ASSESSING INFRARED-VISIBLE FUSION TOWARDS MACHINE PERCEPTION AND UNDERSTANDING PREFERENCES: BENCHMARK, DATASET, AND BASELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite substantial progress in infrared-visible image fusion (IVIF) over the past decades, fused-image evaluation still relies heavily on metrics adapted from generic image quality assessment, which primarily cater to human perceptual preference. Machine-centric evaluation, by contrast, remains underexplored due to the lack of dedicated benchmarks. To address this limitation, we formalize machine preference as the utility of fused images for representative downstream perception and understanding tasks. We then introduce FuM-Score, which jointly quantifies such preference by aggregating fusion gains relative to the infrared and visible sources across these tasks under standardized evaluation protocols. Based on these, we subsequently establish FuM-18K, to our knowledge, the first large-scale dataset for machine-centric assessment in IVIF, comprising over 18,000 images spanning diverse scenes and representative fusion models, along with fine-grained downstream task annotations. Finally, an effective FuM-Net is built to leverage fusion features together with auxiliary source references, achieving automatic FuM-Score prediction for quality assessment. Extensive experiments validate the efficacy of our benchmark in capturing machine preference and demonstrate the superiority of FuM-Net over recent state-of-the-art methods.

1 INTRODUCTION

Complementary to the RGB modality, thermal cues from infrared sensors offer promising robustness against constrained conditions, such as low illumination, adverse weather, and occlusions. By integrating these advantages, infrared-visible image fusion (IVIF) synthesizes a single informative fused image that provides more complete representations to mitigate the limitations of single-modality sensing. Benefiting from this, IVIF has become indispensable for various downstream applications, including nighttime surveillance, autonomous driving, and robotic perception (Ma et al., 2019; Zhang et al., 2021a; Liu et al., 2024a). Many methods in the literature prioritize fusion for visual aesthetics, pursuing to improve the visual appearance and contrast for human observation. However, its inherently ill-posed nature, compounded by the absence of ground-truth (GT) labels, poses significant challenges for reliable fusion image quality assessment (IQA).

To bypass this, existing evaluation protocols predominantly rely on handcrafted, no-reference statistical metrics, such as entropy (EN) (Roberts et al., 2008), mutual information (MI) (Qu et al., 2002), and gradient-based measurements (Su et al., 2026; Xydeas et al., 2000), or repurpose existing IQA metrics from other low-level vision tasks (Ledig et al., 2017; Lim et al., 2017; Zhang et al., 2021b) by treating the fusion result as a pseudo-label and computing structural similarity between the fused image and each source modality. To overcome this problem, recent efforts have increasingly focused on either constructing dedicated datasets (Zhang et al., 2020; Chang et al., 2023) for IVIF quality assessment or developing metrics (Cheng et al., 2026; Liu et al., 2026; Guo et al., 2026) explicitly grounded in the human visual system (HVS) to provide more reliable evaluation. Excellent visual quality under HVS criteria does not necessarily imply superior performance in downstream machine perception tasks (Li et al., 2025; Wang et al., 2025b), as the latter often favors discriminative semantic features (See Figure 1). Accordingly, task-adaptive IVIF methods (Liu et al., 2022; Zhang et al., 2024; Bai et al., 2025a; Wang et al., 2025a; Zhang et al., 2026) have gained increasing attention, which incorporate task-related features or targeted regularization to enhance the downstream utility

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

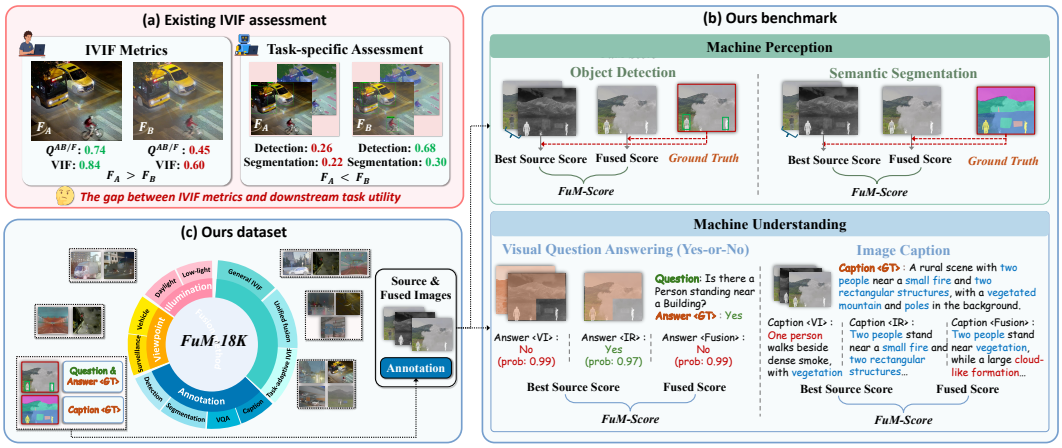

Figure 1: Motivation. (a) The big gap between conventional IVIF metrics and downstream task utility in existing IVIF assessment; (b) Our benchmark considers both machine perception and understanding, using task utility and source performance to assess fusion quality without an ideal fused reference image. (c) FuM-18K provides over 18,000 fused images with diverse illumination conditions, viewpoints, and fusion methods, annotated for four downstream tasks.

of fused images. Despite these advances, these methods suffer from severe fragmentation, relying on heterogeneous datasets, inconsistent annotations, and disparate task backbones. There is no standardized benchmark for rigorously assessing fusion efficacy aligned with machine preferences. While RA-MIQA (Wang et al., 2025b) and MPD (Li et al., 2025) introduce IQA for machine vision (MIQA), they primarily evaluate consistency between distorted and reference images, which are not applicable to IVIF since no reference labels are available.

Based on the above analysis, this work frames IVIF quality assessment from a machine-centric perspective on fusion quality, explicitly formalizing machine preferences as the practical utility of fused images across representative downstream perception and understanding tasks. To this end, we propose FuM-Score to quantify this preference through task-performance gains relative to the infrared and visible source images, aggregating these gains under a standardized evaluation protocol. Even though, establishing such a reliable criteria is non-trivial, which necessitates a comprehensive dataset that pairs diverse source–fusion samples with downstream task outputs obtained under a unified evaluation protocol, with labels capturing the performance gains or potential degradations of fused images relative to both source modalities. Consequently, we then build FuM-18K, a large-scale dataset tailored for machine-centric IVIF assessment, featuring diverse scenes, multiple representative fusion models, and fine-grained downstream annotations to support rigorous analysis and fair comparison. Finally, we develop FuM-Net to learn FuM-Score from fusion-aware representations with auxiliary source references, enabling efficient automatic quality estimation. Experiments verify the reliability and generalizability of our benchmark for MIQA. FuM-Net also delivers superior performance over recent state-of-the-art methods. The main contributions are as follows:

- For the first time, we propose a unified benchmark that characterizes IVIF quality toward machine perception and understanding preferences.
- We introduce FuM-Score to quantify machine preferences and build a large-scale dataset FuM-18K with fine-grained annotations for MIQA.
- We develop FuM-Net to learn FuM-Score from fusion-aware representations with auxiliary source references, enabling efficient automatic IVIF quality assessment.

2 Related Work

Infrared-Visible Image Fusion. Deep learning-based IVIF has undergone remarkable progress from convolutional neural networks (CNNs) (Li & Wu, 2018; Xu et al., 2020; Li et al., 2021) to vision transformers (Zhao et al., 2023a; 2024b; Yi et al., 2024) and diffusion models (Zhao et al., 2023b; Wang et al., 2025d). Earlier methods emphasize thermal target enhancement and visual detail preservation (Li & Wu, 2018; Xu et al., 2020; Li et al., 2021; Zhao et al., 2023a; Wang et al., 2025c)

that HVS prefers. EVAFusion (Liu et al., 2026) further learns a reward model from human feedback to better align fusion outputs with human perceptual preferences. More recently, growing interest in effects on downstream applications has inspired many task-adaptive methods, which leverage auxiliary task-driven objectives (Liu et al., 2022; 2024b; Bai et al., 2025a; Yang et al., 2026), semantic priors (Tang et al., 2023; Liu et al., 2023; Wu et al., 2025), or specific prediction heads (Zhang et al., 2024; Zhao et al., 2025; Wang et al., 2025a), to improve relevant task performance. PALDM (Su et al., 2026) devises direct preference optimization that combines human-centric and task-specific objectives to promote both perceptual quality and cross-task transferability. Such achievements underscore the significance of MIQA in IVIF. This work establishes a unified benchmark across multiple downstream perception and understanding tasks for machine-centric assessment.

Image Quality Assessment. IQA (Wang et al., 2004) is a long-standing research topic that endeavors to quantify image quality, providing a basis for evaluating various image processing techniques including IVIF. Modern IQA methods (Sheikh et al., 2005; Mittal et al., 2012a; Ding et al., 2022; Zhang et al., 2018a; Wang et al., 2023) mostly aim to predict human-perceived quality by assessing attributes such as structural fidelity, texture clarity, and visual realism. As autonomous systems become prevalent, IQA for machines has also emerged as an important perspective, measuring how visual impairments affect machine performance. One representative example is face image quality assessment (Boutros et al., 2023; Ou et al., 2025), which reflects the recognition utility of samples. Wang et al. (2025b) present region-aware IQA for machines, which evaluates machine visual quality on multiple high-level perception tasks through fine-grained spatial degradation analysis. MPD (Li et al., 2025) enriches this connotation by combining classical perception and related vision-language tasks, which compares the performance consistency between distorted and reference images to benchmark machine preference. This is naturally not suitable to IVIF whose ideal reference labels are unavailable. In IVIF, existing approaches (Zhang et al., 2020; Chang et al., 2023; Cheng et al., 2026) simultaneously emphasize the texture preservation degree of source visible images and the structure fidelity of infrared images toward HVS. FuScore (Guo et al., 2026) employs an MLLM to emulate human visual perception and achieves continuous quality scoring. This work responds to the necessity of IVIF-IQA toward machine perception and understanding preferences, providing a standardized evaluation protocol, benchmark, and baseline to advance the field.

