

FUSEAGENT: A VLM-DRIVEN AGENT FOR UNIFIED IN-THE-WILD IMAGE FUSION

Anonymous authors

Paper under double-blind review

ABSTRACT

Fusing multi-source images captured in the wild is often undermined by unpredictable and coupled degradations, including pixel-level misalignment, adverse weather, and dynamic artifacts. Existing solutions face notable limitations: (1) Task-specific models rely on predefined degradation priors and fail to generalize to the complex, coupled degradations present in real-world scenarios. (2) All-in-one methods, while designed for multi-fusion tasks, frequently overlook the degradation inherent in input images, leading to suboptimal performance. To address these challenges, we introduce FuseAgent, a VLM-powered agent system that autonomously identifies degradations in the input images and dynamically coordinates expert models to execute a tailored fusion strategy. FuseAgent undergoes a two-stage training process: an initial supervised fine-tuning (SFT) establishes basic degradation perception and tool-use skill, followed by Group Relative Policy Optimization for fusion (GRPO-F) augmented with multi-dimensional rewards to further enhance its decision-making and tool proficiency. Experimental results demonstrate the superior performance of FuseAgent in handling complex and coupled degradations in real-world scenes, achieving a **20%** average improvement across all evaluation metrics on challenging in-the-wild benchmarks.

1 INTRODUCTION

Image fusion is critical for vision-centric applications—including autonomous driving, computational photography, and remote sensing—where multi-source images acquired in real-world scenarios often suffer from complex and intertwined degradations such as pixel misalignment, adverse weather conditions, and dynamic scene artifacts. As shown in Figure 1, existing solutions can be roughly classified into two categories: 1) Task-specific models show proficiency in individual degradation types, such as geometric deviations Huang et al. (2022); Wang et al. (2022) and motion artifacts Cao et al. (2023); Chen et al. (2025), but struggle with the dynamic and coupled degradations in real-world scenarios. 2) All-in-one approaches attempt to consolidate several fusion tasks into a single pipeline, yet frequently neglect the intrinsic unpredictable and mixed degradations of the input images, thereby limiting their practical applicability. To mitigate these challenges, an intuitive solution—shown in Figure 1(a)—is to combine expert models into a fixed pipeline. Yet two fundamental challenges persist: 1) Dynamic degradations, where inputs are affected by complex and coupled distortions that differ across samples, thereby undermining the effectiveness of predefined pipelines. 2) Task adaptability, where identifying the optimal sequence of expert models for each input is still an open challenge.

Recently, the rapid advancement of Large Language Models (LLMs)—which exhibit remarkable capabilities in reasoning, decision-making, and interacting with diverse environments (Patil et al., 2023; Jain, 2022; Yang et al., 2024; Zhao et al., 2025)—has inspired us to rethink image fusion from an agentic perspective. Specifically, an intelligent fusion agent could be *interpreting scene-specific degradations*, *coordinating expert competencies*, and *assembling optimal fusion workflows*. However, realizing such an agent in practice demands addressing three key challenges: 1) The scarcity of training data, particularly paired clean/degraded images and annotated expert sequences. 2) The difficulty of selecting suitable expert models within an enormous combinatorial space, and 3) The lack of infrastructure designed for training agentic fusion systems.

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

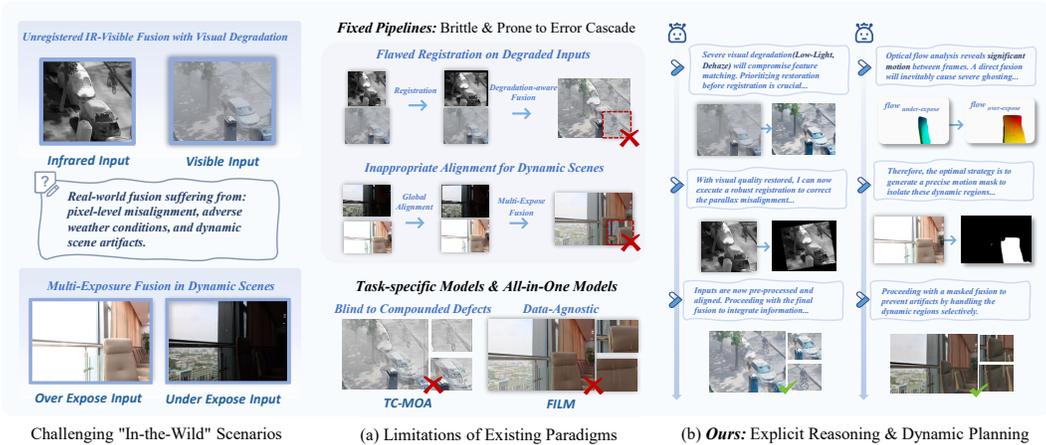


Figure 1: Limitations of task-specific methods, all-in-one methods, and human-defined processing pipelines. 1) Single-task and all-in-one methods fail to effectively handle coupled degradations in real-world scenarios. 2) Static processing pipelines, predefined by humans using multiple expert models, show promise but struggle with dynamic degradations. Moreover, the optimal task sequence remains unpredictable in real-world conditions. In contrast, FuseAgent can dynamically assess scene-specific degradations, coordinate expert models, and plan optimal fusion workflows, yielding superior fusion results.

To translate this agent-driven paradigm into a practical system, as depicted in Figure 1, we establish FuseAgent by employing a Vision-Language Model (VLM) as the agent’s core controller, endowing it with degradation perception, decision-making, and tool-use skills. We adopt a two-stage training regime: (1) Supervised Fine-Tuning (SFT) initially teaches the agent foundational skills, including degradation identification and tool invocation, using a curated set of expert trajectories. (2) Group Relative Policy Optimization for Fusion (GRPO-F) is applied to boost system generalization, reduce hallucinations, and improve decision-making in real-world situations. To ensure stability during the GRPO-F process, we introduce multi-dimensional rewards designed for tool-integrated fusion tasks: the Intrinsic Quality Reward (IQR) and the Relational Quality Reward (RQR). The IQR evaluates the perceptual quality of each fused image generated during the rollout, while the RQR quantifies the improvement in “fusion compatibility” between the source images and the fused output. These rewards—inter-image (IQR) and intra-image (RQR)—provide essential, unsupervised guidance for the agent’s evolution throughout the process.

Our main contributions are as follows:

- We introduce a novel fusion paradigm, FuseAgent, a VLM-powered agent that autonomously identifies degradations and coordinates specialized expert models to effectively handle complex and coupled degradations in real-world environments.
- We propose a novel two-stage framework combining SFT and GRPO-F, augmented with task-specific rewards—Intrinsic Quality Reward (IQR) and Relational Quality Reward (RQR)—to enable label-free reinforcement learning and improve generalizability in real-world settings.
- Comprehensive experiments demonstrate that FuseAgent excels in both image fusion quality and task-level decision-making, significantly outperforming advanced all-in-one baselines and static, manually designed pipelines.

2 RELATED WORK

Task-specific Image Fusion. To bridge the gap between idealized fusion assumptions and the complexities of real-world inputs, a significant body of research has focused on task-specific robust fusion. Early efforts concentrated on geometric robustness, leading to registration-free models (Arar et al., 2020; Wang et al., 2022; Xu et al., 2022). While effective at bypassing explicit registration, their narrow specialization renders them ineffective against photometric or dynamic inconsistencies. Another major line of research targets dynamic scenes, proposing specialized dehazing al-

gorithms (Cao et al., 2023; Chen et al., 2025) to mitigate motion artifacts. However, these methods are often predicated on the assumption of well-registered frames with consistent illumination, limiting their applicability in more varied scenarios. More recently, methods have emerged to tackle perceptual quality degradations. Approaches like Text-IF (Yi et al., 2024) and DRMF (Tang et al., 2024) aim to perform fusion while simultaneously addressing issues such as sensor noise, low resolution, or poor illumination. Despite their advances in joint restoration and fusion, they typically presuppose a static, pre-aligned scene. Collectively, while these task-specific approaches demonstrate expertise on isolated problems, they lack a unified framework to address the compounded and heterogeneous degradations endemic to in-the-wild data, as they are incapable of reasoning about the complex interplay of multiple, co-occurring defects.

All-in-One Image Fusion. All-in-one fusion paradigm aims to create a single, unified framework for diverse fusion tasks, moving beyond the single-problem focus of specialized models. Early works in this area concentrated on designing unified network architectures, employing powerful backbones like Transformers (Ma et al., 2022) or carefully designed objectives (Xu et al., 2020) to create a shared representation for multiple tasks. More recent approaches have evolved towards greater flexibility. One prominent trajectory leverages generative models to reframe fusion as a conditional synthesis task (Liang et al., 2022). Concurrently, another direction has focused on adaptive mechanisms, such as textual prompts (Zhao et al., 2024; Cao et al., 2025) or modular adapters (Zhu et al., 2024b; Cheng et al., 2025), to guide a core model toward specific fusion objectives. Despite their architectural diversity and task versatility, these all-in-one paradigms share a common limitation: their operational logic remains static. They lack an explicit perception and planning capability to dynamically react to unforeseen, input-level degradations (e.g., misalignment, adverse weather), rendering their fixed frameworks insufficiently robust for the challenges of in-the-wild data.

VLM-powered Agents. The limitations of static paradigms in vision tasks have motivated a shift toward dynamic, agent-based systems, a trend fueled by the recent success of Large Language Models (LLMs) in complex reasoning and tool use (Qin et al., 2023; Shen et al., 2024). Foundational works have demonstrated that LLMs and VLMs can be adapted into powerful controllers, capable of orchestrating diverse external tools and foundation models to solve complex user queries (Patil et al., 2023; Wu et al., 2023a). This agent-based paradigm is now being explored in low-level vision, particularly for image restoration. Current approaches in this area can be broadly categorized. One line of research utilizes powerful, off-the-shelf VLMs as zero-shot planners to coordinate restoration tools based on commonsense reasoning (Zhu et al., 2024a; Bai et al., 2025a). Another category focuses on fine-tuning VLMs on domain-specific, often synthetic, data to create expert controllers that can generate explicit, step-by-step execution plans (Chen et al., 2024). While these pioneering efforts validate the potential of visual agents, they have been exclusively focused on single-image restoration. Our work introduces this paradigm to the fundamentally different domain of multi-image fusion. This presents a unique set of challenges, requiring the agent not only to identify degradations but also to reason about complex inter-image relationships, such as geometric alignment, photometric consistency, and complementary information. To address this, FusionAgent is explicitly designed with a training methodology (SFT+GRPO-F) and a relational reward (RQA) tailored for learning these complex relational policies from unlabeled in-the-wild data.

