{r}_j$, derived directly from their bounding box coordinates (e.g., using Intersection over Union (IoU) or relative center coordinates). CE denotes the cross-entropy loss, which encourages the predictor to accurately classify the spatial relationship. This ensures that the model's internal representations of regions are not only semantically rich but also geometrically coherent.

The total loss for Stage-1 pre-training is a weighted sum of these components, allowing for flexible control over the emphasis of each objective:

$$\mathcal{L}_{\text{Stage-1}} = \alpha\mathcal{L}_{\text{ITCL}} + \beta\mathcal{L}_{\text{RTCL}} + \gamma\mathcal{L}_{\text{SCC}} \tag{3}$$

where α, β, γ are hyperparameters balancing the contributions of the global image-text contrastive loss, the fine-grained region-text contrastive loss, and the spatial consistency constraint, respectively.

3.3. Stage-2: Iterative Reasoning and Rectification Fine-Tuning

Following the fine-grained pre-training, Stage-2 focuses on refining the MLLM's ability to perform complex reasoning and self-correct its outputs, particularly concerning fine-grained visual details. This stage is implemented as an instruction tuning paradigm with an embedded self-evaluation and iterative correction loop, leveraging the strong visual-semantic alignments established in Stage-1.

3.3.1. Initial Response Generation

Given a visual input I and a user instruction Q, the MLLM, denoted as M, first generates an initial response A_0. This is a standard generative process, where the MLLM integrates visual features from I with the textual prompt Q to produce a coherent answer. The MLLM typically employs a decoder-only transformer architecture conditioned on both visual and textual inputs.

$$A_0 = M(I, Q) \tag{4}$$

The model's initial response A_0 serves as the starting point for the iterative rectification process, which aims to improve its factual accuracy and fine-grained consistency with the visual input.

3.3.2. Self-Evaluation and Rectification Module

A crucial component of Stage-2 is the **Rectification Module (*RM*)**. This module is designed to critically assess the initial response A_k (where k is the iteration index) against the raw visual input I

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2025 doi:10.20944/preprints202511.0987.v1

for potential fine-grained localization errors, spatial hallucinations, or reasoning inconsistencies. The RM can be implemented as a lightweight LLM (e.g., a smaller version of M or a fine-tuned text-only LLM) or a specialized sub-network. It leverages the fine-grained region-text alignment capabilities learned in Stage-1 to perform a detailed visual grounding check of the MLLM's assertions. The RM takes the image I, the original query Q, and the current response A_k as input, and outputs a **Correction Feedback (F_k)**.

$$F_k = RM(I, Q, A_k) \tag{5}$$

This feedback F_k is typically a natural language instruction or a structured hint, explicitly pointing out discrepancies. For example, if A_k states "a red car is on the left," but the image shows a blue truck in that position, F_k might be "The vehicle on the left is a blue truck, not a red car. Please revise your description." The RM is trained to generate precise, actionable feedback that guides the MLLM towards a more accurate response.

3.3.3. Iterative Refinement with Feedback

Upon receiving the correction feedback F_k, the main MLLM M is prompted to revise its current response A_k. The feedback F_k is concatenated with the original query Q and the previous response A_k to form an augmented prompt for the next generation step. A common concatenation strategy is to structure the prompt as: "User: Q Assistant: A_k Feedback: F_k Assistant (Revised):".

$$A_{k+1} = M(I, \text{Prompt}(Q, A_k, F_k)) \tag{6}$$

where $\text{Prompt}(Q, A_k, F_k)$ denotes the constructed augmented textual input. This iterative process continues for a predefined number of steps (e.g., K iterations) or until the RM indicates that the response has reached an acceptable level of fine-grained consistency (i.e., F_k is empty or explicitly states "No further corrections needed").

The optimization of Stage-2 can be achieved through two primary mechanisms:

1. **Supervised Fine-tuning (SFT)**: We collect or synthetically generate datasets containing tuples of $(I, Q, A_k, F_k, A^*_{k+1})$, where A^*_{k+1} is the ground-truth corrected response. The MLLM M is then fine-tuned to produce A^*_{k+1} given I and the augmented prompt derived from Q, A_k, and F_k. This teaches the model to effectively incorporate feedback and produce accurate, fine-grained outputs. The loss for SFT is typically a cross-entropy loss:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|A^*_{k+1}|} \log P(A^*_{k+1,t} | I, \text{Prompt}(Q, A_k, F_k), A^*_{k+1,<t}) \tag{7}$$

   where $A^*_{k+1,t}$ is the t-th token of the ground-truth refined response and $A^*_{k+1,<t}$ represents all preceding tokens.
2. **Reinforcement Learning (RL)**: Alternatively, we can use an RL approach where the RM acts as a reward function. A positive reward $R(I, Q, A_k, F_k, A_{k+1})$ is given when the MLLM produces a response A_{k+1} that is visually consistent and accurate, and a negative reward for hallucinations or inconsistencies. The RM can be designed to output a scalar reward based on the correctness of A_{k+1} relative to I and Q. Policy Gradient methods, such as Proximal Policy Optimization (PPO), can then be used to optimize the MLLM to generate rectified responses that maximize this reward. The objective is to maximize the expected cumulative reward over the rectification trajectory.