3 Benchmark Construction

3.1 Principle Definition

We benchmark machine preferences by measuring the fusion utility in downstream applications. It should be able to quantify how effectively machines equipped with fused images achieve semantic perception and vision-language understanding. In general, perception concerns object recognition, localization, and structural semantics, whereas understanding focuses on interpreting and describing scene contents. Therefore, we form an evaluation set $T = \{\mathrm{Det}, \mathrm{Seg}, \mathrm{VQA}, \mathrm{Cap}\}$, where Det and Seg respectively denote object detection and semantic segmentation for machine perception. VQA and Cap refer to visual question answering and image captioning for understanding. Then, we introduce FuM-Score, which jointly accounts for the utility on machine tasks and the gain or degradation relative to source images, *i.e.*, fusion gains. Specifically, given an infrared-visible source pair $(I_{\mathrm{ir}}, I_{\mathrm{vi}})$ and its fused image I_f, we define FuM-Score as:

$$Q_t = \mathrm{clip}\left(U_t + \eta G_t, 0, 1\right), \qquad t \in T, \tag{1}$$

where U_t denotes the task utility and G_t denotes the positive/negative gains relative to source B_t on task t. η controls the contribution of G_t, which is empirically set to $\eta = 0.4$.

3.2 Dataset Collection

To compute FuM-Score across selected tasks, it is essential to have source and fused images with shared fine-grained annotations, along with standardized evaluation protocols that ensure task responses are directly comparable. To this end, we construct FuM-18K and establish unified task evaluation, as illustrated in Figure 2.

FuM-18K Dataset. We collect 1,892 registered infrared-visible image pairs from FMB (Liu et al., 2023), LLVIP (Jia et al., 2021), and MFAD (Hu et al., 2025), covering diverse scenes, illumina-

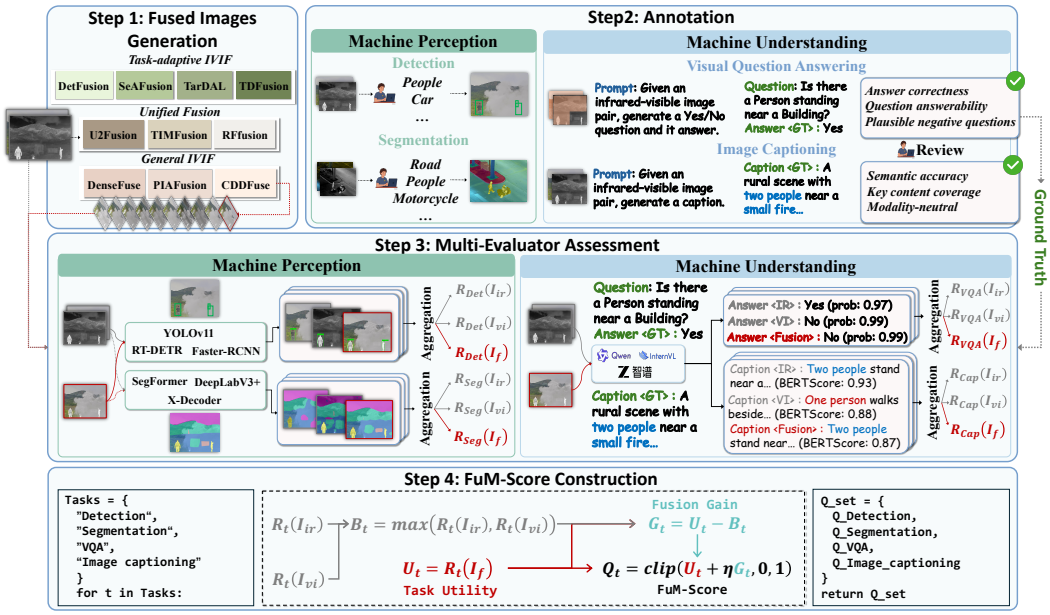

Figure 2: Construction pipeline of FuM-18K. First, we collect and generate fused images using 10 IVIF methods. Second, we unify detection and segmentation annotations under a common label set. For understanding tasks, we first use MLLM to generate VQA and caption references from the source pairs. We jointly consider task utility and fusion gains to obtain FuM-Score.

tion conditions, and viewpoints. Each source pair is processed by 10 representative fusion methods spanning three categories: 1) General IVIF methods including DenseFuse (Li & Wu, 2018), PIAFusion (Tang et al., 2022b), and CDDFuse (Zhao et al., 2023a); 2) Unified fusion frameworks including U2Fusion (Xu et al., 2020), TIMFusion (Liu et al., 2024b), and RFfusion (Wang et al., 2025d); 3) Task-adaptive IVIF methods including DetFusion (Sun et al., 2022), SeAFusion (Tang et al., 2022a), TarDAL (Liu et al., 2022), and TDFusion (Bai et al., 2025a). These methods generate per-scene fusion outputs under shared source images and annotations.

Perception Annotations. In FuM-18K, the IVIF datasets used differ in annotation availability and category definitions. For example, FMB covers scenes mainly from the vehicle perspective with segmentation annotations for 14 categories, while LLVIP mainly comes from street view surveillance images with annotations for a single detection category. This discrepancy hinders their direct use for unified evaluation. To address this problem, we construct a common label set for each task across the collected images, covering 6 detection categories (Liu et al., 2022) and 14 segmentation categories (Liu et al., 2023). We retained existing annotations when available and employed four annotators to complete the missing labels (total 17,794 bounding boxes and 5,467 semantic mask regions) and ensure consistency. In this way, we can yield 18,987 bounding boxes and 45,380 semantic mask regions, with all annotations carefully manually reviewed for evaluating machine perception capability.

Understanding Annotations. To our knowledge, existing IVIF datasets lack complete annotations for understand tasks. This work focuses on VQA and image caption, and annotate on paired infrared and visible images to build reliable labels. For VQA, we adopt a binary VQA setting to assess object presence, counts, and spatial relationships. We use an object-centered cropping strategy and provide the aligned source regions to Qwen3-VL-Plus (Bai et al., 2025b) to generate questions about these attributes. Each region is assigned two complementary Yes/No questions. Reviewers then verify answer correctness, single-image answerability, and the plausibility of negative (No) questions. The retained questions and the corresponding identical crops are reused for both the source and fused images. For image caption, Qwen3-VL-Plus receives each full source pair to describe objects, scene layout, and spatial relationships, producing 1,892 reference captions. Reviewers check semantic accuracy, key content coverage, and modality-neutral wording. The same reference caption is used to evaluate the source images and their fusion results. More details are in Appendix A.1.

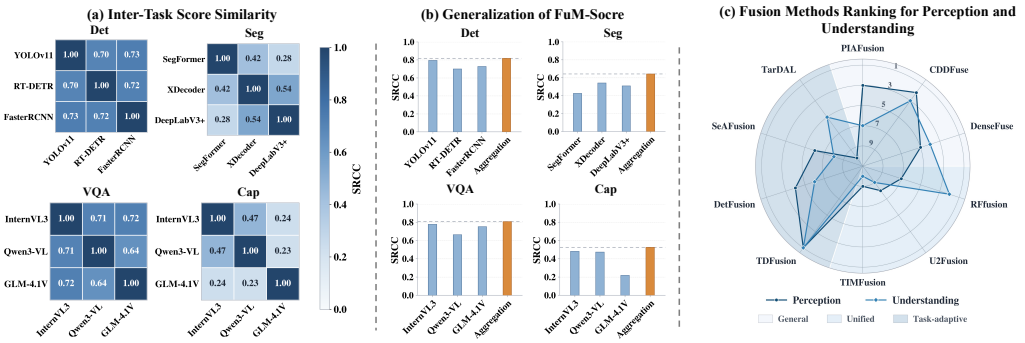

Figure 3: Analysis of machine preference assessment. (a) Inter-task score similarity measured by SRCC of FuM-Scores. (b) Agreement of individual and aggregated assessments with unseen evaluators. (c) Mean ranks of fusion methods across perception and understanding tasks; lower ranks indicate better performance.

3.3 Unified Task Evaluation

Evaluation Settings. For perception tasks, we select commonly used evaluators with different architectures. To be specific, in object detection, we apply the single-stage detector YOLOv11 (Khanam & Hussain, 2024a), the two-stage detector Faster R-CNN (Ren et al., 2016), and the transformer-based detector RT-DETR (Zhao et al., 2024a). Semantic segmentation uses the CNN-based DeepLabV3+ (Chen et al., 2018), the transformer-based SegFormer (Xie et al., 2021), and X-Decoder (Zou et al., 2023) that incorporates visual-language representations. For understanding tasks, we select 3 recent open-source MLLMs, Qwen3-VL-8B (Bai et al., 2025b), GLM-4.1V-9B (Hong et al., 2025), and InternVL3-8B (Zhu et al., 2025b), as evaluators for both VQA and image caption, where each processes the infrared, visible, and fused images independently.

Task Responses. We present FuM-Score (Eq. (1)) that quantify the utility and fusion gains on machine tasks as the measurement of MIQA. Supposing the task t related to object detection, let $r_t^{(e)}(I)$ denote the image-level response of evaluator e to image I. We average the IoU of valid, category-consistent matches over all annotated objects, assigning zero to unmatched objects. Segmentation is evaluated using image-level mIoU. For binary VQA, we apply softmax to the Yes and No logits and utilize the probability assigned to the correct answer. Question scores are then averaged within each region and subsequently aggregated with equal weight across the represented object categories. For image caption, we compare the generated caption with the shared reference caption using BERTScore-F1. All task responses lie in $[0, 1]$, with higher values indicating better task performance. More details about task response collection are in Appendix A.2.

Evaluator Aggregation. The above procedure provides image-level responses from multiple evaluators. To reduce dependence on any single model, we aggregate the three evaluator responses for each task and image using an unweighted mean: $R_t(I) = \frac{1}{3}\sum_{e=1}^{3} r_t^{(e)}(I)$, where $I \in \{I_{\text{ir}}, I_{\text{vi}}, I_f\}$ and $R_t(I)$ is the aggregated response. The task utility U_t and source reference B_t are defined as:

$$U_t = R_t(I_f), \qquad B_t = \max\left(R_t(I_{\text{ir}}), R_t(I_{\text{vi}})\right), \qquad G_t = U_t - B_t, \tag{2}$$

We aggregate the evaluator scores for each source and define the higher score between the two sources B_t. Positive G_t indicates that the fused image outperforms both sources, while negative indicates degradation after fusion. Substituting U_t and G_t into Eq. (1) yields the final FuM-Score.