3 METHODOLOGY

In this section, we first describe the overall workflow of FuseAgent (Sec. 3.1). Next, we introduce our data generation pipeline, which constructs a large-scale dataset, including expert reasoning traces and sequences for agentic fusion tasks. Finally, we detail the two-stage training framework for FuseAgent, comprising supervised fine-tuning and group relative policy optimization for fusion in an unsupervised manner (Sec. 3.3).

3.1 OVERVIEW

FuseAgent is a VLM-powered agent system designed for unified in-the-wild image fusion. As shown in Figure 2, FuseAgent autonomously perceives degradation, decomposes complex fusion tasks into subproblems, and assigns the most suitable expert model to each subtask. The pipeline comprises three principal stages: (1) **Perception and task decomposition**, where the agent first analyzes the input image pair to identify a complex set of coupled degradations, such as misalign-

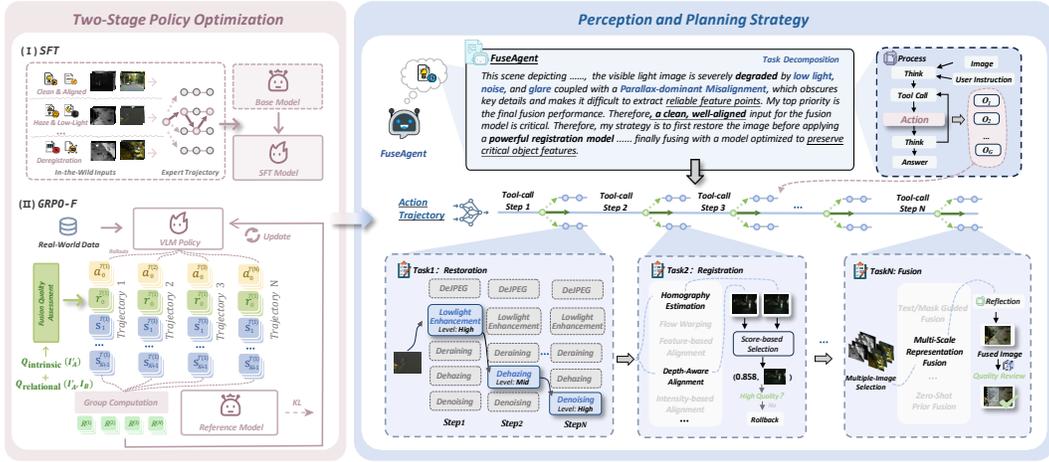


Figure 2: Overview of the two-stage optimization framework for FuseAgent. Initially, FuseAgent undergoes supervised fine-tuning (SFT) on expert-annotated trajectories to establish foundational skills in degradation perception and workflow planning. Following this, the Group Relative Policy Optimization for Fusion (GRPO-F) algorithm is applied to further enhance FuseAgent’s decision-making, tool proficiency, and robustness on unlabeled, in-the-wild data.

ment, low light, and dynamic artifacts. (2) **Reasoning and planning**, where the agent formulates multi-step fusion solutions referred to as the “Action Trajectory,” based on its expert-level fusion knowledge, determining task sequences and selecting the corresponding expert models. (3) **Tool execution**, where the agent executes the planned task and expert tool sequence to obtain the final fused results. Formally, FuseAgent defines a function as follows:

$$f(I_A, I_B) \rightarrow \mathcal{A} = \{a_1, a_2, \dots, a_n\},$$

where I_A and I_B are the source images, and \mathcal{A} is the planned Action Trajectory. Each action a_i represents a specific tool-call (e.g., Register(I_A, I_B) or Dehaze(I_A)). The final fused image is obtained by $I_{\text{fused}} = g(I_A, I_B, \mathcal{A})$, where $g(\cdot)$ represents the tool execution environment.

3.2 DATA GENERATION PIPELINE

Stage I: Generation of image pairs. Our data generation process employs a hybrid strategy, combining authentic degraded image pairs from public benchmarks with systematically synthesized data. While existing real-world datasets provide a crucial foundation of authenticity, we observe that they predominantly feature isolated or single-type degradations (e.g., containing motion artifacts or misalignment, but rarely both). To bridge this gap and ensure our agent is trained and evaluated on the compounded challenges truly representative of in-the-wild conditions, we augment this real data by synthesizing complex, multi-defect scenarios. This is accomplished using a modular degradation library capable of introducing and layering a wide range of fusion-specific defects. These include not only adverse weather and sensor noise, but also challenging geometric misalignments, photometric inconsistencies (e.g., exposure variations), and dynamic artifacts. This hybrid approach yields a dataset that captures a comprehensive spectrum of challenges, from simple, single-defect cases to complex, multi-degradation scenarios.

Stage II: Generation of expert responses (CoT & Actions). For each degraded image pair, we generate the corresponding expert response, which consists of two parts. (a) **CoT rationales:** To generate detailed reasoning, we leverage a powerful low-level vision VLM. Given a degraded pair, we prompt the model to perform a step-by-step analysis, identifying each degradation, explaining its potential impact on fusion, and verbalizing a high-level strategy. (b) **Optimal action trajectory:** To determine the ground-truth sequence of tool calls, we employ an exhaustive search strategy. We explore all viable permutations of applicable tools and model combinations, scoring each complete trajectory using a set of relational quality metrics to identify the optimal path \mathcal{A} . A comprehensive description of this pipeline is provided in Appendix B.

3.3 FUSEAGENT FRAMEWORK

The core of FuseAgent is a VLM-based policy network, π_θ , which is trained to generate an optimal action trajectory. We cultivate its expert-level capabilities via a two-stage optimization process.

3.3.1 SUPERVISED FINE-TUNING (SFT)

Following recent agent development paradigms (Guo et al., 2025; Chen et al., 2024), we first initialize the policy via Supervised Fine-Tuning (SFT) on our generated dataset. This phase instills foundational skills in degradation identification and tool invocation by teaching the agent to mimic expert trajectories. The SFT objective is to maximize the likelihood of the expert’s response $\mathcal{R} = \{\mathcal{C}, \mathcal{A}\}$, optimized via a standard cross-entropy loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = - \sum \log \pi_\theta(\mathcal{R}|I_A, I_B). \quad (1)$$

3.3.2 REINFORCEMENT LEARNING FOR ROBUST POLICY MAKING

Building on the SFT-initialized policy, we introduce a reinforcement learning (RL) stage to enhance the agent’s robustness and generalization on unlabeled, in-the-wild data. We employ Group Relative Policy Optimization for Fusion (GRPO-F) (Shao et al., 2024; Jaech et al., 2024), a policy-gradient algorithm well-suited for this task. Refer to Appendix A.1 for further details on the implementation. The key to our approach is a novel, unsupervised reward signal that is action-dependent, distinguishing between intermediate pre-fusion steps and the final fusion action.

For any intermediate pre-fusion action ($a_t, t < n$), the reward R_t is designed to assess its contribution to improving the conditions for the final fusion. It is a composite signal:

$$R_t = w_i \cdot R_{\text{intrinsic}} + w_r \cdot R_{\text{relational}}.$$

where w_i and w_r are balancing weights.

Intrinsic quality reward. This reward, $R_{\text{intrinsic}} \in [0, 1]$, assesses the standalone perceptual quality of an action’s output image, I'_t . Its primary role is to encourage effective restoration and prevent the introduction of new artifacts. It is formulated as a weighted composite score from a suite of robust, no-reference IQA metrics:

$$R_{\text{intrinsic}}(I'_t) = \sum_{m \in M_{\text{IQA}}} w_m \cdot \hat{\Phi}_m(I'_t). \quad (2)$$

Here, M_{IQA} is the set of selected IQA metrics, w_m is a predefined weight for each metric m to balance its contribution, and $\hat{\Phi}_m$ is the normalized score produced by that metric.

Relational quality reward. The key of our reward design is $R_{\text{relational}}$, a metric that quantifies the improvement in “fusion compatibility” between the source images. It provides a critical guidance signal for all pre-fusion steps. As ground-truth is unavailable, it is defined as a multi-dimensional score evaluating the gain in inter-image consistency. We provide a detailed formulation in Appendix A.2.

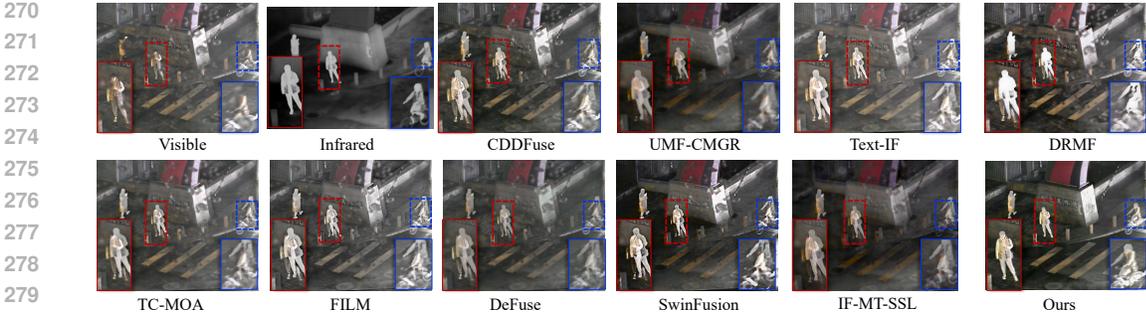
$$R_{\text{relational}} = \sum_{d \in D} w_d \cdot \mathcal{C}_d(I'_t, I_t, I_{\text{other}}). \quad (3)$$

where D is the set of predefined compatibility dimensions, w_d is the corresponding weight for each dimension d , and \mathcal{C}_d is the compatibility score function. The inputs to this function are the output image I'_t , the original image before the action I_t , and the other unprocessed image in the pair, I_{other} .