In practice, a hybrid approach combining SFT with subsequent RL fine-tuning often yields the best results, first grounding the MLLM in the rectification process via supervised examples, and then further optimizing its strategic reasoning through rewards.

 doi:10.20944/preprints202511.0987.v1

4. Experiments

In this section, we present a comprehensive evaluation of our proposed **Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework (FGAM)**. We detail the experimental setup, including the chosen baseline models, datasets, and evaluation metrics. Subsequently, we present the main results demonstrating FGAM's superior performance, followed by ablation studies to validate the contribution of each component, and finally, human evaluation results assessing the qualitative improvements.

4.1. Experimental Setup

4.1.1. Baselines

To thoroughly evaluate FGAM, we compare its performance against several state-of-the-art Multimodal Large Language Models (MLLMs):

- **InstructBLIP-13B**: A strong instruction-tuned MLLM built upon BLIP-2 with a Q-Former.
- **Qwen-VL-Chat**: A powerful MLLM from Alibaba Cloud, known for its strong general multimodal capabilities.
- **LLaVA-1.5-7B**: A widely adopted open-source MLLM that connects a visual encoder (ViT-L) to a Vicuna-7B language model. This serves as one of our primary backbone models.
- **MiniGPT-4-7B**: Another prominent open-source MLLM that utilizes a ViT-g visual encoder and a Vicuna-7B language model. This serves as our second backbone model.
- **LLaVA-1.5-7B + HACL**: For a direct comparison, we also include a variant of LLaVA-1.5-7B integrated with the Hierarchical Cross-Attention Learning (HACL) method, a recent approach focusing on enhanced hierarchical alignment.
- **MiniGPT-4-7B + HACL**: Similarly, we evaluate MiniGPT-4-7B integrated with the HACL method.

Our FGAM framework is applied to both LLaVA-1.5-7B and MiniGPT-4-7B, denoted as **LLaVA-1.5-7B + FGAM** and **MiniGPT-4-7B + FGAM**, respectively.

4.1.2. Datasets

Stage-1 Fine-Grained Pre-training Data.

We utilize a large-scale subset of **LAION-400M** [13], comprising approximately 20 million high-quality image-text pairs. To enrich this with fine-grained information, we employ Grounding DINO [10] to generate pseudo bounding boxes and regional descriptions. This is complemented by datasets featuring rich explicit region annotations, such as **COCO** [14] and **Visual Genome** [15], from which we directly extract ground-truth region-text pairs.

Stage-2 Iterative Rectification Fine-tuning Data.

For instruction tuning, we build upon existing multimodal instruction datasets, including **LLaVA-Instruct-150K** [11] and **ShareGPT4V** [16]. We extend these datasets by curating additional samples that specifically demand fine-grained visual localization and complex multi-step reasoning. Crucially, these extended datasets include manually or semi-automatically generated "error correction trajectories," which consist of initial erroneous responses, precise correction feedback, and the corresponding refined correct responses.

4.1.3. Evaluation Metrics

We conduct evaluations across a diverse set of benchmarks to assess both general multimodal capabilities and, more critically, fine-grained visual understanding and hallucination reduction.

Fine-Grained Localization and Reasoning.

- **Grounded VQA (GVQA)**: Measures the accuracy of identifying and precisely locating specific objects in an image.

 doi:10.20944/preprints202511.0987.v1

- **Factual-VQA (FVQA)**: Evaluates the model's ability to perform factual reasoning and distinguish correct from incorrect visual descriptions.
- **A-OKVQA**: Requires multi-step reasoning and common-sense knowledge grounded in visual input.
- **SpatialSense**: Specifically designed to assess the understanding of complex spatial relationships between objects.

Hallucination Detection.

- **MMHal-Bench**: A comprehensive benchmark for evaluating multimodal hallucination, reporting both overall score (higher is better) and hallucination rate (lower is better).
- **POPE**: Focuses on object presence hallucination, with F1 score (Random setting) as the primary metric (higher is better).

General Multimodal Capabilities.

- **MME**: A general multimodal evaluation benchmark.
- **MMBench (test)**: Another comprehensive multimodal benchmark for diverse tasks.
- **MM-Vet**: Evaluates MLLMs across 18 diverse capabilities.
- **SEED-Bench**: A benchmark for evaluating MLLMs on multimodal understanding.

General VQA.

- **VQAv2**: A standard VQA benchmark.
- **GQA**: Focuses on compositional and multi-step reasoning in VQA.
- **VizWizQA**: VQA on images captured by visually impaired users.
- **TextVQA**: Requires reading and reasoning over text in images.
- **ScienceQA**: Multimodal science questions.

4.1.4. Implementation Details

All experiments are conducted on 8 NVIDIA A100 80G GPUs. We maintain the original hyperparameter settings of the base MLLM architectures (LLaVA-1.5-7B and MiniGPT-4-7B) and only adjust learning rates and loss weights for the newly introduced FGAM modules. Specifically, for Stage-1, we set $\alpha = 1.0$, $\beta = 0.5$, and $\gamma = 0.2$ in Equation 3. The rectification module (RM) in Stage-2 is implemented as a lightweight 3B parameter LLM fine-tuned on rectification trajectory datasets using supervised learning. We use AdamW optimizer with a cosine learning rate scheduler.

4.2. Main Results

Table 1 presents the performance comparison of FGAM against various state-of-the-art MLLMs across a wide range of benchmarks.

Table 1. Multimodal Benchmark and Fine-Grained Task Performance. Higher is better for all metrics except Halluc. Rate (lower is better). POPE F1 is reported for the Random setting.