3.4 Benchmark Analysis

Given the discrepancies across downstream tasks, we compute the inter-task score similarity to measure performance consistency. Figure 3(a) visualizes these similarities via heatmaps, where darker colors indicate higher similarity. We observe that detection and VQA exhibit relatively consistent patterns, whereas segmentation and image caption show more scattered similarities. This suggests that evaluating a single downstream model may not adequately reflect the overall quality of downstream performance. To address this, we adopt downstream task score aggregation. As shown in Figure 3(b), each bar chart corresponds to one downstream task. For that task, we obtain task responses from three downstream task models participating in our dataset and also compute

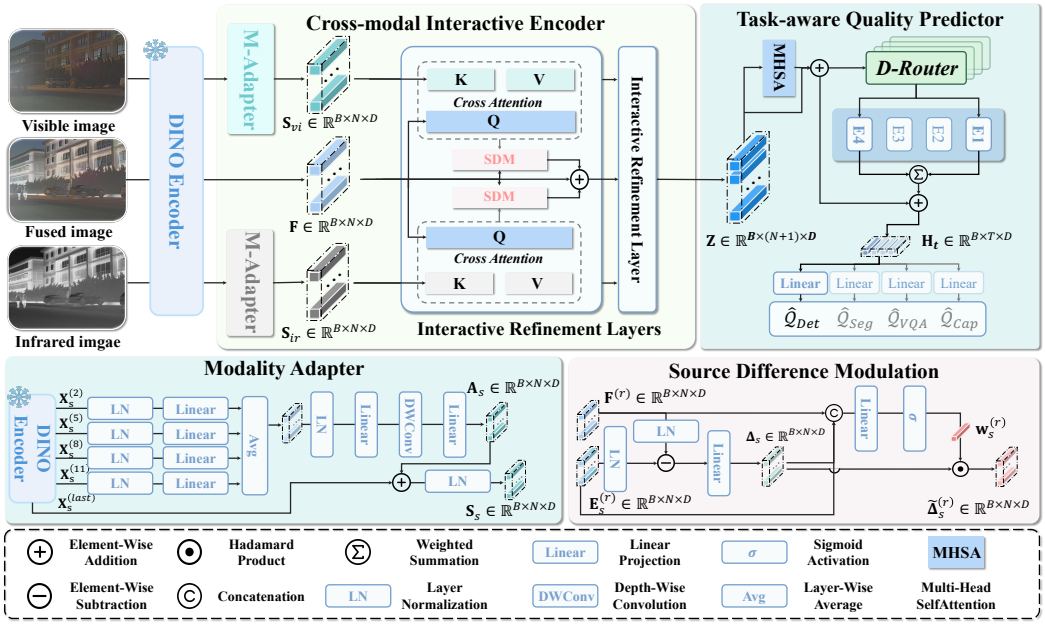

Figure 4: Overview of FuM-Net. It contains two important parts, which are the cross-modal interactive encoder and the task-aware quality predictor. The interactive encoder refines the fused features through cross-attention with the source features and source difference modulation. Then the task-aware quality predictor uses a dynamic router to combine shared experts, followed by separate prediction heads to produce FuM-Scores.

the aggregated FuM-Score by aggregating them. Specifically, we use a held-out task model that was not involved in constructing FuM-18K (YOLOv5 (Khanam & Hussain, 2024b) for detection, DPT-Large (Ranftl et al., 2021) for segmentation, LLaVA-OneVision (Li et al., 2024) for VQA, and Gemma-4 (Team et al., 2026) for caption) to produce its individual FuM-Score. We then measure the similarity between the held-out model's FuM-Score and those of the three individual models, as well as their aggregated FuM-Score. The results demonstrate that aggregation achieves higher similarity with domain-specific task models, leading to a more reliable evaluation.

Moreover, we also analyze whether machines in perception and understanding reveal different strengths and limitations of fusion methods. Figure 3(c) compares their mean ranks across the two task groups. PIAFusion ranks better in perception than in understanding, while RFfusion shows the opposite pattern. TDFusion ranks well in both groups. These results show that some methods perform modestly across tasks, whereas others have obvious advantages in particular tasks. This verifies the necessity of the two dimensions for MIQA.

4 Baseline Framework

We propose FuM-Net, an efficient yet effective framework to predict FuM-Score for MIQA. As illustrated in Figure 4, we first apply a shared frozen DINOv3 (Siméoni et al., 2025) to extract features from the infrared, visible, and fused images. A cross-modal interactive encoder is then designed to exploit complementary information from source modalities and strengthen the fused representations, which are then fed into a task-aware quality predictor to produce the FuM-Score.

Cross-modal Interactive Encoder. Given the patch tokens of the respective infrared, visible, and fused images extracted at a certain layer in DINOv3, there are two parallel modality adapters (M-Adapters), one for each source, to prepare source features for interaction with the fused tokens. Each adapter aggregates projected representations from a set of selected DINO layers and adds a learned residual to the final-layer source tokens, producing adapted features $\mathbf{S}_{ir}$ and $\mathbf{S}_{vi}$.

After that, we conduct interactive refinement to enrich the fusion representations with informative cues from $\mathbf{S}_{ir}$ and $\mathbf{S}_{vi}$. Specifically, we use the output tokens $\mathbf{X}_{fuse}$ of the fused image by the DINO encoder as the query and refine them through cross-attention, formulated as:

$$\mathbf{E}_{vi}^{(r)} = \mathcal{C}_{vi}^{(r)}\left(Q = \mathbf{F}^{(r-1)}, K, V = \mathbf{S}_{vi}\right), \qquad \mathbf{E}_{ir}^{(r)} = \mathcal{C}_{ir}^{(r)}\left(Q = \mathbf{F}^{(r-1)}, K, V = \mathbf{S}_{ir}\right) \tag{3}$$

where $\mathbf{F}^{(r-1)}$ denotes the output tokens from the $(r-1)$-th refinement layer, serving as the input to the r-th layer. We initialize $\mathbf{X}_{fuse}$ as $\mathbf{F}^{(0)}$ for the first layer. $\mathcal{C}_{ir}^{(r)}$ and $\mathcal{C}_{vi}^{(r)}$ are the cross-attention operations corresponding to each modality feature and produce $[\mathbf{E}_{ir}^{(r)}, \mathbf{E}_{vi}^{(r)}]$. We further propose the source difference modulation (SDM), which models the differences between the resultant features and the input $\mathbf{F}^{(r-1)}$. Here, we use only the visible modality for simplicity, where the difference $\mathbf{\Delta}_{vi}^{(r)}$ is obtained by $\mathbf{\Delta}_{vi}^{(r)} = \mathcal{D}_{vi}^{(r)}\left(\mathrm{LN}(\mathbf{E}_{vi}^{(r-1)}) - \mathrm{LN}(\mathbf{F}^{(r-1)})\right)$. By combining $\mathbf{\Delta}_{vi}^{(r)}$ and corresponding visible features through a linear layer $\mathcal{G}_{vi}^{(r)}$, SDM predicts a weight $\mathbf{w}_{vi}^{(r)}$ for each location to control the contribution of $\mathbf{\Delta}_{vi}^{(r)}$

$$\mathbf{w}_{vi}^{(r)} = \sigma\left(\mathcal{G}_{vi}^{(r)}\left[\mathbf{F}^{(r)}; \mathbf{E}_{vi}^{(r)}; \mathbf{\Delta}_{vi}^{(r)}\right]\right), \qquad \widetilde{\mathbf{\Delta}}_{vi}^{(r)} = \mathbf{w}_{vi}^{(r)} \odot \mathbf{\Delta}_{vi}^{(r)}. \tag{4}$$

We multiply each difference token by its weight and add the results from both sources, enabling region-adaptive enhancement. We concatenate the output from all the refinement layers to generate the final feature $\mathbf{Z}$ for later quality prediction.

Task-aware Quality Predictor. To predict a unified FuM-Score across four tasks, FuM-Net leverages shared image representations $\mathbf{Z}$ while modeling task-specific judgment criteria. We construct a dynamic router (D-Router) composed of a shared expert group and task-specific sub-routers. The shared experts provide common feature processing capabilities, and each sub-router learns an input-dependent selection over experts, enabling different tasks to emphasize distinct expert combinations based on their own scoring criteria

$$\mathbf{H} = \mathbf{Z} + \mathrm{MHSA}(\mathrm{LN}(\mathbf{Z})), \qquad \mathbf{H}_t = \mathbf{H} + \sum_{k=1}^{K} \boldsymbol{\pi}_{t,k} \odot \mathcal{E}_k(\mathbf{H}), \tag{5}$$

where MHSA denotes multi-head self-attention for feature enhancement, resulting in $\mathbf{H}$. $\mathcal{R}_t$ is the router for task t, and $\mathcal{E}_k$ is the k-th shared expert in the group of total $K = 4$ experts. $\boldsymbol{\pi}_{t,k}$ denotes the router with the k-th expert. The task-specific outputs are fed into the corresponding linear prediction heads to generate the FuM-Score for each task.

Training Objective. We train FuM-Net with a smooth $L1$ loss and an expert load balancing loss:

$$\mathcal{L} = \frac{1}{|T|}\sum_{t \in T} \mathrm{SmoothL1}\left(\widehat{\mathcal{Q}}_t, \mathcal{Q}_t\right) + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}, \tag{6}$$

where $\widehat{\mathcal{Q}}_t$ and $\mathcal{Q}_t$ are the predicted and reference FuM-Scores for task t, and $T = \{\mathrm{Det}, \mathrm{Seg}, \mathrm{VQA}, \mathrm{Cap}\}$. Smooth L1 penalizes score prediction errors averaged over the four tasks and all images in each training batch. $\mathcal{L}_{\mathrm{bal}}$ encourages the routers to use shared experts evenly. Details about $\mathcal{L}_{\mathrm{bal}}$ are in Appendix D.1.

5 Experiments

5.1 Experimental Settings

Dataset and Metrics. We conduct experiments on the FuM-18K dataset, consisting of 1,892 infrared–visible image pairs and 18,920 fused samples from 10 distinct IVIF methods. This dataset provides fine-grained annotations for object detection, semantic segmentation, visual question answering, and image captioning. We divide the overall dataset into training, validation, and testing subsets at a ratio of 8:1:1. Similar to existing IQA methods, Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC) are employed to evaluate the linear correlation and ranking consistency of our method, respectively, where both scale from -1 to 1, with values close to 1 indicating better performance.