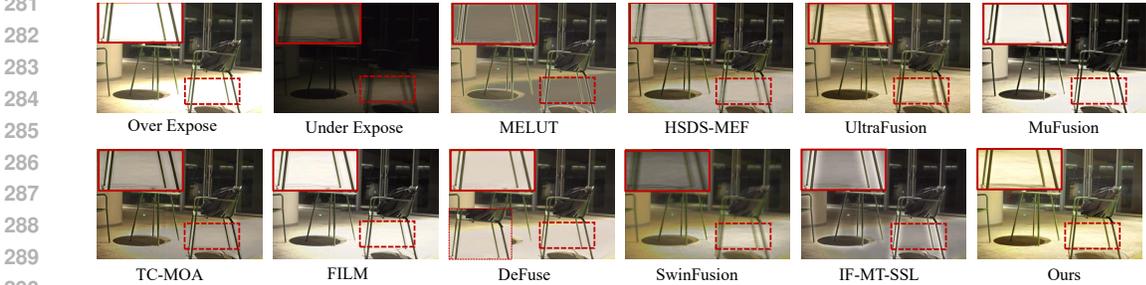
4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

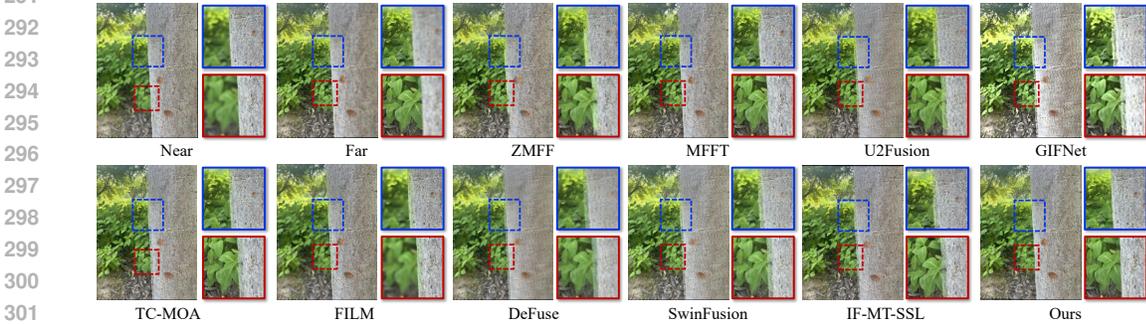
Dataset Setting. To validate performance under authentic in-the-wild conditions, we curated a comprehensive evaluation benchmark from a variety of public datasets. This benchmark was purposefully constructed to span a wide spectrum of real-world challenges, ranging from clean, well-controlled scenes to complex scenarios featuring multiple, coupled degradations. Our scenarios are



280 Figure 3: Visual comparisons of various methods in IVIF task.



291 Figure 4: Visual comparisons of various methods in MEF task.



303 Figure 5: Visual comparisons of various methods in MFF task.

304 organized around three key fusion tasks: (i) Infrared-Visible Fusion (IVIF). Scenarios are drawn
305 from sources like LLVIP (Jia et al., 2021) and anti-UAV (Zhu et al., 2023), selected to cover a
306 range of conditions from well-aligned inputs to challenging cases exhibiting inherent or simulated
307 misalignment, dynamic objects, and adverse weather. (ii) Multi-Exposure Fusion (MEF). Scenarios
308 are based on datasets like RealHDR-V (Shu et al., 2024) and DMEF (Tan et al., 2023), which fea-
309 ture prominent motion and ghosting artifacts. These challenges are often compounded with other
310 common degradations such as blur and poor visibility. (iii) Multi-Focus Fusion (MFF). Scenarios
311 are built upon datasets like RealMFF (Zhang et al., 2020) and EDMF (Li et al., 2024), presenting
312 a diversity of challenges from simple focus stacking to complex cases involving dynamic objects,
313 significant illumination variations, and adverse weather. The final dataset comprises 43,000 training
314 pairs (22k IVIF, 8k MEF, 13k MFF) and a test set of 3,200 pairs (1.5k IVIF, 0.8k MEF, 1k MFF).

315 **Implementation details.** We adopt Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) as the base model
316 for FusionAgent. The supervised fine-tuning is performed on our 43,000 collected training samples,
317 with a batch size of 2, a learning rate of 1e-5, and training for 2 epochs using the Llama-Factory
318 framework (Zheng et al., 2024) on 8 A100 (80G) GPUs. The reinforcement learning, employing
319 the GRPO-F algorithm, is conducted on a diverse subset of 5,000 challenging samples. For each
320 step, we sample a batch of 4, a learning rate of 1e-6, and generate 4 responses per query, training
321 for 2 epochs on 8 A100 (80G) GPUs. The detail hyper-parameter settings for SFT and GRPO-F are
322 provided in Appendix C.1.

323 **Evaluation metrics.** We deviate from traditional fusion metrics (e.g., Q_{abf} , Q_{cv} , VIF), whose
formulations are predicated on idealized assumptions of source image quality and alignment. Such
assumptions are frequently violated by real-world degradations, causing the metric scores to become

Table 1: Comparison of decision-making strategies on multi-degraded fusion tasks. We evaluate different processing strategies across IVIF, MEF and MFF task. The averaged metrics presented are based on the common ones. **Best** results are highlighted.

Strategy	IVIF Task					MEF Task				
	MS-SSIM \uparrow	HyperIQA \uparrow	BRISQUE \downarrow	MUSIQ \uparrow	CLIQQA \uparrow	MEF-SSIM \downarrow	HyperIQA \uparrow	BRISQUE \downarrow	MUSIQ \uparrow	CLIQQA \uparrow
Random Order & Model	0.2815	0.2522	48.3418	41.8833	0.2115	0.4533	0.3150	46.1824	43.1561	0.2898
Random Order & Predict Model	0.3582	0.3340	41.5227	48.2319	0.2640	0.5218	0.4011	40.5299	49.3782	0.3351
Predict Order & Random Model	0.4103	0.4281	33.1892	53.1524	0.3088	0.6304	0.5129	36.4183	56.8813	0.4529
Pre-defined Order & Model	0.4855	0.5891	19.5314	61.0581	0.3804	0.7782	0.6721	29.8371	68.1340	0.6215
Zero-shot VLM Planner	0.4629	0.5466	24.8816	58.7342	0.3571	0.7501	0.6503	31.7240	65.2199	0.5988
Human Expert	0.4988	0.6215	20.0253	63.5822	0.3819	0.7850	0.7018	27.9934	67.5028	0.6053
FuseAgent (Ours)	0.5064	0.6311	16.6587	64.3891	0.4122	0.8068	0.7105	27.2134	72.3164	0.6744

Strategy	MFF Task					Average			
	MS-SSIM \uparrow	HyperIQA \uparrow	BRISQUE \downarrow	MUSIQ \uparrow	CLIQQA \uparrow	HyperIQA \uparrow	BRISQUE \downarrow	MUSIQ \uparrow	CLIQQA \uparrow
Random Order & Model	0.5934	0.2811	49.5220	36.7291	0.2642	0.2828	48.0154	40.5895	0.2552
Random Order & Predict Model	0.6841	0.3205	44.8130	40.1824	0.3017	0.3519	42.2885	45.9308	0.3003
Predict Order & Random Model	0.7725	0.4593	38.2917	51.3888	0.3478	0.4668	35.9664	53.8075	0.3698
Pre-defined Order & Model	0.8870	0.6533	24.1821	64.9210	0.6401	0.6382	24.5169	64.7044	0.5473
Zero-shot VLM Planner	0.8653	0.6120	28.5294	62.0015	0.5905	0.6030	28.3783	61.9852	0.5155
Human Expert	0.8996	0.6686	22.3117	67.1305	0.6791	0.6640	23.4435	66.0718	0.5554
FuseAgent (Ours)	0.9051	0.6890	21.7826	68.0078	0.6870	0.6769	21.8849	68.2378	0.5912

decoupled from perceptual reality. Our quantitative assessment therefore first utilizes a suite of no-reference metrics robust to diverse degradations: HyperIQA (Su et al., 2020), BRISQUE (Mittal et al., 2012), MUSIQ (Ke et al., 2021) and CLIQQA (Wang et al., 2023) to assess overall perceptual quality and semantic fidelity. Moreover, we employ specialized metrics for task-specific structural fidelity: (i) For IVIF, we use Spatial Frequency (SF) to measure information richness and the multi-scale MS-SSIM (Wang et al., 2003) to preserve multi-scale structural similarity (ii) For MFF, the multi-scale MS-SSIM is applied. (iii) For MEF involving motion, we adopt the dynamic-aware MEF-SSIMd (Fang et al., 2020).

Tool settings. FusionAgent’s planned workflows are executed through a versatile suite of expert tools. Our implementation includes tools for registration (e.g., LoFTR (Sun et al., 2021)), optical flow estimation (e.g., RAFT (Teed & Deng, 2020)), image inpainting (e.g., LaMa (Suvorov et al., 2022)), illumination adjustment (e.g., IAT (Cui et al., 2022)), and image fusion (e.g., CDDFuse (Zhao et al., 2023)). Notably, we select representative and efficient models instead of the latest state-of-the-art models to simplify the validation process of our proposed paradigm. Incorporating more advanced models could further enhance performance. Refer to Appendix C.2 for a detailed description of the model settings.

4.2 DECISION MAKING CAPABILITY

Compared methods. To analyze the planning and decision-making capabilities of FusionAgent, we conducted a comparative study against a range of baseline strategies: (i) Random Order & Model, where both the workflow sequence and the expert model for each step are selected randomly. (ii) Random Order & Predicted Model, where the workflow order is random, but the model for each step is chosen by FusionAgent. (iii) Predicted Order & Random Model, where the workflow order is determined by FusionAgent, but the expert model for each step is chosen randomly. (iv) Pre-defined Order & Model, which represents a rigid, rule-based pipeline (e.g., restore \rightarrow register \rightarrow fuse) with a fixed set of models. (v) Zero-shot VLM Planner, where we use the base VLM (ChatGPT, Gemini) without any fine-tuning to generate the workflow. (vi) Human Expert, where an experienced specialist manually designs a bespoke workflow for each image. This baseline represents a strong, practical upper bound based on human intelligence.