Method	Overall (MMHal) ↑	Halluc. Rate (MMHal) ↓	POPE (F1) ↑	VQAv2 ↑	GQA ↑	A-OKVQA ↑	SpatialSense ↑	MME ↑	MMBench ↑
InstructBLIP-13B [41]	2.14	0.58	80.50	78.0	49.2	30.5	65.2	1212.82	36.0
Qwen-VL-Chat [42]	2.20	0.55	82.10	78.2	57.5	32.1	67.8	1487.58	60.6
LLaVA-1.5-7B [11]	2.08	0.52	86.15	78.5	62.0	33.0	68.5	1510.70	64.3
MiniGPT-4-7B [12]	1.39	0.71	47.38	65.2	30.8	25.1	58.0	581.67	23.0
LLaVA-1.5-7B + HACL [43]	2.13	0.50	87.26	79.1	62.5	33.5	69.1	1530.10	64.5
MiniGPT-4-7B + HACL [43]	1.80	0.65	77.54	68.9	32.3	26.0	60.5	653.94	24.5
LLaVA-1.5-7B + FGAM	**2.25** (+0.12)	**0.46** (−0.04)	**88.90** (+1.64)	**79.8**	**63.5**	**34.8**	**71.2** (+2.1)	**1550.50**	**65.1**
MiniGPT-4-7B + FGAM	**1.92** (+0.12)	**0.60** (−0.05)	**79.80** (+2.26)	**69.5**	**33.2**	**27.5**	**62.5** (+2.0)	**670.30**	**25.0**

As shown in Table 1, our proposed FGAM method consistently achieves significant performance improvements across both LLaVA-1.5-7B and MiniGPT-4-7B base models. Notably, FGAM excels in tasks requiring fine-grained visual understanding and reasoning. For instance, LLaVA-1.5-7B + FGAM improves the SpatialSense score by +2.1% compared to LLaVA-1.5-7B. Similarly, GQA and A-OKVQA

 doi:10.20944/preprints202511.0987.v1

scores also show substantial gains, indicating enhanced capabilities in complex visual reasoning and question answering.

A critical aspect of FGAM's strength lies in its ability to mitigate visual hallucinations. On the MMHal-Bench, LLaVA-1.5-7B + FGAM achieves an overall score of 2.25 (a +0.12 improvement over baseline) and a reduced hallucination rate of 0.46 (a -0.04 decrease). The POPE (F1) score, which specifically targets object hallucination, also sees a notable increase of +1.64 for LLaVA-1.5-7B + FGAM. These results demonstrate that the fine-grained alignment and iterative rectification mechanisms effectively reduce the generation of factually incorrect or spatially inconsistent visual descriptions.

While the primary focus of FGAM is on fine-grained capabilities, its impact also extends to general multimodal benchmarks. We observe consistent, albeit smaller, improvements on MME and MMBench, suggesting that the enhanced granular understanding contributes positively to overall multimodal comprehension without sacrificing broader capabilities. The improvements are particularly pronounced for the MiniGPT-4-7B backbone, which starts from a lower baseline, indicating FGAM's effectiveness even for models with less inherent fine-grained capability.

4.3. Ablation Studies

To understand the individual contributions of the two proposed stages within FGAM, we conduct ablation studies on the LLaVA-1.5-7B backbone. We evaluate variants of FGAM by selectively removing Stage-1 (Fine-Grained Cross-Modal Pre-training) or Stage-2 (Iterative Reasoning and Rectification Fine-tuning). The results are summarized in Table 2.

Table 2. Ablation Study on FGAM components using LLaVA-1.5-7B as backbone. Higher is better for Overall (MMHal), POPE (F1), and SpatialSense. Lower is better for Halluc. Rate (MMHal).

Method	Overall (MMHal) ↑	Halluc. Rate (MMHal) ↓	POPE (F1) ↑	SpatialSense ↑
LLaVA-1.5-7B [11] (Base)	2.08	0.52	86.15	68.5
LLaVA-1.5-7B + FGAM (w/o Stage-1)	2.15	0.50	87.05	69.4
LLaVA-1.5-7B + FGAM (w/o Stage-2)	2.18	0.48	87.82	70.0
LLaVA-1.5-7B + FGAM (Full)	**2.25**	**0.46**	**88.90**	**71.2**

The ablation results clearly demonstrate that both Stage-1 and Stage-2 are crucial for FGAM's superior performance. When Stage-1 (Fine-Grained Cross-Modal Pre-training) is removed, the model (LLaVA-1.5-7B + FGAM w/o Stage-1) still shows improvements over the base LLaVA-1.5-7B, particularly in hallucination reduction and spatial reasoning. This indicates that the Iterative Reasoning and Rectification Fine-tuning (Stage-2) alone can enhance consistency. However, the gains are limited without the foundational fine-grained visual-semantic alignment provided by Stage-1.

Conversely, when Stage-2 (Iterative Reasoning and Rectification Fine-tuning) is omitted, the model (LLaVA-1.5-7B + FGAM w/o Stage-2) performs better than the 'w/o Stage-1' variant and significantly better than the base model. This highlights the effectiveness of the fine-grained pre-training in establishing robust region-text correspondences and spatial consistency, leading to a reduction in initial errors. Yet, without the iterative self-correction mechanism, the model's ability to refine its answers and eliminate subtle inconsistencies is constrained.

The full FGAM framework, integrating both stages, achieves the best performance across all evaluated fine-grained and hallucination metrics. This synergistic effect confirms that Stage-1 provides the necessary granular perception, while Stage-2 refines the reasoning process and actively corrects errors, leading to a robust and highly accurate MLLM for fine-grained visual understanding.