Implementation Details. In FuM-Net, all images are resized and center-cropped to 512×512 for training and testing. We optimize the model using AdamW for 20 epochs with a batch size of 8 and a weight decay of 0.01. The learning rate increases linearly to 1×10^{-4} during the first five epochs and then follows a cosine decay. All experiments are performed on an NVIDIA RTX 4090 GPU.

Table 1: Comparison with existing assessment methods on FuM-18K. The best and second-best results are **highlighted** and underlined, respectively.

Type	Method	Det		Seg		VQA		Cap		Overall	
		PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
Fusion Metrics	EN (Roberts et al., 2008)	0.209	0.222	0.079	0.070	0.031	0.008	0.127	0.146	0.173	0.165
	SD (Rao, 1997)	0.186	0.185	0.061	0.091	0.045	0.007	0.073	0.091	0.157	0.157
	AG (Cui et al., 2015)	0.217	0.191	0.427	0.410	-0.008	-0.046	0.268	0.289	0.292	0.245
	SF (Eskicioglu & Fisher, 1995)	0.106	0.091	0.329	0.304	-0.056	-0.084	0.269	0.278	0.170	0.137
	SCD (Aslantas & Bendes, 2015)	0.232	0.258	0.247	0.278	0.146	0.134	0.010	-0.024	0.287	0.303
	SSIM (Wang et al., 2004)	-0.341	-0.357	-0.200	-0.250	-0.082	-0.079	-0.076	-0.071	-0.317	-0.323
	VIF (Han et al., 2013)	-0.117	-0.111	0.036	0.008	0.024	0.038	-0.082	-0.096	-0.054	-0.051
	$Q^{AB/F}$ (Xydeas et al., 2000)	0.014	0.020	0.230	0.194	-0.012	0.010	0.079	0.059	0.089	0.085
FR-IQA	LPIPS (Zhang et al., 2018a)	0.530	0.539	0.571	0.538	0.380	0.372	0.116	0.137	0.447	0.463
	DISTS (Ding et al., 2022)	0.696	0.708	0.873	0.862	0.463	0.480	0.462	0.448	0.702	0.709
	AHIQ (Lao et al., 2022)	0.704	0.721	0.904	0.891	0.498	0.507	0.509	0.501	0.729	0.727
	DeepDC (Zhu et al., 2025a)	0.187	0.165	0.013	0.003	-0.055	-0.086	0.179	0.172	0.106	0.095
	A-FINE (Chen et al., 2025)	0.128	0.095	0.053	0.044	0.136	0.113	-0.027	-0.008	0.149	0.109
NR-IQA	NIQE (Mittal et al., 2012b)	0.178	0.226	0.042	0.062	0.090	0.120	0.027	0.023	0.160	0.207
	HyperIQA (Su et al., 2020)	0.754	0.763	0.935	0.926	0.583	0.603	0.165	0.590	0.778	0.778
	DBCNN (Zhang et al., 2018b)	0.766	0.778	0.906	0.896	0.569	0.565	0.402	0.377	0.783	0.778
	LIQE (Zhang et al., 2023)	0.759	0.768	0.892	0.913	0.602	0.614	0.656	0.650	0.751	0.742
	Q-Align (Wu et al., 2023)	0.799	0.811	0.902	0.921	0.592	0.605	**0.661**	**0.663**	0.745	0.741
MIQA & Generic Backbones	ResNet50 (NR) (He et al., 2016)	0.709	0.718	0.896	0.880	0.477	0.475	0.517	0.532	0.719	0.713
	ResNet50 (FR) (He et al., 2016)	0.689	0.703	0.834	0.841	0.481	0.499	0.418	0.434	0.696	0.712
	ViT-S (NR) (Dosovitskiy et al., 2020)	0.681	0.681	0.902	0.885	0.494	0.502	0.562	0.539	0.725	0.722
	ViT-S (FR) (Dosovitskiy et al., 2020)	0.675	0.675	0.862	0.841	0.490	0.480	0.225	0.198	0.695	0.697
	RA-MIQA (Wang et al., 2025b)	0.797	0.809	0.939	0.932	0.576	0.580	0.635	0.633	0.826	0.822
Ours	**FuM-Net (Ours)**	**0.828**	**0.834**	**0.944**	**0.933**	**0.620**	**0.632**	0.632	0.646	**0.848**	**0.841**

Table 2: Trainable parameters (M) and runtime (ms) comparisons using 512×512 resolution.

Method	$Q^{AB/F}$	HyperIQA	DBCNN	LIQE	Q-Align	AHIQ	DeepDC	A-FINE	RA-MIQA	**FuM-Net**
Params. (M) ↓	–	27.375	15.311	151.277	148.582	27.169	–	–	22.186	**12.179**
Time (ms) ↓	126.511	39.972	31.406	41.116	316.464	302.388	16.867	**15.112**	46.349	16.277

Competing Methods. We compare our FuM-Net with 21 mainstream IQA metrics comprising 1) 8 metrics typically used in IVIF: EN (Roberts et al., 2008), SD (Rao, 1997), AG (Cui et al., 2015), SF (Eskicioglu & Fisher, 1995), SCD (Aslantas & Bendes, 2015), SSIM (Wang et al., 2004), VIF (Han et al., 2013), and $Q^{AB/F}$ (Xydeas et al., 2000); 2) 5 full-reference metrics (FR-IQA): LPIPS (Zhang et al., 2018a), DISTS (Ding et al., 2022), AHIQ (Lao et al., 2022), DeepDC (Zhu et al., 2025a), and A-FINE (Chen et al., 2025); 3) 5 Non-reference metrics (NR-IQA): NIQE (Mittal et al., 2012b), HyperIQA (Su et al., 2020), DBCNN (Zhang et al., 2018b), LIQE (Zhang et al., 2023), and Q-Align (Wu et al., 2023); and 4) 2 representative backbones and 1 machine-centric IQA (MIQA) method: ResNet-50 (He et al., 2016), ViT-small (Dosovitskiy et al., 2020), and RA-MIQA (Wang et al., 2025b). To ensure fair comparisons, all methods are retrained on our FuM-18K, with the exception of typical non-deep learning approaches, DeepDC and A-FINE.

5.2 Main Results

Quality Prediction Performance. Table 1 reports the performance of MIQA on four perception and understanding tasks. Conventional fusion metrics based on information statistics show limited effectiveness for MIQA tasks. Notably, even FR-IQA methods remain weak on MIQA, despite full-supervised training on FuM-18K. This suggests that these approaches mainly capture low-level information but ignore the preservation of semantic consistency. LIQE learns textual templates that provide auxiliary scene knowledge and Q-Align instructs MLLMs with discrete text-defined quality levels to introduce scene knowledge, which achieve relatively stronger VQA and image caption performance than other settings, yet remain obviously inferior to our FuM-Net in detection and segmentation. Although RA-MIQA devises a region-aware model with fine-grained degradation analysis for MIQA, it still yields suboptimal results. In contrast, FuM-Net delivers consistently superior performance, outperforming existing methods by a substantial margin across all tasks. Furthermore, we calculate the overall performance by aggregating the average PLCC/SRCC values over four tasks for each image, where the results validate the superiority of FuM-Net.

Model Complexity. We quantitatively compare model complexity across multiple IQA methods in terms of trainable model parameters and inference speed, using the same NVIDIA RTX 4090 GPU. As shown in Table 2, FuM-Net achieves the fewest parameters and the fastest runtime among nearly

Table 3: FuM-Score validation on five unseen fusion methods across four tasks.

Fusion Method	Det		Seg		VQA		Cap		FuM-Score
	IoU ↑	mAP50 ↑	mIoU ↑	mAcc ↑	Acc ↑	Recall ↑	BERTScore ↑	METEOR ↑	
Text-IF (Yi et al., 2024)	**0.5015**	**0.3977**	0.4195	**0.4898**	**0.8429**	**0.7851**	0.9065	**0.2453**	**0.6488**
UniFusion (Li et al., 2026)	0.4825	0.3753	**0.4226**	0.4730	0.8383	0.7716	0.9061	0.2414	0.6410
FreeFusion (Zhao et al., 2025)	0.4802	0.3380	0.4080	0.4619	0.8316	0.7593	**0.9069**	**0.2453**	0.6337
DCEvo (Liu et al., 2025)	0.4749	0.3569	0.4149	0.4673	0.8280	0.7471	0.9062	0.2412	0.6336
DPOFusion (Su et al., 2026)	0.4458	0.3004	0.3636	0.4321	0.8041	0.7175	0.9066	0.2437	0.5966

Table 4: Ablations on input in FuM-Net. IR, VIS, and F denote infrared, visible, and fused.

Input			Overall	
IR	VIS	F	PLCC ↑	SRCC ↑
–	–	✓	0.839	0.834
✓	–	✓	0.832	0.826
–	✓	✓	0.826	0.821
✓	✓	✓	**0.848**	**0.841**

Table 5: Ablations on the components of in the cross-modal interactive encoder.

Configuration	Overall	
	PLCC ↑	SRCC ↑
w/ all	**0.848**	**0.841**
w/o M-Adapter	0.840	0.831
w/o Interactive Refinement	0.839	0.834
w/o SDM	0.835	0.825

all competing methods. Though A-FINE is marginally faster, it leads to reduced quality prediction accuracy (see Table 1). Such a comparison demonstrates that FuM-Net offers optimal trade-off between efficiency and effectiveness.

Consistency of FuM-Score. In Table 3, we evaluate 5 recent IVIF methods, Text-IF (Yi et al., 2024), UniFusion (Li et al., 2026), FreeFusion (Zhao et al., 2025), DCEvo (Liu et al., 2025), and DPOFusion (Su et al., 2026), by testing their fused images on four tasks. The results validate a high consistency between the overall FuM-Score and downstream task performance.

5.3 Ablation Study

Effect of Input Modalities in FuM-Net. Table 4 shows that the fused image alone provides a strong baseline, while adding either source individually reduces performance. By combining both source images with the fused one, the model achieves the best in PLCC and SRCC.