Results. As shown in Table 1, we compare the decision-making capabilities of FuseAgent with baseline strategies. Three key observations emerge: 1) Random strategies exhibit significantly poor performance, highlighting the need for both strategic planning and proper model selection. 2) The “Predict Order & Random Model” strategy outperforms the “Random Order & Predict Model” strategy, suggesting that the optimal workflow sequence is more critical than the quality of individual tools. 3) Rule-based and expert-driven strategies, such as the Pre-defined pipeline (Avg. HyperIQA 0.6382) and the Human Expert (Avg. HyperIQA 0.6640), yield strong scores. However, our

Table 2: Quantitative comparison with state-of-the-art *task-specific methods* on multi-degraded fusion tasks. We evaluate across IVIF, MEF, and MFF benchmarks. We highlight the **best** and **second-best** results.

Method	Task type	MS-SSIM \uparrow	HyperIQA \uparrow	MUSIQ \uparrow	CLIQQA \uparrow	SF \uparrow
TarDAL	IVIF	0.4636	0.4237	51.3410	0.2067	7.3909
CDDFuse	IVIF	0.4786	0.4800	57.1356	0.2789	7.2951
C-MPDR	IVIF	0.4928	0.2812	37.5873	0.1251	4.9812
UMF-CMGR	IVIF	0.4953	0.2844	36.7453	0.1310	4.6752
Text-IF	IVIF	0.3834	0.4739	57.3989	0.2715	8.2082
DRMF	IVIF	0.4851	0.4432	52.7181	0.3018	6.8013
FuseAgent (Ours)	IVIF	0.5064	0.6311	64.3891	0.4122	8.4295
		MS-SSIM \uparrow	HyperIQA \uparrow	MUSIQ \uparrow	CLIQQA \uparrow	BRISQUE \downarrow
ZMFF	MFF	0.7942	0.3410	50.2858	0.4062	41.2423
MFFT	MFF	0.8192	0.4222	52.0305	0.3293	32.2150
MDLSR-RFM	MFF	0.8107	0.4997	59.5366	0.4113	27.3327
FuseAgent (Ours)	MFF	0.9051	0.6890	68.0078	0.6870	21.7826
		MEF-SSIM $_d\uparrow$	HyperIQA \uparrow	MUSIQ \uparrow	CLIQQA \uparrow	BRISQUE \downarrow
MEFLUT	MEF	0.6970	0.5823	68.4800	0.4411	23.8959
HSDS-MEF	MEF	0.6228	0.5364	64.7659	0.3916	25.3278
UltraFusion	MEF	0.7740	0.5910	68.9866	0.4303	28.8180
FuseAgent (Ours)	MEF	0.8068	0.7105	72.3164	0.6744	25.2134

FuseAgent achieves the highest performance (Avg. HyperIQA **0.6769**), surpassing all baseline approaches. In conclusion, FuseAgent, augmented with the SFT+GRPO paradigm, discovers more optimal and generalizable strategies than static pipelines or case-by-case human heuristics for complex, real-world fusion problems. Detailed case studies that visualize this complete reasoning and planning process can be found in Appendix D.

4.3 GENERAL FUSION ABILITY

Compared Methods. To demonstrate the effectiveness and superiority of FuseAgent, we conduct comprehensive comparisons against a wide range of state-of-the-art (SOTA) methods, which can be broadly divided into two categories. The first category, **task-specific models**, includes methods designed for a single fusion task or a particular degradation, such as TarDAL (Liu et al., 2022), CDDFuse (Zhao et al., 2023), C-MPDR (Wang et al., 2024a), UMF-CMGR (Wang et al., 2022), Text-IF (Yi et al., 2024), DRMF (Tang et al., 2024), MEFLUT (Jiang et al., 2023), HSDS-MEF (Wu et al., 2024), UltraFusion (Chen et al., 2025), MDLSR-RFM (Wang et al., 2024c), ZMFF (Hu et al., 2023), and MFFT (Zhai et al., 2024). The second category consists of **all-in-one methods**, which aim to handle multiple fusion tasks within a single framework. This group includes U2Fusion (Xu et al., 2020), DeFuse (Liang et al., 2022), SwinFusion (Ma et al., 2022), IF-MT-SSL (Wang et al., 2024d), GIFNet (Cheng et al., 2025), MUFusion (Cheng et al., 2023), TC-MOA (Zhu et al., 2024b), and FILM (Zhao et al., 2024). For all compared methods, we use their officially released codes and follow the recommended settings to ensure a fair and rigorous comparison.

Results. As shown in Tables 2 and 3, FuseAgent consistently outperforms both task-specific and all-in-one models across our challenging in-the-wild benchmarks. While competing methods struggle with compounded defects, FuseAgent establishes a new state-of-the-art. For instance, in the IVIF task, it achieves a **36.6%** improvement in CLIQQA score over the strongest specialized competitor, DRMF. Similarly, its performance in MEF (MEF-SSIM $_d$: **0.8068** vs. 0.7740) and MFF (HyperIQA: **0.6890** vs. 0.4946) surpasses the top competitors in each category. The visual comparisons in Figures 3-5 further confirm this superiority; unlike static methods that produce artifacts like ghosting and distortion, FuseAgent’s dynamic, perception-driven planning yields clean and structurally coherent results.

4.4 ABLATION STUDY

Training Strategy. To validate our two-stage training paradigm, we compare three configurations in Table 4 (rows 2–4). The SFT-only approach establishes a reasonable baseline but is insufficient for optimizing complex scenarios that deviate from the expert trajectories. Conversely, training with

Table 3: Quantitative comparison with state-of-the-art *all-in-one methods* on multi-degraded fusion tasks. We evaluate across IVIF, MEF, and MFF benchmarks. We highlight the **best** and **second-best** results.

Method	IVIF					MEF Task				MFF Task			
	SF \uparrow	MS-SSIM \uparrow	HyperQA \uparrow	MUSIQ \uparrow	CLIPQA \uparrow	MEF-SSIM \uparrow	HyperQA \uparrow	MUSIQ \uparrow	CLIPQA \uparrow	MS-SSIM \uparrow	HyperQA \uparrow	MUSIQ \uparrow	CLIPQA \uparrow
U2Fusion	6.9949	0.5054	0.3948	53.5210	0.2771	0.6768	0.4782	62.2200	0.4246	0.8560	0.4078	53.7902	0.4408
DeFuse	5.7597	0.4757	0.4720	54.7675	0.2388	0.7461	0.4893	59.7320	0.3972	0.8813	0.3981	49.9259	0.3668
SwinFusion	7.3620	0.4764	0.4049	51.8339	0.2646	0.6129	0.4170	55.5563	0.3208	0.8413	0.3614	52.6268	0.3283
IF-MT-SSL	6.2778	0.5042	0.4521	51.9412	0.2529	0.6863	0.5894	66.9227	0.4979	0.8372	0.4946	59.5262	0.4501
GIFNet	7.6257	0.4996	0.3956	55.1952	0.2760	0.5483	0.5312	63.6882	0.4414	0.7783	0.4414	58.3438	0.4049
MUFusion	7.3898	0.4372	0.2572	41.4161	0.1361	0.6852	0.3997	51.0261	0.3182	0.8170	0.4243	58.2549	0.4629
TC-MOA	6.3922	0.4753	0.4010	52.2921	0.2289	0.7041	0.5744	66.8931	0.4775	0.8289	0.4228	53.8276	0.3914
FILM	7.6726	0.4603	0.5136	59.5164	0.2956	0.7518	0.5763	67.3389	0.5084	0.8322	0.4935	57.1397	0.4210
FuseAgent (Ours)	8.4295	0.5064	0.6311	64.3891	0.4122	0.8068	0.7105	72.3164	0.6744	0.9051	0.6890	68.0078	0.6870

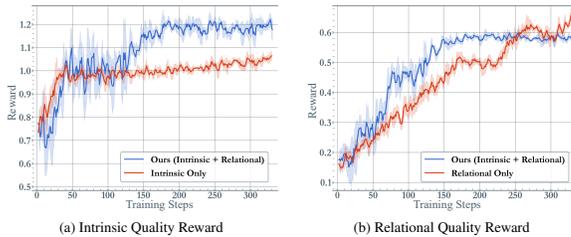


Figure 6: The reward trends across training steps.

GRPO-F from scratch improves perceptual metrics at the cost of structural fidelity. This trade-off indicates that while RL can explore better perceptual solutions, it struggles to acquire foundational fusion knowledge without proper initialization. The proposed SFT+GRPO-F pipeline attains superior performance across all metrics, confirming that SFT provides a crucial policy initialization, creating a robust foundation upon which GRPO-F effectively refines the policy for real-world complexities without sacrificing structural coherence.

Reward Design. The efficacy of our multi-dimensional reward is analyzed in Table 4 (rows 6–8) and Figure 6. Isolating the relational quality reward ($R_{\text{relational}}$) yields the highest structural fidelity, confirming its role in optimizing for geometric and structural consistency, but provides insufficient signal for improving standalone perceptual quality. Conversely, optimizing for only the intrinsic quality reward ($R_{\text{intrinsic}}$) leads to premature convergence; as illustrated in Figure 6(a), the policy rapidly improves but plateaus at a suboptimal level after approximately 50 training steps, indicating an inability to escape local optima. The full reward combination ($R_{\text{intrinsic}} + R_{\text{relational}}$) achieves the best overall performance, excelling in both perceptual quality and artifact reduction. This synergy validates our design: the intrinsic term ensures perceptual quality, while the relational term provides the necessary guidance to escape local optima and enforce inter-image compatibility, leading to solutions that are both visually pleasing and structurally coherent.