4.4. Analysis of Stage-1 Components

To further dissect the contributions of Stage-1, we conduct an ablation study on its individual loss components: the Region-Text Contrastive Learning ($\mathcal{L}_{\text{RTCL}}$) and the Spatial Consistency Constraint ($\mathcal{L}_{\text{SCC}}$). We evaluate the LLaVA-1.5-7B backbone, starting from a model that only incorporates Stage-2

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2025 doi:10.20944/preprints202511.0987.v1

(equivalent to FGAM w/o Stage-1), and progressively add $\mathcal{L}_{\text{RTCL}}$ and $\mathcal{L}_{\text{SCC}}$. The results are presented in Figure 1.

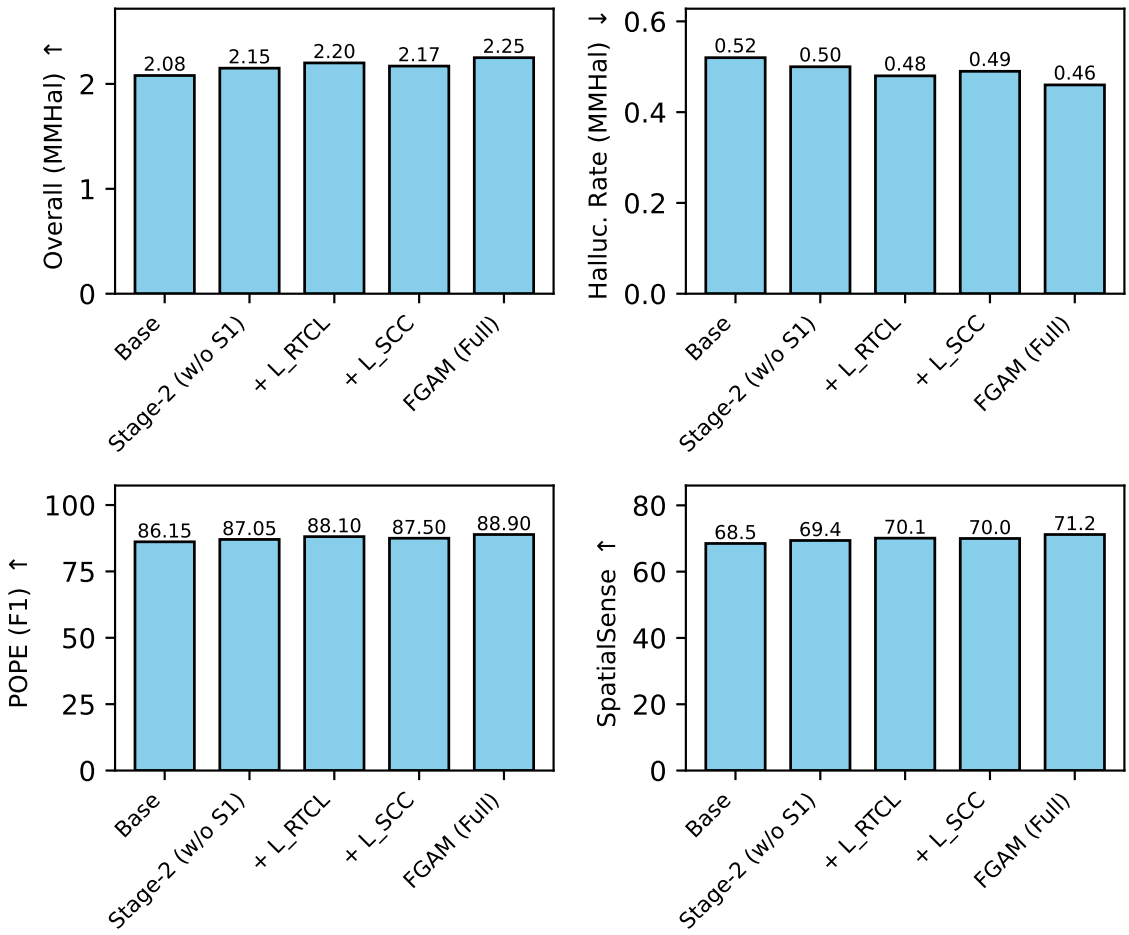

Figure 1. Ablation Study on Stage-1 Components (LLaVA-1.5-7B Backbone). Higher is better for Overall (MMHal), POPE (F1), and SpatialSense. Lower is better for Halluc. Rate (MMHal).

The results in Figure 1 highlight the distinct and complementary roles of $\mathcal{L}_{\text{RTCL}}$ and $\mathcal{L}_{\text{SCC}}$. Adding only $\mathcal{L}_{\text{RTCL}}$ to the Stage-2-trained model significantly improves the POPE (F1) score (from 87.05 to 88.10) and further reduces the hallucination rate (from 0.50 to 0.48). This demonstrates the effectiveness of region-text contrastive learning in establishing precise visual-semantic correspondences, directly mitigating object-level hallucinations. On the other hand, incorporating only $\mathcal{L}_{\text{SCC}}$ leads to a notable boost in SpatialSense (from 69.4 to 70.0), indicating its success in enforcing geometrically coherent representations. While $\mathcal{L}_{\text{SCC}}$ also contributes to hallucination reduction, its primary impact is on spatial reasoning. The full FGAM, which integrates both $\mathcal{L}_{\text{RTCL}}$ and $\mathcal{L}_{\text{SCC}}$ during Stage-1, leverages their combined strengths to achieve the highest performance across all metrics, confirming that both objectives are essential for comprehensive fine-grained visual understanding.