Components in Cross-Modal Interactive Encoder. Table 5 analyzes the components of the cross-modal interactive encoder. As shown, removing the Source Adapter leads to decreased prediction accuracy. When the interactive refinement layers are removed, the model degenerates to the fusion-only configuration, result in suboptimal performance due to the missing source information. The ablation on SDM also validate the importance of the difference modeling between the source and fusion features.

Table 6: Ablations on the components in the task-aware quality predictor.

Configuration	Overall	
	PLCC ↑	SRCC ↑
Shared MLP	0.833	0.828
Task-Private MLP	0.826	0.827
Shared Expert Group + Shared Router	0.822	0.824
Shared Expert Group + Task-specific Sub-Routers	**0.848**	**0.841**

Task-Aware Quality Predictor. Table 6 compares alternative prediction modules. Shared MLP uses a common transformation across tasks, whereas task-private MLP assigns a separate transformation to each task. Their overall SRCC values are similar. Using shared experts with a shared router yield the worse performance with a SRCC of 0.824. FuM-Net combines shared experts with task-specific routers and achieves the highest SRCC/PLCC, supporting the routing configuration in FuM-Net. More ablations are in Appindix D.2

6 Conclusion

In this paper, we present a machine-centric perspective for IVIF by relating fusion quality to downstream perception and understanding performance. To measure this preference in a unified way, we introduce FuM-Score, which aggregates fusion gains over representative tasks under standardized evaluation. Based on FuM-Score, we construct FuM-18K, a large-scale IVIF benchmark that includes diverse scenes, representative fusion models, and fine-grained task-oriented annotations. We further propose FuM-Net to predict FuM-Score from fusion representations augmented with auxiliary source references, enabling efficient automatic quality assessment. Extensive experiments validate that our benchmark captures machine preference effectively and that FuM-Net achieves consistent and stronger performance than recent state-of-the-art approaches.

AI USE STATEMENT

In this work, we used the Qwen3-VL-Plus API to generate reference visual question-answer pairs and image captions, followed by manual review. We also used locally deployed MLLMs to collect VQA and captioning responses for infrared, visible, and fused images. We take responsibility for the final content of this work, including the AI-assisted annotations and evaluation results.

REFERENCES

Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of Electronics and Communications*, 69 (12):1890–1896, 2015.

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7457–7468, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19593–19604, 2023.

Zhihao Chang, Shuyuan Yang, Zhixi Feng, Quanwei Gao, Shengzhe Wang, and Yuyong Cui. Semantic-relation transformer for visible and infrared fused image quality assessment. *Information Fusion*, 95:454–470, 2023.

Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: relaxing the perfect reference quality assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12742–12752, June 2025.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, pp. 833–851, 2018.

Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Tao Zhou, Hui Li, Zhangyong Tang, and Josef Kittler. Evanet: Towards more efficient and consistent infrared and visible image fusion assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 9914–9929, 2026.

Guangmang Cui, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

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

Ahmet M Eskicioglu and Paul S Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965, 1995.

Yuchen Guo, Junli Gong, Yao Lu, Xintong Xu, Yiuming Cheung, and Weifeng Su. Bringing multimodal large language models to infrared-visible image fusion quality assessment. *arXiv preprint arXiv:2605.06969*, 2026.

Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ke Hu, Yudong He, Yuan Li, Jiayu Zhao, Song Chen, and Yi Kang. Ei^{2}det: Edge-guided illumination-aware interactive learning for visible-infrared object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):7101–7115, 2025.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3489–3497, 2021.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024a.

Rahima Khanam and Muhammad Hussain. What is yolov5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:2407.20892*, 2024b.

Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1139–1148, 2022.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 105–114, 2017.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Chunyi Li, Yuan Tian, Xiaoyue Ling, Zicheng Zhang, Haodong Duan, Haoning Wu, Ziheng Jia, Xiaohong Liu, Xiongkuo Min, Guo Lu, et al. Image quality assessment: From human to machine preference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7570–7581, 2025.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

Xingyuan Li, Songcheng Du, Yang Zou, Haoyuan Xu, Zhiying Jiang, and Jinyuan Liu. Unifusion: A unified image fusion framework with robust representation and source-aware preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33869–33880, 2026.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1132–1140, 2017.

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. 5792–5801, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 8081–8090, 2023.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2024a.

Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2226–2235, 2025.

Jinyuan Liu, Xingyuan Li, Qingyun Mei, Haoyuan Xu, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Bridging human evaluation to infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12322–12333, 2026.

Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024b.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012b.

Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. Mr-fiqa: Face image quality assessment with multi-reference representations from synthetic data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12915–12925, 2025.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):1, 2002.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12159–12168, 2021.

Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement Science and Technology*, 8(4):355–375, 1997.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3664–3673, 2020.

Weijian Su, Songqian Zhang, Yuqi Han, Jian Zhuang, Yongdong Huang, and Qiang Zhang. Fusion in your way: Aligning image fusion with heterogeneous demands via direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41499–41509, 2026.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4003–4011, 2022.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022a.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92, 2022b.

Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity. *Information Fusion*, 99:101870, 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Liying Wang, Xiaoli Zhang, Chuanmin Jia, and Siwei Ma. Mafs: Masked autoencoder for infrared-visible image fusion and semantic segmentation. *IEEE Transactions on Image Processing*, 34: 6490–6505, 2025a.

Xiaoqi Wang, Yun Zhang, and Weisi Lin. Image quality assessment for machines: Paradigm, large-scale database, and models. *arXiv preprint arXiv:2508.19850*, 2025b.

Zeyu Wang, Jizheng Zhang, Haiyu Song, Mingyu Ge, Jiayu Wang, and Haoran Duan. Highlight what you want: Weakly-supervised instance-level controllable infrared-visible image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12637–12647, 2025c.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. *Advances in Neural Information Processing Systems*, 38: 22927–22948, 2025d.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17882–17891, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2020.

Costas S Xydeas, Vladimir Petrovic, et al. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

Zengyi Yang, Yu Liu, Juan Cheng, Zhiqin Zhu, Yafei Zhang, and Huafeng Li. Customized fusion: A closed-loop dynamic network for adaptive multi-task-aware infrared-visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 188–198, 2026.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27016–27025, 2024.

Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information fusion*, 76:323–336, 2021a.

Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26964–26973, 2024.

Hao Zhang, Yanping Zha, Zizhuo Li, Meiqi Gong, and Jiayi Ma. Magicfuse: Single image fusion for visual and semantic reinforcement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26551–26560, 2026.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018a.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018b.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081. IEEE, 2023.

Xingchen Zhang, Ping Ye, and Gang Xiao. Vifb: A visible and infrared image fusion benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 468–478, 2020.

Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024a.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5906–5916, 2023a.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 8048–8059, 2023b.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25912–25921, 2024b.

Hanwei Zhu, Baoliang Chen, Lingyu Zhu, Shiqi Wang, and Weisi Lin. Deepdc: Deep distance correlation as a perceptual image quality evaluator. *IEEE Transactions on Image Processing*, 34: 7859–7873, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023.

A FuM-18K Construction and Annotation Details

A.1 Task Annotation

Object detection. Detection annotations specify object categories and bounding boxes using six categories following M3FD (Liu et al., 2022): Bus, Car, Motorcycle, People, Lamp, and Truck. We use Person as the textual name for the human category. Existing annotations are retained when available, and missing annotations are added and manually reviewed. Each object is stored in YOLO format as a category index and normalized bounding-box center, width, and height.

Semantic Segmentation. Segmentation annotations provide a semantic label for each annotated pixel under the 14 category label space of FMB (Liu et al., 2023), including Road, Sidewalk, Building, Traffic Light, Traffic Sign, Vegetation, Sky, Person, Car, Truck, Bus, Motorcycle, Bicycle and Pole. We retain available segmentation annotations and supplement missing masks using the same category definitions. Newly added annotations are manually reviewed.

Visual Question Answering: Region Selection. VQA concern annotated objects within up to three regions per source pair, with at most one region per category. Following prior work, we initially collected yes-or-no questions using full images. However, complex scenes, limited image resolution, and small objects made the relevant visual evidence difficult to identify. We therefore adopt object-centered crops to focus question generation and evaluation on local regions. Up to three object anchors are selected from the detection annotations of each source pair. We first retain the largest bounding box within each category and prioritize these category representatives by global category rarity and then box area. If fewer than three anchors have been selected, the remaining boxes are considered in the same order, allowing repeated categories at this stage. For an anchor box of width w and height h, the square crop has side length

$$L = \mathrm{clip}(\mathrm{round}(2.5 \max(w, h)), 128, 512), \tag{7}$$

where w, h, and L are measured in pixels. The crop is centered on the anchor and shifted inside the image when it crosses an image boundary; black padding is used only when necessary. Identical crop coordinates are applied to both source images and all ten fusion results. We also record the annotated objects whose overlap with the crop covers at least 20% of their bounding-box area. These records provide category and count information for subsequent question validation.

Visual Question Answering: Question Generation. For each aligned infrared-visible crop pair, we use Qwen3-VL-Plus (Bai et al., 2025b) to generate two binary questions, targeting one Yes answer and one No answer. The questions assess local semantic content and are designed for single-image evaluation. Figure 7 presents the prompt instructions.

Visual Question Answering: Manual Review. Reviewers inspect the paired source images and the corresponding local regions. They check answer correctness, single-image answerability, and the plausibility of negative questions, and revise question wording or answers when necessary. Here, single-image answerability means that a question can be posed without access to a second image or modality-specific metadata; it does not require both source modalities to provide equally strong visual evidence. The same question and reference answer are associated with the aligned infrared, visible, and fused crops. The one-Yes/one-No requirement is a generation target rather than a constraint imposed again after validation and manual revision. Finally, we annotate 5,422 regions and 10,844 generated questions. Rule-based validation retains 10,392 questions. These counts distinguish the initially generated questions from those passing validation and should not be interpreted as exactly two retained questions for every region.

Image Captioning: Reference Generation. We use Qwen3-VL-Plus to generate one English reference caption from each infrared-visible source pair. The caption describes major objects, scene layout, and spatial relationships using modality-neutral wording. Figure 8 presents the prompt instructions.

Image Captioning: Manual Review. Reviewers inspect each candidate caption together with the paired source images. Review focuses on semantic accuracy, key content coverage, and modality-neutral wording. Unsupported objects, unreliable counts, incorrect relations, and modality-specific or subjective descriptions are identified for correction. The review interface records the review

decision, revised caption, validity flag, and comments. The reference caption is associated with the source pair and shared across its source images and fusion results, without creating method-specific references.