5 CONCLUSION

In this paper, we present FuseAgent, a VLM-powered agent system designed to address the challenges of in-the-wild image fusion. FuseAgent autonomously perceives complex, coupled degradations and dynamically orchestrates expert models into tailored workflows. The training process follows a two-stage optimization paradigm: 1) Supervised Fine-Tuning (SFT) on a curated dataset of expert trajectories instills foundational skills in degradation perception and tool usage. 2) Group Relative Policy Optimization for Fusion (GRPO-F), augmented with novel, unsupervised rewards—the Intrinsic Quality Reward (IQR) and Relational Quality Reward (RQR)—refines the agent’s decision-making and generalization abilities. Comprehensive experiments on challenging in-the-wild benchmarks demonstrate that FuseAgent significantly outperforms both specialized and all-in-one models in fusion quality and decision-making.

Table 4: Ablation studies on different training strategies and reward design.

Configurations	SF \uparrow	HyperQA \uparrow	BRISQUE \downarrow	CLIPQA \uparrow
Training strategy				
only SFT	8.10	0.45	25.43	0.39
only RL	7.79	0.55	25.88	0.44
SFT + RL (Ours)	8.63	0.68	21.88	0.59
Reward design				
only IQR	8.45	0.61	23.02	0.54
only RQR	8.81	0.52	24.15	0.4
IQR + RQR (Ours)	8.63	0.68	21.88	0.59

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. We confirm that all authors have read and agreed to abide by its principles. Our research does not involve human subjects or sensitive personal data, and no potentially harmful insights are presented. The datasets used are publicly available and appropriately cited. We have no known conflicts of interest, financial or otherwise, related to this work. All methodologies were designed with fairness and transparency in mind, and we have taken steps to ensure reproducibility and research integrity. No IRB approval was required for this study.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our research. In the main paper, we provide a detailed description of the model’s architecture, experimental setup, and evaluation methods. To provide deep insight into the model’s internal operational mechanisms, we have included an interactive visualization interface in the supplementary materials. This interface dynamically presents the model’s end-to-end inference process, from receiving the initial input to generating the final output, thereby revealing its complex decision-making logic.

REFERENCES

- Moab Arar, Yiftach Ginger, Dov Danon, Amit H. Bermano, and Daniel Cohen-Or. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025a.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.
- Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23555–23564, 2023.
- Zi-Han Cao, Yu-Jie Liang, Liang-Jian Deng, and Gemine Vivone. An efficient image fusion network exploiting unifying language and mask guidance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *arXiv preprint arXiv:2407.18035*, 2024.
- Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16111–16121, 2025.
- Chunyang Cheng, Tianyang Xu, and Xiao-Jun Wu. Mufusion: A general unsupervised image fusion network based on memory unit. *Information Fusion*, 92:80–92, 2023.
- Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 28102–28112, June 2025.
- Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, ZhengKai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: a light weight transformer for image

- 540 enhancement and exposure correction. In *33rd British Machine Vision Conference 2022, BMVC*
541 *2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL [https://bmvc2022.](https://bmvc2022.mpi-inf.mpg.de/0238.pdf)
542 [mpi-inf.mpg.de/0238.pdf](https://bmvc2022.mpi-inf.mpg.de/0238.pdf).
- 543
- 544 Yuming Fang, Hanwei Zhu, Kede Ma, Zhou Wang, and Shutao Li. Perceptual evaluation for multi-
545 exposure image fusion of dynamic scenes. *IEEE Transactions on Image Processing*, 29(1):1127–
546 1138, Dec. 2020.
- 547
- 548 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
549 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
550 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 551
- 552 Xingyu Hu, Junjun Jiang, Xianming Liu, and Jiayi Ma. Zmff: Zero-shot multi-focus image fusion.
553 *Information Fusion*, 92:127–138, 2023.
- 554
- 555 Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet:
556 Recurrent correction network for fast and efficient multi-modality image fusion. In *European*
557 *conference on computer Vision*, pp. 539–555. Springer, 2022.
- 558
- 559 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
560 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv*
561 *preprint arXiv:2412.16720*, 2024. URL <https://arxiv.org/abs/2412.16720>.
- 562
- 563 Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging*
564 *face library and models to solve problems*, pp. 51–67. Springer, 2022.
- 565
- 566 Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Lvip: A visible-infrared
567 paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on*
568 *Computer Vision*, pp. 3496–3504, 2021.
- 569
- 570 Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightdiffusion: Unsuper-
571 vised low-light image enhancement with latent-retinex diffusion models. In *European Conference*
572 *on Computer Vision*, 2024.
- 573
- 574 Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate
575 hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international*
576 *conference on computer vision*, pp. 9772–9781, 2021.
- 577
- 578 Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, and Shuaicheng Liu. Meflut: Un-
579 supervised 1d lookup tables for multi-exposure image fusion. In *Proceedings of the IEEE/CVF*
580 *International Conference on Computer Vision (ICCV)*, pp. 10542–10551, October 2023.
- 581
- 582 Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale im-
583 age quality transformer. In *Proceedings of the IEEE/CVF international conference on computer*
584 *vision*, pp. 5148–5157, 2021.
- 585
- 586 Hui Li, Tianyu Shen, Zeyang Zhang, Xuefeng Zhu, and Xiaoning Song. Edmf: A new benchmark
587 for multi-focus images with the challenge of exposure difference. *Sensors (Basel, Switzerland)*,
24(22):7287, 2024.
- 588
- 589 Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-
590 supervised decomposition approach for image fusion. In *European Conference on Computer*
591 *Vision*, pp. 719–735. Springer, 2022.
- 592
- 593 Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan
Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to
fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on*
Computer Vision and Pattern Recognition, pp. 5802–5811, 2022.
- David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh*
IEEE international conference on computer vision, volume 2, pp. 1150–1157. Ieee, 1999.

- 594 Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. Swinfusion: Cross-
595 domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of*
596 *Automatica Sinica*, 9(7):1200–1217, 2022.
- 597 Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assess-
598 ment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.
- 600 Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation
601 with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024.
- 602 Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model
603 connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.
- 604 Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru
605 Tang, Bill Qian, et al. Toollm: Facilitating large language models to master 16000+ real-world
606 apis. *arXiv preprint arXiv:2307.16789*, 2023.
- 607 Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-
608 time image enhancement. *Journal of VLSI signal processing systems for signal, image and video*
609 *technology*, 38(1):35–44, 2004.
- 610 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
611 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical
612 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 613 Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugging-
614 gpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information*
615 *Processing Systems*, 36, 2024.
- 616 Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world hdr video
617 reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *Proceed-*
618 *ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2879–2888,
619 2024.
- 620 Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly
621 assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the*
622 *IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.
- 623 Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local
624 feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer*
625 *vision and pattern recognition*, pp. 8922–8931, 2021.
- 626 Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha,
627 Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempit-
628 sky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint*
629 *arXiv:2109.07161*, 2021.
- 630 Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha,
631 Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky.
632 Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the*
633 *IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.
- 634 Xiao Tan, Huaian Chen, Rui Zhang, Qihan Wang, Yan Kan, Jinjin Zheng, Yi Jin, and Enhong Chen.
635 Deep multi-exposure image fusion for dynamic scenes. *IEEE Transactions on Image Processing*,
636 32:5310–5325, 2023.
- 637 Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. Drmf:
638 Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings*
639 *of the ACM International Conference on Multimedia*, pp. 8546–8555, 2024.
- 640 Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European*
641 *conference on computer vision*, pp. 402–419. Springer, 2020.

- 648 Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible
649 image fusion via cross-modality image generation and registration. In *IJCAI*, pp. 3508–3515,
650 2022.
- 651 Di Wang, Jinyuan Liu, Long Ma, Risheng Liu, and Xin Fan. Improving misaligned multi-modality
652 image fusion with one-stage progressive dense registration. *IEEE Transactions on Circuits and
653 Systems for Video Technology*, 34(11):10944–10958, 2024a.
- 654 Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel
655 of images. In *AAAI*, 2023.
- 656 Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploit-
657 ing diffusion prior for real-world image super-resolution. 2024b.
- 658 Jiwei Wang, Huaijing Qu, Zhisheng Zhang, and Ming Xie. New insights into multi-focus image
659 fusion: A fusion method based on multi-dictionary linear sparse representation and region fusion
660 model. *Information Fusion*, pp. 102230, 2024c. ISSN 1566-2535. doi: [https://doi.org/10.1016/
661 j.inffus.2024.102230](https://doi.org/10.1016/j.inffus.2024.102230). URL [https://www.sciencedirect.com/science/article/
662 pii/S1566253524000083](https://www.sciencedirect.com/science/article/pii/S1566253524000083).
- 663 Wu Wang, Liang-Jian Deng, and Gemine Vivone. A general image fusion framework using multi-
664 task semi-supervised learning. *Information Fusion*, pp. 102414, 2024d.
- 665 Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind
666 super-resolution with pure synthetic data. In *International Conference on Computer Vision Work-
667 shops (ICCVW)*.
- 668 Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality
669 assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*,
670 volume 2, pp. 1398–1402. Ieee, 2003.
- 671 Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Vi-
672 sual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint
673 arXiv:2303.04671*, 2023a.
- 674 Guanyao Wu, Hongming Fu, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Hybrid-supervised
675 dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In *Pro-
676 ceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5985–5993, 2024.
- 677 Ruiqi Wu, Zhengpeng Duan, Chunle Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real
678 image dehazing via high-quality codebook priors. In *Proceedings of the IEEE/CVF Conference
679 on Computer Vision and Pattern Recognition*, 2023b.
- 680 Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer.
681 *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- 682 Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised
683 image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- 684 Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mu-
685 tually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF
686 Conference on Computer Vision and Pattern Recognition*, pp. 19679–19688, 2022.
- 687 Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching
688 large language model to use tools via self-instruction. *Advances in Neural Information Processing
689 Systems*, 36, 2024.
- 690 Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text
691 guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF
692 Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024.
- 693 Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint
694 anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