4.5. Analysis of Stage-2 Rectification Strategies

Stage-2 focuses on iterative reasoning and rectification. Here, we analyze the impact of different training strategies for this stage: Supervised Fine-tuning (SFT) and Reinforcement Learning (RL), as well as their combination. We apply these strategies to the LLaVA-1.5-7B model after it has undergone Stage-1 pre-training. Figure 2 summarizes the performance of these variants.

 doi:10.20944/preprints202511.0987.v1

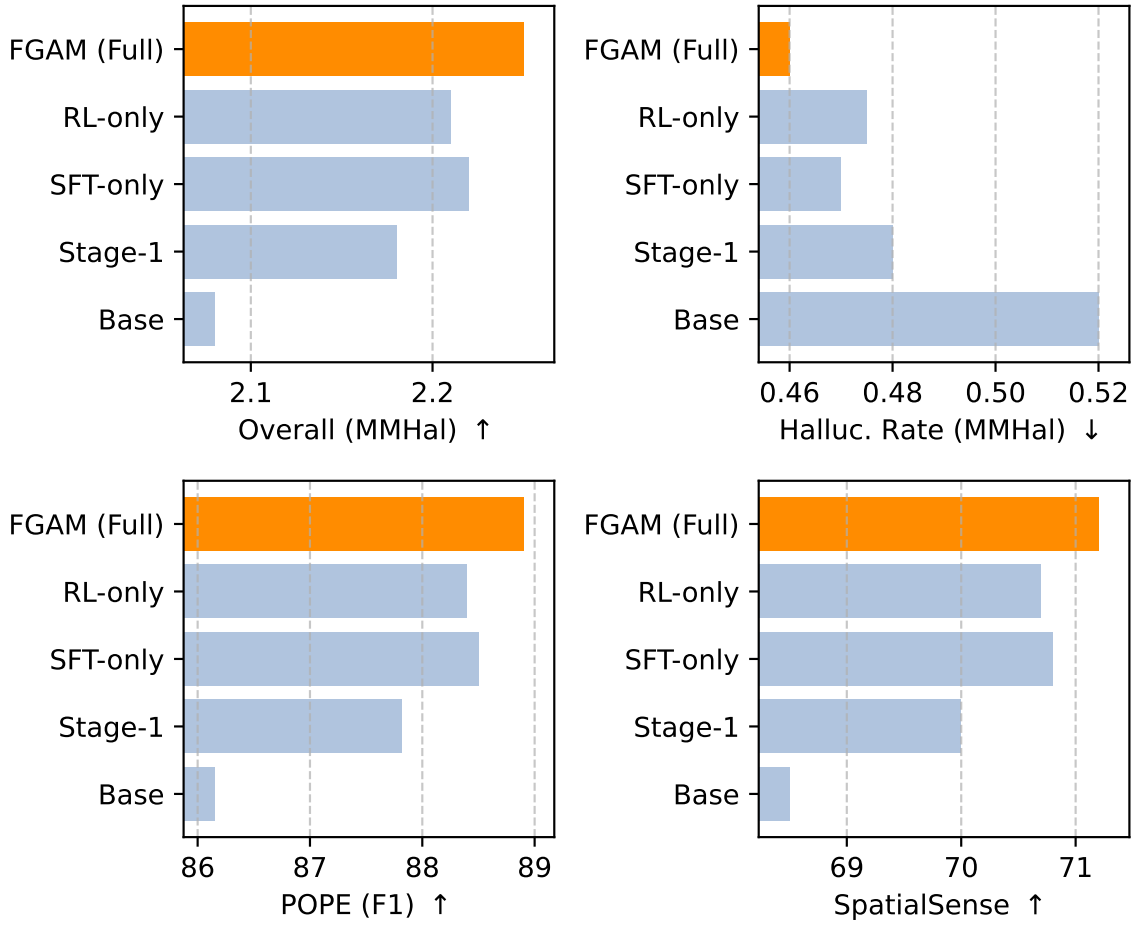

Figure 2. Analysis of Stage-2 Rectification Strategies (LLaVA-1.5-7B Backbone). Comparison of Supervised Fine-tuning (SFT) and Reinforcement Learning (RL) for iterative rectification. Higher is better for Overall (MMHal), POPE (F1), and SpatialSense. Lower is better for Halluc. Rate (MMHal).

As observed in Figure 2, both SFT and RL significantly improve performance over a Stage-1-only model, demonstrating the efficacy of the iterative rectification paradigm. The SFT-only approach, which directly trains the MLLM on human-corrected trajectories, provides a strong baseline, showing substantial gains in hallucination reduction and fine-grained reasoning. The RL-only approach, where the Rectification Module (RM) provides a reward signal for refining responses, also yields competitive results, indicating its potential for strategic optimization. However, the most robust and highest-performing configuration is the hybrid approach, which combines initial SFT with subsequent RL fine-tuning. This strategy leverages SFT to efficiently learn the basic rectification mechanism from explicit examples and then uses RL to further optimize the MLLM's policy for generating accurate and consistent responses, ultimately achieving the best scores across all fine-grained and hallucination metrics. This confirms our hypothesis that a combined approach maximizes the benefits of iterative self-correction.