A.2 Response Collection

For each task t, let $r_t(I)$ denote the image-level response of a fixed evaluator to image I. We omit the evaluator index for simplicity. Each evaluator processes the infrared, visible, and fused images independently using the same task annotations and response calculation. For VQA and captioning, the task-specific prompts are preceded by shared instructions that emphasize observable semantic and spatial content, standardized vocabulary, and modality-neutral wording.

Object Detection. For object detection, evaluators predict bounding boxes and category labels. For each ground-truth box, we select the candidate with the highest IoU among unused predictions. A match is accepted only when the categories agree and the IoU exceeds 0.5. The image-level response averages the matched IoU over all ground-truth objects, assigning zero to objects without a valid match:

$$r_{\mathrm{Det}}(I) = \frac{1}{N}\sum_{i=1}^{N} s_i, \qquad s_i = \begin{cases} \mathrm{IoU}(b_i, \hat{b}_i), & \text{if the match is valid,} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where N is the number of annotated objects, b_i is the ground-truth box, and $\hat{b}_i$ is its selected prediction. Accepted matches are one-to-one. Images without annotated objects receive a missing detection response.

Semantic Segmentation. For semantic segmentation, each evaluator produces a semantic mask, which is resized to the annotation resolution before evaluation. Pixels labeled 255 are ignored. The image-level response is the mean intersection-over-union over the included categories:

$$r_{\mathrm{Seg}}(I) = \frac{1}{|C|}\sum_{c\in C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{9}$$

where C denotes the set of categories included in the image-level evaluation. For each category c, TP_c is the number of pixels correctly predicted as c, FP_c is the number of pixels from other categories incorrectly predicted as c, and FN_c is the number of ground-truth pixels of c incorrectly assigned to other categories.

Visual Question Answering. Each evaluator receives one image region and one retained question, without the reference answer. The same question and crop coordinates are used across the infrared, visible, and fused images. Figure 10 presents the evaluation prompt, which requests a Yes-or-No answer based on the image evidence. We extract the Yes and No logits at the first generated-token position and apply softmax over these two logits only. For question k in region j, the response is the probability assigned to the reference answer:

$$p_{jk} = \frac{\exp\left(z_{jk}(y_{jk})\right)}{\exp\left(z_{jk}(\mathrm{Yes})\right) + \exp\left(z_{jk}(\mathrm{No})\right)}. \tag{10}$$

where y_{jk} is the reference answer, and $z_{jk}(a)$ is the next-token logit for answer a, conditioned on the image region and question. We first average valid question scores within each region and then average the region scores:

$$r_{\mathrm{VQA}}(I) = \frac{1}{M}\sum_{j=1}^{M} \frac{1}{n_j}\sum_{k=1}^{n_j} p_{jk}. \tag{11}$$

where M is the number of regions with valid questions, and n_j is the number of valid questions in region j. Since each region represents a distinct object category, this procedure assigns equal weight to each represented category.

Image Caption. Each evaluator receives a single full image and generates one English caption without access to the reference caption. Figure 9 presents the evaluation prompt, which is shared across the infrared, visible, and fused images and requests a description of observable semantic content and spatial relationships. The generated caption is compared with the shared reference caption using BERTScore-F1, computed for English text without baseline rescaling. The image-level response is

$$r_{\mathrm{Cap}}(I) = \mathrm{BERTScore}_{\mathrm{F1}}(\hat{y}, y). \tag{12}$$

where $\hat{y}$ and y denote the generated and reference captions, respectively. The reference caption is used only for scoring and is shared across the source images and their corresponding fusion results.

B Task Evaluators and FuM-Score Analysis

B.1 Evaluator Uncertainty and Threshold Calibration

We estimate evaluator disagreement to calibrate the task-specific thresholds used in the benchmark analysis. For each task, we compare the quantity obtained from one evaluator with that obtained from the remaining two evaluators. The leave-one-evaluator-out response is computed as

$$R_t^{(-e)}(I) = \frac{1}{2} \sum_{e' \neq e} r_t^{(e')}(I), \tag{13}$$

where e denotes the excluded evaluator. We compute the stronger source reference B_t, the infrared–visible response difference D_t, and the fusion gain G_t separately from the single-evaluator responses and the leave-one-evaluator-out responses. In particular, the stronger source is determined after averaging the retained source responses, following the same aggregation order as the formal labels.

For quantity G, we pool the absolute differences across the corresponding samples and the three choices of excluded evaluator. We use the 25th percentile of this distribution as an empirical tolerance around zero gain. We empirically set the threshold to the 25th percentile of this distribution to define a tolerance around zero fusion gain.:

$$\epsilon_t = \mathrm{Quantile}_{0.25}\left(\left\{ \left| G_{t,s}^{(e)} - G_{t,s}^{(-e)} \right| \right\}_{s,e} \right), \tag{14}$$

where s indexes the samples used to calibrate the corresponding quantity. Calibration uses the common set of 1,892 source pairs and their fusion results. The thresholds provide an empirical scale for evaluator disagreement in the descriptive analyses. Fusion behavior is categorized using the gain threshold:

$$\mathrm{Category}_t = \begin{cases} \text{beneficial}, & G_t > \epsilon_t, \\ \text{redundant}, & |G_t| \leq \epsilon_t, \\ \text{harmful}, & G_t < -\epsilon_t. \end{cases} \tag{15}$$

These thresholds are used for the benchmark analyses; the FuM-Score labels retain the original response scale and are computed directly from the averaged evaluator responses.

B.2 Source-referenced fusion behaviors

Our evaluation protocol uses source-image performance as a reference when an ideal fused reference image is unavailable, allowing us to examine whether fusion improves, preserves, or reduces task performance. We identify **beneficial**, **redundant**, and **harmful** outcomes from G_t, using uncertainty thresholds estimated from evaluator disagreement. Fused images with similar task utility can have different gains: some outperform their stronger source, while others fall below it. This difference reflects the varying performance of their source images, providing a basis for interpreting the three types of fusion behavior in FuM-18K.

Table 7: Distance to the ideal point for different values of η. Lower values indicate a better balance between gain-order agreement and evaluator stability.

η	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	1.0
$d(\eta) \downarrow$	1.000	0.745	0.588	0.513	**0.510**	0.561	0.639	0.820	1.000

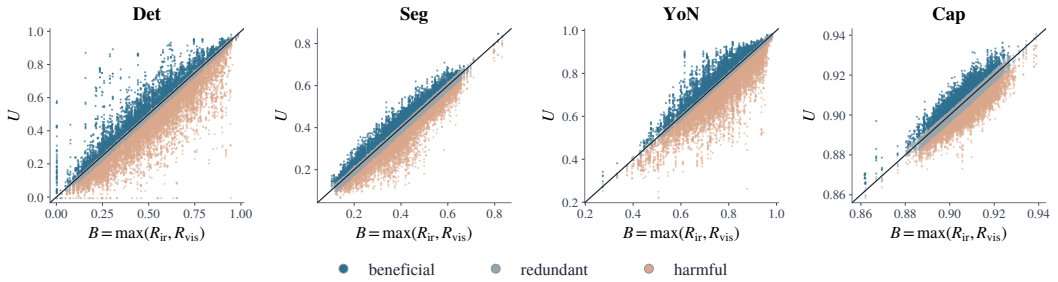

Figure 5: Fusion gains relative to the stronger source across four tasks. Fused-image utility U_t is compared with the stronger source response B_t, with the difference $G_t = U_t - B_t$ indicating fusion gain. Colors distinguish beneficial, redundant, and harmful fusion using task-specific thresholds estimated from evaluator disagreement. Fused images with similar utility can yield different gains depending on source performance.

B.3 Selection of η

The first criterion, $A_t(\eta)$, measures whether beneficial fusion receives higher quality scores than harmful fusion at similar utility levels. These images may achieve similar downstream performance, yet one improves upon its stronger source while the other falls below it. This criterion therefore assesses whether quality rankings reflect the improvement brought by fusion.

The second criterion, $S_t(\eta)$, measures ranking agreement across evaluators. It compares quality rankings constructed from two evaluators with those independently obtained from the held-out evaluator. Higher agreement indicates that the evaluation conclusions depend less on the particular evaluators used.

$$d(\eta) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[(1 - A_t(\eta))^2 + (1 - S_t(\eta))^2\right]}, \tag{16}$$

where $T = 4$, and $A_t(\eta)$ and $S_t(\eta)$ denote the gain-order agreement and held-out evaluator agreement for task t, respectively. As shown in Table 7, $\eta = 0.4$ achieves the smallest distance among the evaluated candidates and is used for all four tasks.

C Database Examples

Figure 6 presents 6 examples from FuM-18K across different scenes and illumination conditions. Each example includes an infrared-visible source pair and 3 selected fusion results, together with the source responses and task-specific U_t, G_t, and Q_t labels. These examples illustrate how fusion quality depends on both the source responses and the downstream task.

D Implementation Details

D.1 FuM-Net Implementation Details

Cross-Modal Interactive Encoder The infrared, visible, and fused images are processed by a shared frozen visual encoder. The trainable modules operate on 384-dimensional features. Follow-

Table 8: Overall and task-specific ablation results on the inputs of FuM-Net. The best and second-best results are **highlighted** and underlined, respectively.

Input			Det		Seg		VQA		Cap		Overall	
IR	VIS	F	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
–	–	✓	0.827	**0.835**	0.934	0.921	0.601	0.599	0.598	0.595	0.839	0.834
✓	–	✓	0.806	0.812	0.938	0.929	**0.638**	**0.638**	0.617	0.616	0.832	0.826
–	✓	✓	0.814	0.816	0.941	0.932	0.585	0.590	0.630	0.626	0.826	0.821
✓	✓	✓	**0.828**	0.834	**0.944**	**0.933**	0.620	0.632	**0.632**	**0.646**	**0.848**	**0.841**

Table 9: Overall and task-specific ablation results on the components of cross-modal interactive encoder.