- 702 Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-
703 Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*,
704 2022.
- 705 Hao Zhai, Wenyi Zheng, Yuncan Ouyang, Xin Pan, and Wanli Zhang. Multi-focus image fusion
706 via interactive transformer and asymmetric soft sharing. *Engineering Applications of Artificial*
707 *Intelligence*, 133:107967, 2024.
- 708
- 709 Juncheng Zhang, Qingmin Liao, Shaojun Liu, Haoyu Ma, Wenming Yang, and Jing-Hao Xue. Real-
710 mff: A large realistic multi-focus image dataset with ground truth. *Pattern Recognition Letters*,
711 138:370–377, 2020.
- 712 Kai Zhang, Yawei Li, Jingyun Liang, Jiezhong Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu
713 Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data syn-
714 thesis. *Machine Intelligence Research*, 20(6):822–836, 2023. doi: 10.1007/s11633-023-1466-0.
715 URL <https://doi.org/10.1007/s11633-023-1466-0>.
- 716
- 717 Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and
718 Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference*
719 *on Computer Vision*, pp. 187–204. Springer, 2025.
- 720 Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte,
721 and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-
722 modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
723 *Pattern Recognition (CVPR)*, pp. 5906–5916, June 2023.
- 724
- 725 Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin,
726 Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-
727 language model. In *Proceedings of the International Conference on Machine Learning (ICML)*,
728 2024.
- 729 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyao Luo, Zhangchi Feng, and
730 Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Pro-*
731 *ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*
732 *3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguis-
733 tics. URL <http://arxiv.org/abs/2403.13372>.
- 734 Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for
735 complex image restoration problems. *arXiv preprint arXiv:2410.17809*, 2024a.
- 736
- 737 Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for
738 general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*
739 *recognition*, pp. 7099–7108, 2024b.
- 740 Xue-Feng Zhu, Tianyang Xu, Jian Zhao, Jia-Wei Liu, Kai Wang, Gang Wang, Jianan Li, Zhihao
741 Zhang, Qiang Wang, Lei Jin, et al. Evidential detection and tracking collaboration: New problem,
742 benchmark and algorithm for robust anti-uav system. *arXiv preprint arXiv:2306.15767*, 2023.
- 743
- 744
- 745
- 746
- 747
- 748
- 749
- 750
- 751
- 752
- 753
- 754
- 755

APPENDIX

Our Appendices includes the following sections:

- Use of LLMs.
- Sec.A Additional Method Details.
 - Group Relative Policy Optimization.
 - Details of Reward Calculation.
- Sec.B Details of Dataset.
 - Stage I: Image Pair Generation and Degradation Synthesis.
 - Stage II: Expert Response Generation.
- Sec.C Additional Experiment Details.
 - Hyper-parameter Settings.
 - Tool settings.
 - Zero-Shot VLM Planner Baseline.
- Sec.D Case Studies of the Decision-Making Process.

USE OF LLMs

We used a large language model (LLM) solely for proofreading purposes, such as correcting grammatical errors and improving sentence clarity in the final draft. The LLM did not contribute to the research ideation, experimental design, theoretical analysis, or any substantive content of this work. All scientific content, claims, and results are the sole responsibility of us.

A ADDITIONAL METHOD DETAILS

A.1 GROUP RELATIVE POLICY OPTIMIZATION

Our reinforcement learning stage employs Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a policy-gradient algorithm that operates without a critic model. In GRPO, given an input image pair, the policy model π_θ generates a set of N potential action trajectories $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_N\}$. Each trajectory is executed, and a final reward is computed for each, resulting in a set of rewards $\{R_1, R_2, \dots, R_N\}$. Unlike PPO, which relies on a value function, GRPO directly compares the rewards within the group to determine the relative quality of each trajectory. The relative advantage A_i of the i -th trajectory is calculated by normalizing its reward against the statistics of the entire group:

$$A_i = \frac{R_i - \text{Mean}(\{R_1, R_2, \dots, R_N\})}{\text{Std}(\{R_1, R_2, \dots, R_N\}) + \epsilon},$$

where Mean and Std represent the mean and standard deviation of the rewards, and ϵ is a small constant for numerical stability. This normalization allows GRPO to capture nuanced differences between candidate trajectories. The policy update is constrained by minimizing the KL divergence between the current and reference models, ensuring stable learning. For more details, we refer the reader to (Guo et al., 2025; Jaech et al., 2024).

A.2 DETAILS OF REWARD CALCULATION

This section provides the detailed formulations for the two primary reward components introduced in Sec. 3.3.2: the Intrinsic Quality Reward and the Relational Quality Reward.

- **Intrinsic Quality Reward** ($R_{\text{intrinsic}}$). This reward assesses the standalone perceptual quality of a single image I'_i resulting from a unilateral action (e.g., restoration, initial enhancement). It encourages artifact removal and visual plausibility. The reward is a weighted

810 composite score from a suite of no-reference Image Quality Assessment (IQA) metrics,
811 each chosen to capture different aspects of quality:

$$812 R_{\text{intrinsic}}(I'_t) = \sum_{m \in M_{\text{IQA}}} w_m \cdot \hat{\Phi}_m(I'_t),$$

813 where $M_{\text{IQA}} = \{\text{HyperIQA}, \text{BRISQUE}, \text{MUSIQ}, \text{CLIQQA}\}$. For instance, BRISQUE is
814 sensitive to common compression and sensor artifacts, MUSIQ assesses overall aesthetic
815 quality, and CLIQQA provides a semantic evaluation of fidelity.

- 816 • **Relational Quality Reward ($R_{\text{relational}}$)**. This reward is the core of our RQA paradigm and
817 evaluates any action that modifies or depends on the relationship between the two images
818 in a pair. It is a multi-dimensional assessment whose specific instantiation depends on the
819 action’s goal.

820 **Enhancement Compatibility.** For actions that enhance one image (e.g., visible light en-
821 hancement producing I'_v) to better match another (e.g., an infrared image I_i), the reward
822 measures the improvement in fusion compatibility. It is a weighted sum over several di-
823 mensions:

$$824 R_{\text{enhance}}(I'_v, I_v, I_i) = \sum_{d \in D_{\text{enhance}}} w_d \cdot \mathcal{C}_d(I'_v, I_v, I_i),$$

825 where key dimensions $d \in D_{\text{enhance}}$ include *structural compatibility* (via SSIM), *gradient*
826 *compatibility* (via gradient correlation), and *contrast compatibility* (via histogram distance),
827 ensuring the enhancement prepares the image for a more effective final fusion.

828 **Alignment and Deghosting Accuracy.** For actions that address geometric or temporal
829 misalignments (e.g., registration, deghosting via optical flow and masking), the reward
830 assesses the resulting consistency. As ground-truth transformations are unavailable, the
831 reward is a composite score evaluating alignment from multiple perspectives:

$$832 R_{\text{align}}(I'_{\text{warped}}, I_{\text{ref}}) = \sum_{d \in D_{\text{align}}} w_d \cdot \mathcal{A}_d(I'_{\text{warped}}, I_{\text{ref}}),$$

833 where dimensions $d \in D_{\text{align}}$ include *structural consistency* (measured by NCC), *feature*
834 *similarity* (cosine similarity of deep features from a pre-trained VGG network), and *geo-*
835 *metric precision* (e.g., penalizing unrealistic distortions).

836 **Overall Fusion Quality.** For the final fusion action which produces the output I_f , this re-
837 ward serves as the comprehensive evaluation of the entire workflow. It holistically measures
838 how well I_f integrates information from both sources, I_v and I_i . It combines refer-
839 ence-based metrics that assess structural and informational fidelity with no-reference metrics for
840 perceptual quality:

$$841 R_{\text{fusion}}(I_f, I_v, I_i) = \sum_{m \in M_{\text{fusion}}} w_m \cdot \Phi_m(I_f, I_v, I_i),$$

842 where M_{fusion} includes information-theoretic metrics (SF, AG), structural metrics (MS-
843 SSIM), and the full suite of IQA metrics.

853 B DETAILS OF DATASET

854 This section provides a detailed account of the three-stage pipeline used to construct the training and
855 evaluation dataset for FuseAgent, as introduced in Sec. 3.2. Our goal was to create a large-scale,
856 diverse dataset of complex fusion scenarios paired with high-quality expert reasoning (Chain-of-
857 Thought, CoT) and optimal action trajectories.

860 B.1 STAGE I: IMAGE PAIR GENERATION AND DEGRADATION SYNTHESIS

861 Our data generation process employs a hybrid strategy, combining authentic degraded image pairs
862 from public benchmarks with systematically synthesized data to ensure both realism and compre-
863 hensive coverage of challenges.

Image Collection. We source our data from a variety of public datasets covering our three main fusion tasks (IVIF, MEF, MFF), including real-world captures that already exhibit a single, specific degradation (e.g., misalignment from anti-UAV, dynamic artifacts from RealHDR-V). These serve as the foundation for our benchmark.

Degradation Synthesis. To create complex, compounded challenges that are underrepresented in existing datasets, we developed a modular degradation library. This library is applied to both clean source images and real degraded images to layer additional defects. Key simulated degradations include:

- **Geometric Misalignment:** We apply random homography and affine transformations to one image in a pair to simulate camera shake and parallax errors.
- **Dynamic Artifacts:** To simulate motion, we first segment a chosen object, inpaint the background using a high-fidelity inpainting model Suvorov et al. (2022), and then re-insert the object with a slight transformation in one of the images.
- **Photometric Inconsistencies:** We model exposure and white balance variations by applying randomized color and brightness adjustments based on physical camera response functions.
- **Adverse Weather & Sensor Noise:** We utilize established physical models and generative techniques to synthesize realistic weather effects (e.g., rain, haze) and sensor noise patterns.

This hybrid and modular approach allows us to generate a rich dataset spanning a wide spectrum of complexity, from simple single-defect cases to challenging scenarios with multiple, coupled degradations.

B.2 STAGE II: EXPERT RESPONSE GENERATION

For each degraded image pair, we generate a corresponding expert response, which consists of the optimal action trajectory and the underlying CoT rationale.

Optimal Action Trajectory. To establish a ground-truth sequence of tool-calls (\mathcal{A}), we employ an exhaustive search strategy. For each scenario, we define a set of applicable pre-processing and fusion tools from our library. We then explore all valid permutations of tool sequences, executing each full trajectory. Each final fused image is scored using our Holistic Quality Assessment (HQA) reward function (detailed in Appendix A.2). The trajectory that yields the highest HQA score is designated as the ground-truth optimal path.