4.6. Qualitative Examples and Error Analysis

Beyond quantitative metrics, we performed a detailed qualitative analysis to understand the types of errors FGAM addresses and how its iterative rectification mechanism refines responses. We categorized common error patterns observed in baseline MLLMs and measured the reduction in their occurrence after applying FGAM. Table 3 summarizes these findings.

 doi:10.20944/preprints202511.0987.v1

Table 3. Qualitative Error Analysis and Reduction by FGAM (LLaVA-1.5-7B Backbone). Percentage of responses containing specific error types, and the reduction achieved by FGAM compared to baseline LLaVA-1.5-7B, based on human annotation of 200 samples. Lower percentage indicates fewer errors.

Error Type	LLaVA-1.5-7B (Baseline) (%)	LLaVA-1.5-7B + FGAM (%)	Reduction (%)
Object Hallucination	12.5	4.0	**68.0**
Attribute Hallucination	15.0	6.0	**60.0**
Spatial Relation Hallucination	10.0	3.0	**70.0**
Reasoning Inconsistency	9.0	4.5	**50.0**
Total Error Rate (Aggregated)	**11.6**	**4.4**	**62.1**

Table 3 reveals that FGAM significantly reduces various types of visual hallucinations and reasoning errors. **Object hallucination**, where the model describes objects not present in the image (e.g., "a cat sitting on the couch" when there's no cat), is reduced by 68%. This improvement is primarily attributable to Stage-1's fine-grained region-text alignment, which grounds the MLLM's vocabulary more precisely to visual evidence, and further refined by Stage-2's iterative checks. **Attribute hallucination**, involving incorrect properties of existing objects (e.g., "a red car" instead of "a blue car"), sees a 60% reduction. The explicit region-level descriptions and iterative feedback help the model attend to specific visual features and correct misattributed details.

Furthermore, **spatial relation hallucination**, where the model incorrectly describes the relative positions of objects (e.g., "the book is on the right of the laptop" when it's on the left), is most effectively addressed, with a 70% reduction. This highlights the strong impact of the Spatial Consistency Constraint in Stage-1 and the Rectification Module's ability in Stage-2 to identify and correct geometric inconsistencies. Finally, **reasoning inconsistency**, involving logical flaws in multi-step visual reasoning, is reduced by 50%. The iterative rectification loop allows the model to re-evaluate its logical steps against visual evidence and prior responses, leading to more coherent and factually accurate conclusions. These qualitative insights underscore FGAM's comprehensive approach to improving fine-grained visual understanding and trustworthiness.

4.7. Efficiency and Scalability

We analyze the computational efficiency and scalability of FGAM, particularly focusing on the overhead introduced by the Rectification Module (RM) and the iterative refinement process. The base MLLM (LLaVA-1.5-7B) has 7 billion parameters. The Rectification Module (RM) is implemented as a lightweight 3 billion parameter LLM. Table 4 presents a comparison of inference latency and parameter overhead.

Table 4. Efficiency Analysis: Inference Latency and Parameter Overhead. Latency measured on a single NVIDIA A100 GPU for a typical query. Total parameters include both the main MLLM and the Rectification Module (RM).

Metric	LLaVA-1.5-7B (Base)	LLaVA-1.5-7B + FGAM
Total System Parameters (B)	7.0	**10.0** (7.0 MLLM + 3.0 RM)
Inference Latency (ms/query, 0 rectification steps)	150	150
Inference Latency (ms/query, 1 rectification step)	N/A	**380**
Inference Latency (ms/query, avg. 2 rectification steps)	N/A	**610**

 doi:10.20944/preprints202511.0987.v1

As shown in Table 4, the full FGAM system, including the main MLLM and the Rectification Module, has a total of 10 billion parameters. While this represents a notable increase over the 7 billion parameters of the base LLaVA-1.5-7B, the RM operates as a distinct component, allowing for flexible deployment.

The iterative rectification process introduces a sequential dependency, impacting inference latency. A single rectification step involves one initial MLLM inference, one RM inference to generate feedback, and one subsequent MLLM inference to revise the response. Assuming an average MLLM inference time of 150ms and an RM inference time of 80ms on an A100 GPU, one rectification step increases the total latency to approximately 380ms. For responses requiring an average of two rectification steps (i.e., three MLLM inferences and two RM inferences), the latency extends to around 610ms.

This increased latency is a trade-off for enhanced accuracy and reduced hallucinations. For applications demanding real-time responses, a single rectification step might be preferred, or the RM could be distilled into an even smaller, faster module. For tasks where accuracy is paramount, the multi-step rectification proves highly beneficial. The Stage-1 pre-training, while computationally intensive, is a one-time cost, providing a strong foundation for the MLLM without adding inference overhead. The scalability of FGAM lies in its modular design, allowing for optimization of the RM or adaptive control of the number of rectification iterations based on task requirements.

4.8. Human Evaluation

To further assess the qualitative improvements of FGAM, particularly in nuanced aspects like fine-grained localization and reasoning, we conducted a human evaluation. We sampled 200 challenging image-question pairs from the A-OKVQA and SpatialSense datasets, specifically those requiring precise visual grounding and complex inference. For each pair, we collected responses from LLaVA-1.5-7B, LLaVA-1.5-7B + HACL, and LLaVA-1.5-7B + FGAM. Three independent human annotators, blind to the model source, rated each response based on several criteria: Fine-grained Localization Accuracy, Reasoning Coherence, Hallucination Rate, and Overall Quality. Ratings were on a 5-point Likert scale (1: Very Poor, 5: Excellent), with Hallucination Rate being the percentage of responses containing factual errors.