Configuration	Det		Seg		VQA		Cap		Overall	
	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
w/ all	**0.828**	0.834	**0.944**	0.933	**0.620**	**0.632**	0.632	0.646	**0.848**	**0.841**
w/o M-Adapter	0.824	0.830	0.943	**0.934**	0.595	0.595	**0.667**	**0.654**	0.840	0.831
w/o Interactive Refinement	0.827	**0.835**	0.934	0.921	0.601	0.599	0.598	0.595	0.839	0.834
w/o SDM	0.820	0.824	0.937	0.923	0.606	0.604	0.631	0.628	0.835	0.825

ing the notation in the main text, $\mathbf{X}_m^{(l)}$ denotes the patch tokens extracted from layer l for modality $m \in \{\text{ir}, \text{vi}, f\}$. For each source $s \in \{\text{ir}, \text{vi}\}$, the modality adapter (M-Adapter) aggregates features from $L = \{2, 5, 8, 11\}$ through layer-specific projections:

$$\overline{\mathbf{X}}_s = \frac{1}{|L|} \sum_{l \in L} \mathcal{P}_s^{(l)} \left(\text{LN}(\mathbf{X}_s^{(l)}) \right). \tag{17}$$

The aggregated features are processed by a local adapter $\mathcal{A}_s$ and added to the final source features:

$$\mathbf{S}_s = \text{LN} \left(\mathbf{X}_s^{(last)} + \alpha_s \mathcal{A}_s(\overline{\mathbf{X}}_s) \right), \tag{18}$$

where $\mathcal{P}_s^{(l)}$ is the layer projection and α_s is a learnable scalar initialized to zero. The local adapter consists of a bottleneck projection, a 3×3 depthwise convolution, and an output projection. The two source adapters have separate parameters. The fused branch does not use M-Adapter, and its final patch tokens initialize refinement as $\mathbf{F}^{(0)} = \mathbf{X}_f^{(last)}$.

Task-Aware Quality Prediction. Each task uses a linear router to select two of the four shared MLP experts for each feature. The resulting task-specific feature sequence is summarized through attention pooling. A single learned query attends to the sequence through 6 heads cross-attention, with layer normalization applied to the query and the task-specific feature used as keys and values. The attention output is added to the query, followed by a feed-forward block with pre-layer normalization and a residual connection. A final layer normalization produces the pooled representation. The query and pooling parameters are shared across tasks, while each task has its own linear prediction head with a sigmoid output.

Training loss. The regression loss is Smooth L1 with $\beta = 1$, averaged over all samples and four task-specific FuM-Score labels in each batch. The load balancing term discourages the routers from repeatedly concentrating on a small subset of experts:

$$\mathcal{L}_{\text{bal}} = \frac{1}{|T|} \sum_{t \in T} \frac{K}{2} \sum_{k=1}^{K} P_{t,k} F_{t,k}, \tag{19}$$

where $K = 4$ is the number of shared experts, $P_{t,k}$ is the mean routing probability assigned to expert k before Top-2 selection, and $F_{t,k}$ is the fraction of features for which that expert is selected. Both quantities are averaged over the tokens in the training batch. The factor of two accounts for selecting two experts per token. We set $\lambda_{\text{bal}} = 10^{-3}$ in the total loss defined in the main text.

Table 10: Overall and task-specific ablation results on the task-aware quality prediction module.

Configuration	Det		Seg		VQA		Cap		Overall	
	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	SRCC ↑
Shared MLP	0.814	0.823	0.941	0.931	0.619	0.617	0.632	0.634	0.833	0.828
Task-Private MLP	0.810	0.817	0.940	0.930	0.576	0.578	**0.644**	**0.654**	0.826	0.827
Shared Experts Group + Shared Router	0.799	0.808	0.936	0.925	0.587	0.573	0.628	0.625	0.822	0.824
Shared Experts Group + Task-Specific Sub-Router	**0.828**	**0.834**	**0.944**	**0.933**	**0.620**	**0.632**	0.632	0.646	**0.848**	**0.841**

D.2 Experiment Implementation Details

Comparison with existing assessment methods. We compute the typical fusion metrics following the implementations used by Liu et al. (2024a). The NR-IQA and FR-IQA baselines are trained on FuM-18K using the FuM-Score labels, except for NIQE, DeepDC, and A-FINE. NIQE uses its original statistical model, DeepDC uses fixed pretrained features, and A-FINE retains its pretrained fidelity-naturalness calibration obtained through staged training. The reported FR-IQA results use the fused image as the assessed image and the infrared source as the reference. ResNet50 and ViT-S are trained from scratch with a shared backbone and four task-specific regression heads. Their NR variants receive only the fused image, while their FR variants encode the infrared, visible, and fused images separately using shared backbone weights. The three image representations are concatenated and projected into a shared feature space before task-specific prediction. ResNet50 uses global average pooling, while ViT-S uses the class token as the image representation. RA-MIQA retains its original architecture and is trained separately for each task, with both input views derived from the same fused image. Its primary encoder is initialized with ImageNet-pretrained weights, the pretrained SERE encoder remains frozen, and the scalar regressor is initialized randomly.

Ablation study. Tables 8, 9, and 10 supplement the ablations with results for all four tasks. In the input ablations, **F+IR** and **F+VIS** retain only the corresponding source branch, including its M-Adapter and source-specific operations in interactive refinement layers. The **F-only** configuration removes both source branches and passes the fused token sequence directly to the Quality Encoder; it is also reported as **w/o Interactive Refinement**. For the component ablations, **w/o M-Adapter** uses source patch tokens from the frozen visual encoder directly in interactive refinement layers. **w/o SDM** writes the retrieved source features back to the fused representation instead of constructing difference corrections. It retains token-wise scalar weights, computed from the current fused tokens and retrieved source features. For Task-aware quality prediction, complete model achieves the best Overall PLCC and SRCC and ranks first or second on every task-specific metric within each reported ablation group, supporting the effectiveness of the complete design in jointly predicting quality for different tasks.

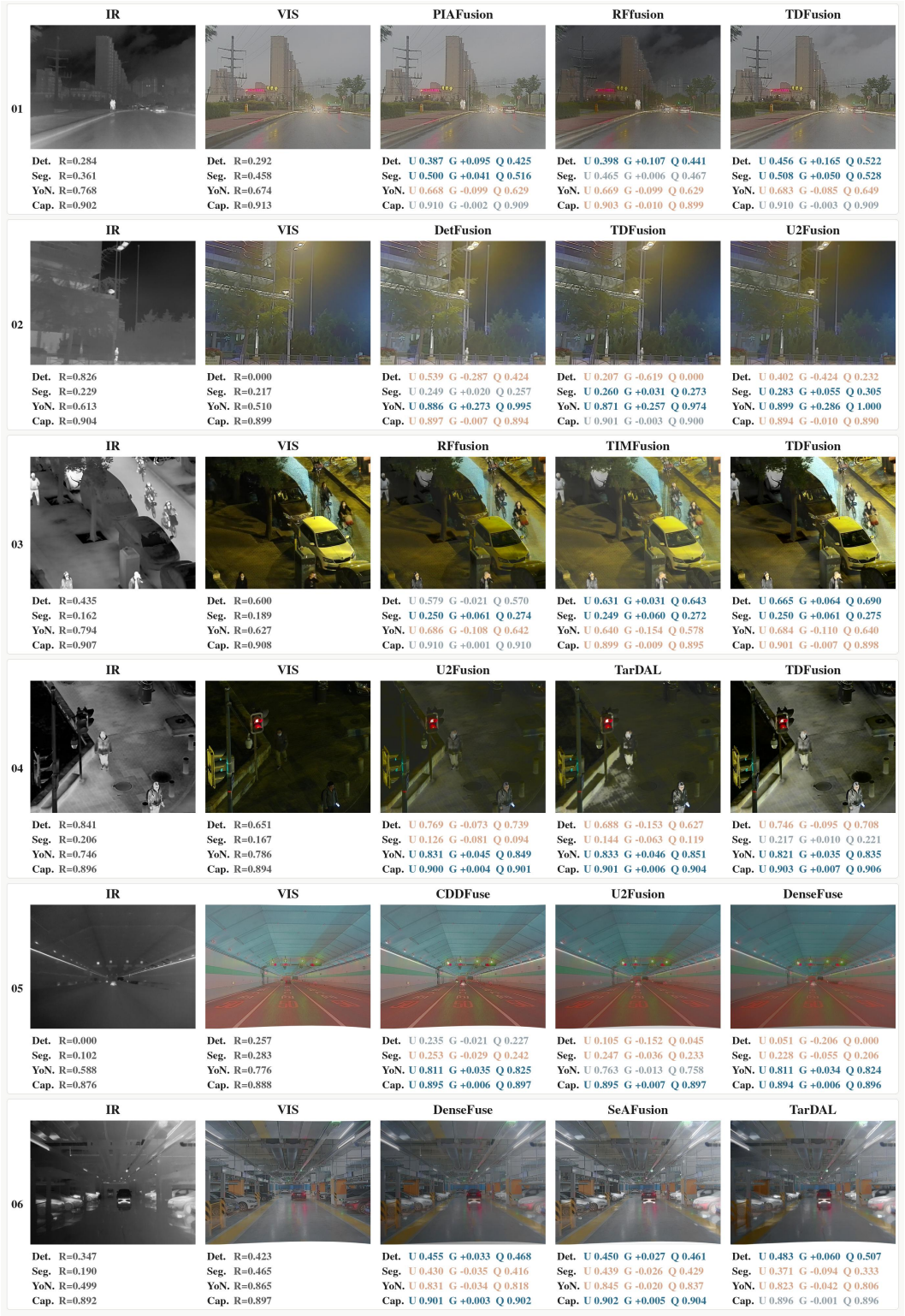

Figure 6: Examples from FuM-18K. Each row presents an infrared-visible source pair and 3 selected fusion results, with source responses and task-specific U_t, G_t, and Q_t labels shown below. Colored labels indicate beneficial, redundant, and harmful fusion according to the task-specific gain thresholds.