Chain-of-Thought (CoT) Rationale. To generate a human-like reasoning process (\mathcal{C}) that explains the logic behind the optimal trajectory, we "reverse-engineer" the thought process using a powerful VLM. We provide the VLM with the degraded source images and the pre-determined optimal action trajectory. The VLM is then prompted to produce a concise, first-person rationale explaining *why* this sequence of actions is necessary and logical. This process ensures that the CoT is not only coherent but also perfectly aligned with the ground-truth actions. The detailed prompt used for this generation process is presented in Table 8.

C ADDITIONAL EXPERIMENT DETAILS

C.1 HYPER-PARAMETER SETTINGS

In Table 5, we detail the hyper-parameter settings for our experiments.

C.2 TOOL SETTINGS

Table 6 lists the comprehensive suite of tools integrated into our framework. For registration and alignment, we utilize both a detector-free model, LoFTR Sun et al. (2021), and the classic SIFT Lowe (1999) algorithm as a robust baseline. Motion analysis is handled by GMA Jiang et al. (2021) and the iterative RAFT Teed & Deng (2020) for accurate optical flow estimation. For inpainting and filling masked regions, we employ LaMa Suvorov et al. (2021) and Inpaint-Anything Yu et al. (2023) for high-resolution inpainting.

Table 5: Hyper-parameter settings for SFT and GRPO-F

Hyper-parameter	SFT	GRPO-F
Batch size	2	4
Learning rate	1e-5	1e-6
Weight decay	0	0
Optimizer	AdamW	AdamW
Warmup ratio	0.1	0.1
LR scheduler	cosine	cosine
Training samples	43K	5K
Training epochs	2	2
Precision	bfloat16	bfloat16
KL coefficient	-	0.1
Reward coefficients	-	<i>IQR:1, RQR:1</i>
Number of generations	-	4
GPU resources	8×A100 (~384 GPU hours)	8×A100 (~2076 GPU hours)

Our denoising module consists of the hybrid SCUnet Zhang et al. (2023) and the efficient transformer-based Restormer Zamir et al. (2022). To address adverse weather conditions, we use IDT Xiao et al. (2022) for deraining and RIDCP Wu et al. (2023b) for dehazing. A versatile set of tools is available for general restoration tasks like deblurring and super-resolution, including the diffusion-based StableSR-turbo Wang et al. (2024b) and the fast GAN-based Real-ESRGAN Wang et al..

For challenges in illumination and exposure, our framework is equipped with a wide array of methods: a fast low-light enhancer in Img2img-turbo-night Parmar et al. (2024), a lightweight transformer IAT Cui et al. (2022), a diffusion-based model LightenDiff Jiang et al. (2024), and two classical histogram-based techniques, Histogram Matching and CLAHE Reza (2004).

Finally, the critical task of merging processed outputs is handled by a diverse set of fusion operators. In addition to the versatile SwinFusion Ma et al. (2022), our suite includes models for various scenarios: Text-IF Yi et al. (2024) for text-guided interactive fusion, CDDFuse Zhao et al. (2023) for multi-modal tasks, MEFLUT Jiang et al. (2023) for efficient multi-exposure fusion, as well as MFFT Zhai et al. (2024) and the zero-shot ZMFF Hu et al. (2023) for multi-focus fusion.

Notably, some models lack weights corresponding to certain tasks but are inherently adaptable; we collect appropriate data to retrain them. It is also important to note that we are not necessarily utilizing the latest state-of-the-art tools, suggesting considerable potential for future enhancements to our models.

C.3 ZERO-SHOT VLM PLANNER BASELINE

To evaluate the contribution of our two-stage training paradigm (SFT+GRPO-F), we established a strong baseline using a zero-shot VLM planner. This baseline utilizes the base VLM (e.g., Qwen2.5-VL-7B-Instruct, GPT 4V) without any of the domain-specific fine-tuning described in our main paper.

The agent’s behavior in this setting is guided solely by a comprehensive textual prompt that instructs it to act as an image fusion expert. This prompt, detailed in Table 8, provides the model with the high-level task objective, the required output format (including the use of `<think>` and `<answer>` tags), and the complete library of available expert tools and their functions. The VLM is then tasked with generating a full action trajectory based on its general-purpose, pre-existing reasoning capabilities, without any in-domain examples or specialized training. This allows us to fairly measure the performance gains achieved through our targeted SFT and GRPO-F stages.

Table 6: The expert tool suite available to FuseAgent, covering a wide range of pre-processing and fusion operations.

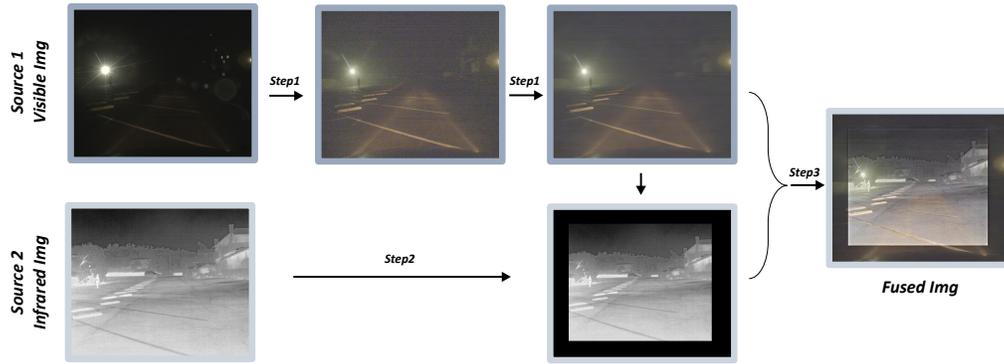
Task	Tools	Model Description
Registration & Alignment	LoFTR	Detector-free local feature matching model that excels at finding correspondences in challenging conditions.
	SIFT	Classic Scale-Invariant Feature Transform algorithm, serving as a robust baseline for keypoint matching.
Optical Flow Estimation	GMA	A global motion aggregation-based network for accurate optical flow estimation, crucial for motion analysis in dynamic scenes.
	RAFT	An iterative deep learning model for optical flow that uses a recurrent GRU-based operator to refine predictions.
Inpainting & De-occlusion	LaMa	High-resolution inpainting model using fast Fourier convolutions, effective at filling large masked regions after dehosting.
	Inpaint-Anything	A mask-free inpainting framework based on SAM that enables users to remove, fill, or replace objects via simple clicks.
Denoising	SCUNet	Hybrid UNet-based model combining convolution and transformer blocks for robust real-world denoising.
	Restormer	Efficient Transformer-based model for high-quality image denoising and restoration.
Deraining & Dehazing	IDT	Transformer-based model for unified de-raining and raindrop removal.
	RIDCP	Efficient dehazing model utilizing high-quality codebook priors for complex real-world haze.
Deblurring & Super-resolution	StableSR-turbo	Utilizes pre-trained diffusion models for high-quality super-resolution, deblurring, and artifact removal.
	Real-ESRGAN	Fast GAN-based model for super-resolution and deblurring, handling complex real-world degradations efficiently.
Illumination & Exposure Correction	Img2img-turbo-night	Fast and efficient model based on SD-turbo, designed for low-light enhancement.
	ITA	Lightweight transformer for efficient low-light and exposure correction.
	LightenDiff	Diffusion-based framework for unsupervised low-light enhancement leveraging Retinex theory.
	Histogram Matching	Classical technique that modifies an image’s histogram to match a reference, used for standardizing exposure.
	CLAHE	Contrast Limited Adaptive Histogram Equalization, a robust method for enhancing local contrast in images.
General-Purpose Fusion	SwinFusion	A Swin Transformer-based model serving as a powerful and versatile final fusion operator.
	Text-IF	A novel fusion model that leverages semantic text guidance for degradation-aware and interactive fusion tasks.
	CDDFuse	A Correlation-Driven feature Decomposition Fusion network using a dual-branch Transformer-CNN to model cross-modality features.
	MEFLUT	An efficient Multi-Exposure Fusion method that encodes fusion weights into a 1D lookup table (LUT) using attention.
	MFFT	A Multi-Focus Fusion method using an interactive transformer and asymmetric soft sharing to produce all-in-focus images.
	ZMFF	A zero-shot, untrained Multi-Focus Fusion framework that models the deep prior of the image and focus map.

D CASE STUDIES OF THE DECISION-MAKING PROCESS

This section provides two detailed case studies that visualize the complete reasoning and planning workflow of FuseAgent when confronted with complex, in-the-wild fusion scenarios. Each case study (Figure 7 and 8) is structured to offer a transparent view of the agent’s process, from perception to execution. The visual workflow at the top of each figure illustrates the step-by-step application of the planned action trajectory, showing the intermediate images generated at each stage. Below the visuals, the agent’s corresponding textual output reveals its Chain-of-Thought (CoT) reasoning within `<think>` tags, followed by the final, executable plan in `<answer>` tags.

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051

Action Details



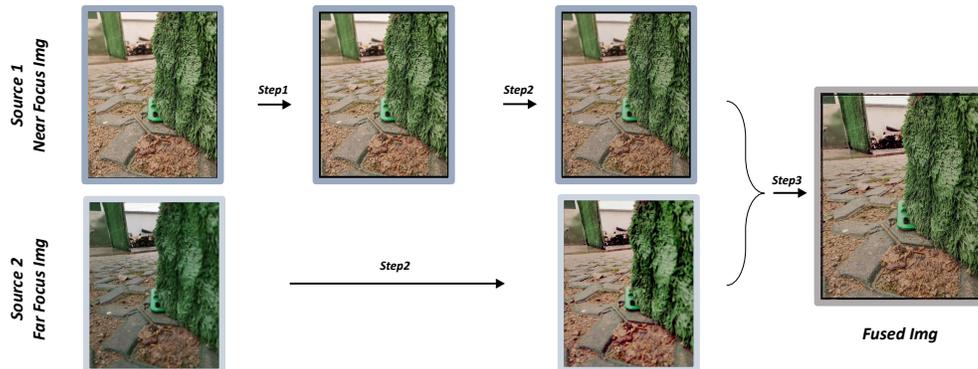
FuseAgent:

<think> This scene, depicting a person in a dark street with a bright streetlight, presents a challenging image fusion task. My initial analysis reveals several issues. First, the visible light image is severely degraded by *low light, noise, and glare*, which obscures key details and makes it difficult to extract reliable feature points. This prevents a straightforward registration process. Therefore, I must first perform restoration to enhance the image quality and prepare it for alignment. </think>
<answer> [task:deflare] [model:flare7k] → [task:lighten] [model:hvicident] → [task:denoise] [model:scunet] </answer>
<think> Given the *lack of robust features* in both the visible and infrared images, a traditional feature-based registration approach would likely fail. Instead, I'll need to use a more powerful model, such as a deep learning-based method, to accurately predict the dense deformation field. </think>
<answer> [task:registration] [model:lofir] </answer>
<think> Finally, while the visible light image is restored, it *remains an under-exposed image*. This requires the use of a specialized fusion model that is optimized for low-light conditions to effectively combine the textural information from the visible spectrum with the thermal properties captured by the infrared sensor. </think>
<answer> [[task:fusion] [model:textif] </answer>

Figure 7: FuseAgent’s Strategy for a Complex In-the-Wild Scenario: Tackling Compounded Degrations in Nighttime Fusion.