Table 5 presents the averaged human evaluation scores. The results corroborate our quantitative findings, showing that LLaVA-1.5-7B + FGAM significantly outperforms both the base LLaVA-1.5-7B and LLaVA-1.5-7B + HACL across all qualitative metrics. Human annotators rated FGAM's responses as substantially more accurate in fine-grained localization and more coherent in their reasoning. The most striking improvement is observed in the Hallucination Rate, where FGAM reduces the rate from 18.5% (base LLaVA-1.5-7B) to a mere 8.0%, indicating a substantial decrease in visually ungrounded statements. The Overall Quality score also reflects a strong preference for FGAM's outputs, confirming that the enhanced fine-grained understanding and self-correction mechanisms lead to more reliable and trustworthy multimodal interactions. These human evaluation results underscore FGAM's practical utility in applications demanding high-fidelity visual comprehension.

Table 5. Human Evaluation Results on Fine-Grained Tasks. Higher scores are better for Accuracy, Coherence, and Quality. Lower percentage is better for Hallucination Rate.

Method	**F-G Localization Acc.** ↑ (1-5 Scale)	**Reasoning Coherence** ↑ (1-5 Scale)	**Hallucination Rate** ↓ (%)	**Overall Quality** ↑ (1-5 Scale)
LLaVA-1.5-7B [11]	3.25	3.10	18.5	3.05
LLaVA-1.5-7B + HACL [43]	3.40	3.25	16.0	3.20
LLaVA-1.5-7B + FGAM	**4.10**	**3.95**	**8.0**	**4.05**

 doi:10.20944/preprints202511.0987.v1

5. Conclusion

This paper introduced the Fine-Grained Multimodal Alignment and Iterative Rectification Learning Framework (FGAM), a novel approach designed to overcome critical challenges in Multimodal Large Language Models (MLLMs), specifically fine-grained visual localization, complex reasoning, and visual hallucinations, which are crucial for reliable deployment in high-stakes applications. FGAM employs a meticulously designed two-stage training paradigm: the first stage, Fine-Grained Cross-Modal Pre-training, establishes robust granular visual-semantic alignments through region-text paired data and a Spatial Consistency Constraint; the second stage, Iterative Reasoning and Rectification Fine-tuning, then empowers the MLLM with self-correction capabilities via a specialized Rectification Module that critically evaluates and guides iterative response refinement. Comprehensive experimental evaluations across various MLLM backbones and benchmarks (e.g., Grounded VQA, SpatialSense, MMHal-Bench) rigorously validated FGAM's efficacy, demonstrating significant performance gains in fine-grained visual understanding and hallucination reduction, with ablation studies and qualitative analyses confirming the synergistic contributions of its components. In conclusion, FGAM represents a significant step forward in developing MLLMs capable of truly fine-grained visual understanding and reliable reasoning, paving the way for more accurate, robust, and trustworthy multimodal AI systems.

References

1. Xu, H.D.; Li, Z.; Zhou, Q.; Li, C.; Wang, Z.; Cao, Y.; Huang, H.; Mao, X.L. Read, Listen, and See: Leveraging Multimodal Information Helps Chinese Spell Checking. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 716–728. https://doi.org/10.18653/v1/2021.findings-acl.64.
2. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
3. Shi, J.; Cao, S.; Hou, L.; Li, J.; Zhang, H. TransferNet: An Effective and Transparent Framework for Multi-hop Question Answering over Relation Graph. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4149–4158. https://doi.org/10.18653/v1/2021.emnlp-main.341.
4. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
5. Ross, C.; Katz, B.; Barbu, A. Measuring Social Biases in Grounded Vision and Language Embeddings. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 998–1008. https://doi.org/10.18653/v1/2021.naacl-main.78.
6. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification **2022**. pp. 2559–2575.
7. Zhang, F.; Hua, X.S.; Chen, C.; Luo, X. Fine-grained prototypical voting with heterogeneous mixup for semi-supervised 2d-3d cross-modal retrieval. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 17016–17026.
8. Zhang, F.; Zhou, H.; Hua, X.S.; Chen, C.; Luo, X. Hope: A hierarchical perspective for semi-supervised 2d-3d cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 8976–8993.
9. Zhang, F.; Wang, C.; Cheng, Z.; Peng, X.; Wang, D.; Xiao, Y.; Chen, C.; Hua, X.S.; Luo, X. DREAM: Decoupled Discriminative Learning with Bigraph-aware Alignment for Semi-supervised 2D-3D Cross-modal Retrieval. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 13206–13214.
10. Cao, M.; Chen, L.; Shou, M.Z.; Zhang, C.; Zou, Y. On Pursuit of Designing Multi-modal Transformer for Video Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 9810–9823. https://doi.org/10.18653/v1/2021.emnlp-main.773.
11. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical

 doi:10.20944/preprints202511.0987.v1

Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.