1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077

Action Details



FuseAgent:

<think> The user wants to perform multi-focus image fusion. I've been given two images of a moss-covered object on a tiled ground, likely taken at different focus points. I see a *slight misalignment* between the two images, which is probably due to camera shake during capture. The first step is to perform pixel-level alignment by adjusting the angle of one of the images. </think>
<answer> [task:registration] [model:sift] </answer>
<think> Next, I'll address the *differences in exposure and white balance* caused by the camera's auto-adjustment. To prevent ghosting artifacts in the final fused image, I'll need to adjust the brightness and color balance of both images to ensure they are consistent. </think>
<answer> [task:expose correction] [model:Histogram Matching] </answer>
<think> After these preparatory steps, the images will *meet the necessary conditions for a clean fusion*. I can then proceed with the fusion process to create a single, all-in-focus image. </think>
<answer> [task:fusion] [model:MFF7] </answer>

Figure 8: FuseAgent’s Strategy for a Complex In-the-Wild Scenario: Tackling Geometric and Photometric Inconsistencies in Multi-Focus Fusion.

Table 7: Prompt for Producing Refined CoT Annotations

Prompt for Producing Refined CoT Annotations	
As an image processing specialist, you're tasked with reverse-engineering the thought process behind an \langle Fusion Type \rangle . Given the source images, the pre-fusion processing pipeline, and the fusion algorithm, deduce the rationale for the entire workflow.	
Requirements:	<ol style="list-style-type: none"> 1. Your response should be a concise, single-paragraph analysis, written in English, limited to 400 characters. 2. Please omit any mention of specific method names, as they are part of the internal thought process and not required in the final response. 3. To start the analysis, provide a brief introduction to the pre-fusion image. For example, both images to be fused show a girl on the grass, playing with her dog. 4. For each processing operation, the introduction should emphasize why it is necessary. Present the rationale before describing how it is performed. 5. The final fused image should not be mentioned; it serves only as a reference to inform our speculation about the expert's original thinking. 6. The final output should sound like natural spoken English, not written text, and be from the original expert's perspective (first-person perspective).
Notes:	<ul style="list-style-type: none"> • The images, in order, are a visible light shot, an infrared shot, and the fused result. • Common pre-fusion image processing methods: <ol style="list-style-type: none"> 1. Registration & Alignment: (Method name: characteristics) <ul style="list-style-type: none"> - LoFTR: Detector-free local feature matching model that excels at finding correspondences in challenging conditions. - SIFT: Classic Scale-Invariant Feature Transform algorithm, serving as a robust baseline for keypoint matching. - ... 2. Optical Flow Estimation: (Method name: characteristics) <ul style="list-style-type: none"> - GMA: A global motion aggregation-based network for accurate optical flow estimation, crucial for motion analysis in dynamic scenes. - RAFT: An iterative deep learning model for optical flow that uses a recurrent GRU-based operator to refine predictions. - ... 3. Inpainting & De-occlusion: (Method name: characteristics) <ul style="list-style-type: none"> - LaMa: High-resolution inpainting model using fast Fourier convolutions, effective at filling large masked regions after deghosting. - Inpaint-Anything: A mask-free inpainting framework based on SAM that enables users to remove, fill, or replace objects via simple clicks. - ... 4. Denoising: (Method name: characteristics) <ul style="list-style-type: none"> - SCUnet: Hybrid UNet-based model combining convolution and transformer blocks for robust real-world denoising. - Restormer: Efficient Transformer-based model for high-quality image denoising and restoration. - ... 5. Deraining & Dehazing: (Method name: characteristics) <ul style="list-style-type: none"> - IDT: Transformer-based model for unified de-raining and raindrop removal. - RIDCP: Efficient dehazing model utilizing high-quality codebook priors for complex real-world haze. - ... 6. Deblurring & Super-resolution: (Method name: characteristics) <ul style="list-style-type: none"> - StableSR-turbo: Utilizes pre-trained diffusion models for high-quality super-resolution, deblurring, and artifact removal. - Real-ESRGAN: Fast GAN-based model for super-resolution and deblurring, handling complex real-world degradations efficiently. - ... 7. Illumination & Exposure Correction: (Method name: characteristics) <ul style="list-style-type: none"> - Img2img-turbo-night: Fast and efficient model based on SD-turbo, designed for low-light enhancement. - ITA: Lightweight transformer for efficient low-light and exposure correction. - LightenDiff: Diffusion-based framework for unsupervised low-light enhancement leveraging Retinex theory. - Histogram Matching: Classical technique that modifies an image's histogram to match a reference, used for standardizing exposure. - CLAHE: Contrast Limited Adaptive Histogram Equalization, a robust method for enhancing local contrast in images. - ... 8. General-Purpose Fusion: (Method name: characteristics) <ul style="list-style-type: none"> - SwinFusion: A Swin Transformer-based model serving as a powerful and versatile final fusion operator. - Text-IF: A novel fusion model that leverages semantic text guidance for degradation-aware and interactive fusion tasks. - CDDFuse: A Correlation-Driven feature Decomposition Fusion network using a dual-branch Transformer-CNN to model cross-modality features. - MEFLUT: An efficient Multi-Exposure Fusion method that encodes fusion weights into a 1D lookup table (LUT) using attention. - MFFT: A Multi-Focus Fusion method using an interactive transformer and asymmetric soft sharing to produce all-in-focus images. - ZMFF: A zero-shot, untrained Multi-Focus Fusion framework that models the deep prior of the image and focus map. - ...
	... The pre-fusion processing on the first image: \langle Method \rangle ; The pre-fusion processing on the second image: \langle Method \rangle ; Method for fusion: \langle Method \rangle ;

Table 8: The prompt used for the Zero-shot VLM Planner baseline.

Prompt for Zero-Shot VLM Planner	
1134	
1135	
1136	
1137	
1138	
1139	
1140	
1141	
1142	
1143	Table 8: The prompt used for the Zero-shot VLM Planner baseline.
1144	
1145	Prompt for Zero-Shot VLM Planner
1146	
1147	
1148	As an expert in image fusion and processing, your task is to analyze a given pair of source images for an <code><Fusion Type></code> task. You must identify any degradations or inconsistencies and then generate an optimal, step-by-step workflow to produce a high-quality fused image using the provided tool library.
1149	Requirements:
1150	1. Your response must consist of two parts: a reasoning process and a final action plan.
1151	2. First, provide your step-by-step analysis and reasoning within <code><think></code> tags. Explain the problems you see and why your proposed plan is the correct approach.
1152	3. Second, provide the final, executable workflow as a sequence of tool-calls within <code><answer></code> tags. Each tool-call should be in the format <code>[task:task_name][model:model_name]</code> .
1153	4. You must only select tools and models from the Tool Library provided below. Do not invent new ones.
1154	5. The order of operations is critical. Your plan should address necessary pre-processing steps (like registration or restoration) before the final fusion.
1155	
1156	Tool Library:
1157	• You will be provided with two source images for a specific fusion task.
1158	• The available tools and their functions are listed below:
1159	1. Registration & Alignment: (Tool name: function)
1160	– LoFTR: Detector-free local feature matching model that excels at finding correspondences in challenging conditions.
1161	– SIFT: Classic Scale-Invariant Feature Transform algorithm, serving as a robust baseline for keypoint matching.
1162	– ...
1163	2. Optical Flow Estimation: (Tool name: function)
1164	– GMA: A global motion aggregation-based network for accurate optical flow estimation, crucial for motion analysis in dynamic scenes.
1165	– RAFT: An iterative deep learning model for optical flow that uses a recurrent GRU-based operator to refine predictions.
1166	– ...
1167	3. Denoising: (Tool name: function)
1168	– SCUnet: Hybrid UNet-based model combining convolution and transformer blocks for robust real-world denoising.
1169	– Restormer: Efficient Transformer-based model for high-quality image denoising and restoration.
1170	– ...
1171	4. Deraining & Dehazing: (Tool name: function)
1172	– IDT: Transformer-based model for unified de-raining and raindrop removal.
1173	– RIDCP: Efficient dehazing model utilizing high-quality codebook priors for complex real-world haze.
1174	– ...
1175	5. Deblurring Super-resolution: (Tool name: function)
1176	– StableSR-turbo: Utilizes pre-trained diffusion models for high-quality super-resolution, deblurring, and artifact removal.
1177	– Real-ESRGAN: Fast GAN-based model for super-resolution and deblurring, handling complex real-world degradations efficiently.
1178	– ...
1179	6. Illumination Exposure Correction: (Tool name: function)
1180	– Img2img-turbo-night: Fast and efficient model based on SD-turbo, designed for low-light enhancement.
1181	– CLAHE: Contrast Limited Adaptive Histogram Equalization, a robust method for enhancing local contrast in images.
1182	– ...
1183	...
1184	7. General-Purpose Fusion: (Tool name: function)
1185	– SwinFusion: A Swin Transformer-based model serving as a powerful and versatile final fusion operator.
1186	– Text-IF: A novel fusion model that leverages semantic text guidance for degradation-aware and interactive fusion tasks.
1187	– CDDFuse: A Correlation-Driven feature Decomposition Fusion network using a dual-branch Transformer-CNN.
	– ...