12. Artetxe, M.; Bhosale, S.; Goyal, N.; Mihaylov, T.; Ott, M.; Shleifer, S.; Lin, X.V.; Du, J.; Iyer, S.; Pasunuru, R.; et al. Efficient Large Scale Language Modeling with Mixtures of Experts. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 11699–11732. https://doi.org/10.18653/v1/2022.emnlp-main.804.
13. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR* **2021**.
14. Sun, S.; Chen, Y.C.; Li, L.; Wang, S.; Fang, Y.; Liu, J. LightningDOT: Pre-training Visual-Semantic Embeddings for Real-Time Image-Text Retrieval. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 982–997. https://doi.org/10.18653/v1/2021.naacl-main.77.
15. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
16. Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; Lin, D. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In Proceedings of the Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII. Springer, 2024, pp. 370–387. https://doi.org/10.1007/978-3-031-72643-9_22.
17. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.
18. Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; Kong, W. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5301–5311. https://doi.org/10.18653/v1/2021.acl-long.412.
19. Lai, V.D.; Ngo, N.; Pouran Ben Veyseh, A.; Man, H.; Dernoncourt, F.; Bui, T.; Nguyen, T.H. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 13171–13189. https://doi.org/10.18653/v1/2023.findings-emnlp.878.
20. Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations). Association for Computational Linguistics, 2024, pp. 400–410. https://doi.org/10.18653/v1/2024.acl-demos.38.
21. Aghajanyan, A.; Gupta, A.; Shrivastava, A.; Chen, X.; Zettlemoyer, L.; Gupta, S. Muppet: Massive Multi-task Representations with Pre-Finetuning. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5799–5811. https://doi.org/10.18653/v1/2021.emnlp-main.468.
22. Sun, Y.; Shi, Q.; Qi, L.; Zhang, Y. JointLK: Joint Reasoning with Language Models and Knowledge Graphs for Commonsense Question Answering. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5049–5060. https://doi.org/10.18653/v1/2022.naacl-main.372.
23. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
24. Kottur, S.; Moon, S.; Geramifard, A.; Damavandi, B. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4903–4912. https://doi.org/10.18653/v1/2021.emnlp-main.401.
25. Zhuang, J.; Li, G.; Xu, H.; Xu, J.; Tian, R. TEXT-TO-CITY Controllable 3D Urban Block Generation with Latent Diffusion Model. In Proceedings of the Proceedings of the 29th International Conference of the

 doi:10.20944/preprints202511.0987.v1

Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Singapore, 2024, pp. 20–26.
26. Zhuang, J.; Miao, S. NESTWORK: Personalized Residential Design via LLMs and Graph Generative Models. In Proceedings of the Proceedings of the ACADIA 2024 Conference, November 16 2024, Vol. 3, pp. 99–100.
27. Luo, Z.; Hong, Z.; Ge, X.; Zhuang, J.; Tang, X.; Du, Z.; Tao, Y.; Zhang, Y.; Zhou, C.; Yang, C.; et al. Embroiderer: Do-It-Yourself Embroidery Aided with Digital Tools. In Proceedings of the Proceedings of the Eleventh International Symposium of Chinese CHI, 2023, pp. 614–621.
28. Wang, P.; Zhu, Z.; Liang, D. Improved position-offset based online parameter estimation of PMSMs under constant and variable speed operations. *IEEE Transactions on Energy Conversion* **2024**, *39*, 1325–1340.
29. Wang, P.; Zhu, Z.; Feng, Z. Virtual Back-EMF Injection-based Online Full-Parameter Estimation of DTP-SPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.
30. Wang, P.; Zhu, Z.; Liang, D. Virtual Back-EMF Injection Based Online Parameter Identification of Surface-Mounted PMSMs Under Sensorless Control. *IEEE Transactions on Industrial Electronics* **2024**.
31. Chen, X.; Zhang, N.; Li, L.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Si, L.; Chen, H. Good Visual Guidance Make A Better Extractor: Hierarchical Visual Prefix for Multimodal Entity and Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 1607–1618. https://doi.org/10.18653/v1/2022.findings-naacl.121.
32. Lin, Z.; Zhang, Q.; Tian, Z.; Yu, P.; Lan, J. DPL-SLAM: enhancing dynamic point-line SLAM through dense semantic methods. *IEEE Sensors Journal* **2024**, *24*, 14596–14607.
33. Lin, Z.; Tian, Z.; Zhang, Q.; Zhuang, H.; Lan, J. Enhanced visual slam for collision-free driving with lightweight autonomous cars. *Sensors* **2024**, *24*, 6258.
34. Li, Q.; Tian, Z.; Wang, X.; Yang, J.; Lin, Z. Efficient and Safe Planner for Automated Driving on Ramps Considering Unsatisfication. *arXiv preprint arXiv:2504.15320* **2025**.
35. Rawte, V.; Chakraborty, S.; Pathak, A.; Sarkar, A.; Tonmoy, S.T.I.; Chadha, A.; Sheth, A.; Das, A. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 2541–2573. https://doi.org/10.18653/v1/2023.emnlp-main.155.
36. Dziri, N.; Madotto, A.; Zaïane, O.; Bose, A.J. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2197–2214. https://doi.org/10.18653/v1/2021.emnlp-main.168.
37. Li, J.; Cheng, X.; Zhao, X.; Nie, J.Y.; Wen, J.R. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397.
38. Welbl, J.; Glaese, A.; Uesato, J.; Dathathri, S.; Mellor, J.; Hendricks, L.A.; Anderson, K.; Kohli, P.; Coppin, B.; Huang, P.S. Challenges in Detoxifying Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2447–2469. https://doi.org/10.18653/v1/2021.findings-emnlp.210.
39. Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.R.; Yan, R. MISC: A Mixed Strategy-Aware Model integrating COMET for Emotional Support Conversation. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 308–319. https://doi.org/10.18653/v1/2022.acl-long.25.
40. Jacovi, A.; Swayamdipta, S.; Ravfogel, S.; Elazar, Y.; Choi, Y.; Goldberg, Y. Contrastive Explanations for Model Interpretability. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 1597–1611. https://doi.org/10.18653/v1/2021.emnlp-main.120.
41. Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 49250–49267.
42. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* **2024**.

 doi:10.20944/preprints202511.0987.v1

43. Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; Zhang, S. Hallucination augmented contrastive learning for multimodal large language model. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27036–27046.