Prompt for Question-Answer Generation

```
You are an expert in infrared-visible image fusion quality assessment.

You will observe a local infrared image patch (Image 1) and its aligned local visible image patch
(Image 2). These two patches are cropped from the same spatial region of a registered infrared-
visible image pair.

Your task is to generate 2 Yes-or-No questions and their ground-truth answers.

The questions will later be asked about a single image, such as an infrared patch, a visible patch,
or a fused patch from the same spatial region. Therefore, each question must be answerable from
one image alone.

Important:
The final question text must NOT mention infrared, visible, thermal, heat, source image, Image 1,
Image 2, fusion, patch, crop, cropped region, local region, anchor, bounding box, mask, or center.

Core goal:
Generate questions that test whether the evaluated image preserves task-relevant semantic and
spatial information in this local scene area.

Answer balance rule:
You MUST generate exactly 2 questions:

* Question 1 must have answer "Yes".
* Question 2 must have answer "No".

Use the standardized object vocabulary only:
Car, Person, Truck, Bus, Motorcycle, Lamp, Road, Sidewalk, Building, Traffic Light, Traffic Sign,
Vegetation, Sky, Pole

Question design requirements:

1. The question must ask about observable semantic content.
2. The question must be specific, natural, and verifiable.
3. The question must be answerable from a single image alone.
4. The question must not rely on information outside the shown image area.
5. The question must not mention modality, sensor type, image quality, fusion, crop, patch, or
region construction.
6. The question should focus on one of the following reasoning types:

   * object existence
   * object count
   * object location
   * object-scene relation
   * object-object relation
   * observable state, such as parked Car or standing Person, only when the state is visually
evident
7. Prefer important and visually stable objects in the shown image area.
8. Prefer questions involving different object categories when possible.
9. Use exact counts only when the count is visually reliable.
10. Avoid tiny, heavily occluded, truncated, or ambiguous objects.
11. Avoid questions about readable text, fine symbols, color, facial details, license plates, or
unstable visual details.
12. Avoid over-mechanical phrasing such as “the object near the center” unless it is necessary to
remove ambiguity.
13. Do not force every question to ask about the same object. If multiple stable objects are
visible, use different objects or relations.

How to design the "Yes" question:

The "Yes" question should describe a true and non-trivial fact visible in the shown image area.
Prefer stable objects or simple spatial relations.

Good strategies:

* Ask whether a specific object category is present.
* Ask whether a specific object is located on or near a Road, Sidewalk, Building, Pole, or other
stable scene element.
* Ask whether two visible object categories have a clear spatial relationship.
* Ask a simple count question only when the count is reliable.
```

Prompt for Question-Answer Generation

```
Good strategies:
* Ask whether a specific object category is present.
* Ask whether a specific object is located on or near a Road, Sidewalk, Building, Pole, or other
stable scene element.
* Ask whether two visible object categories have a clear spatial relationship.
* Ask a simple count question only when the count is reliable.
How to design the "No" question:
The "No" question must be plausible but factually incorrect. It should not be absurd, trivially
false, or based on an object category unrelated to the shown image area.

Good strategies:

* Wrong object type: ask whether a visually similar category is present when it is not.
* Wrong count: ask whether there are two Cars when there is one Car.
* Wrong location: ask whether a Person is on the Sidewalk when the Person is actually on the Road.
* Wrong relation: ask whether a Truck is beside a Car when that relation is not visible.

Bad "No" question strategies:

* Do not ask about objects that are obviously unrelated to the scene.
* Do not use negative wording such as "Is there no Car...?"
* Do not ask about thermal state, image quality, fusion quality, or sensor-specific appearance.
* Do not create a false question by relying on details outside the shown image area.

Output JSON only.
Output format:

```json
{
  "questions": [
    {
      "question": "A specific Yes-or-No question whose answer is Yes.",
      "answer": "Yes",
      "question_type": "existence / count / location / object-scene-relation / object-object-
relation / state",
      "source_pair_evidence": "Briefly state the visual evidence from the two aligned source
patches that supports the answer.",
      "fusion_evaluable_reason": "Why this question can be answered from a single fused image of
the same area.",
      "risk_level": "Low / Medium / High",
      "task_difficulty": "Easy / Medium / Hard"
    },
    {
      "question": "A specific Yes-or-No question whose answer is No.",
      "answer": "No",
      "question_type": "existence / count / location / object-scene-relation / object-object-
relation / state",
      "source_pair_evidence": "Briefly state the visual evidence from the two aligned source
patches that makes the question false.",
      "fusion_evaluable_reason": "Why this question can be answered from a single fused image of
the same area.",
      "risk_level": "Low / Medium / High",
      "task_difficulty": "Easy / Medium / Hard"
    }
  ]
}
```

Good examples:

```json
{
  "question": "Is there a Person standing beside the Road?",
  "answer": "Yes",
  "question_type": "object-scene-relation",
  "source_pair_evidence": "The aligned source patches show a Person standing beside the Road.",
  "fusion_evaluable_reason": "The Person and Road-side relation should remain visible in a fused
image of the same area.",
  "risk_level": "Low",
  "task_difficulty": "Medium"
}
```
```

Prompt for Question-Answer Generation

````
```json
{
  "question": "Is there a Bus in the image?",
  "answer": "No",
  "question_type": "existence",
  "source_pair_evidence": "The aligned source patches show a Car, but no Bus is visible.",
  "fusion_evaluable_reason": "The presence or absence of a Bus can be judged from a single fused
image of the same area.",
  "risk_level": "Low",
  "task_difficulty": "Easy"
}
```

```json
{
  "question": "Are there two Cars visible on the Road?",
  "answer": "No",
  "question_type": "count",
  "source_pair_evidence": "The aligned source patches show one Car on the Road, not two.",
  "fusion_evaluable_reason": "The number of clearly visible Cars can be evaluated from a single
fused image of the same area.",
  "risk_level": "Medium",
  "task_difficulty": "Medium"
}
```


Bad examples:

* "Is the object near the center of this cropped region a Car?"
  Reason: Mentions crop construction and uses mechanical phrasing.

* "Is there a Person with elevated thermal signature on the Road?"
  Reason: Uses thermal-specific wording.

* "Is the Car clearly visible?"
  Reason: Uses subjective quality wording.

* "Does the visible image show a Car?"
  Reason: Refers to a source modality.

* "Is there no Car in the image?"
  Reason: Negative wording makes evaluation less stable.

Final reminders:

1. Generate exactly one Yes question and one No question.
2. The final question text must be modality-neutral.
3. The final question text must not mention patch, crop, region, center, anchor, mask, or bounding
box.
4. The final question must be answerable from a single image.
5. Prefer natural questions about stable local semantic content.
6. Use only standardized object vocabulary.
7. Do not ask about thermal cues, image quality, color fidelity, sensor type, or fusion.

Now analyze the provided aligned infrared and visible image patches and generate the 2 questions.
````

Figure 7: Template for generating reference question prompts

Prompt for Reference Caption Generation

```
You are an expert annotator for infrared-visible image fusion evaluation.
You will observe an infrared image (Image 1) and a visible image (Image 2). Your task is to write one English
reference caption describing the semantic scene content that should be recoverable from the source image pair.
This reference caption will later be compared with captions generated from single evaluated images, including
fused images. Therefore, the caption must be modality-neutral and must not prescribe how the fused image should
look.

Core goal:
Generate a factual reference caption for machine-centric semantic evaluation. The caption should describe objects,
scene layout, and spatial relationships, not visual aesthetics or modality-specific appearance.

Use the standardized object vocabulary only.

Critical rules:

1. Do not write "ideal fusion", "fused image should", "infrared image", "visible image", "Image 1", or "Image 2"
in the caption.
2. Do not mention thermal information, heat response, infrared appearance, visible-light appearance, body heat,
warm engine, or thermal signature.
3. Do not describe image quality, naturalness, sharpness, clarity, contrast quality, noise, or visual comfort.
4. Describe only semantic facts that are observable from the source pair.
5. Prefer stable semantic content: object categories, object counts, object locations, scene regions, and spatial
relations.
6. Use exact counts only when reliable. If the count is uncertain, omit the count instead of using vague
quantifiers.
7. Avoid fine-grained or unstable details such as readable text, small symbols, license plates, facial details,
clothing details, and color-specific descriptions.
8. Do not infer object states that are not visually evident.
9. Keep the wording consistent with the single-image caption prompt to reduce synonym noise in automatic caption
evaluation.

What to include:

- Scene type: urban Road, rural Road, parking area, indoor corridor, intersection, tunnel, or similar setting when
evident.
- Major semantic objects using standardized vocabulary.
- Object quantities when reliable.
- Spatial positions: left, right, center, upper, lower, foreground, background.
- Scene regions: on the Road, beside the Sidewalk, near the Building, under the Sky, near Vegetation, beside a
Pole, near a Traffic Sign or Traffic Light.
- Object relationships: beside, behind, in front of, adjacent to, between, closest to.

Style requirements:

- Continuous prose only.
- One paragraph.
- Direct factual tone.
- No bullet points.
- No uncertainty words such as might, maybe, possibly, seems, appears to.
- No vague quantifiers such as some, several, a few, many.

Output the caption text only. Do not output JSON.

Good examples:

"Urban Road scene with one Car in the center region and one Person near the left Sidewalk. Building facades line
the Background, with Pole structures along the Road edge. The main spatial layout includes the Road in the
foreground, Sidewalk areas near the Buildings, and open Sky above."

"Daytime parking area with two Cars and one Bus positioned on the Road surface. Vegetation borders the right side,
while Buildings occupy the Background under open Sky. The objects are arranged across the center and right regions
of the scene."

Bad examples:

- "An ideal fusion image should show both thermal and visible details with good quality."
Reason: Prescriptive and subjective.

- "Three Persons with elevated thermal signatures cross the Road."
Reason: Uses thermal-specific wording.

- "A clear daytime scene with several vehicles on the street."
Reason: Uses quality word, vague count, and non-standard vocabulary.

Now observe the provided infrared and visible images and generate the reference caption.
```

Figure 8: Template for generating reference caption prompts

Prompt for Image Captioning Evaluation

```
You are analyzing a single image for a machine-centric image understanding task.

Write one English caption describing the semantic content and spatial layout of the image.

The image may be an infrared image, a visible image, or an infrared-visible fused image. Do not identify the image
type. Do not mention infrared, visible, thermal, heat, sensor, camera, or fusion.

Core goal:
Describe what semantic information is observable in the image. Focus on objects, scene layout, and spatial
relationships.

Use the standardized object vocabulary only.

Critical rules:

1. Describe only what is visible in the image.
2. Do not infer information from common sense.
3. Do not mention image quality or visual style.
4. Do not mention modality-specific information.
5. Use exact counts only when the count is reliable. If the count is uncertain, omit the count instead of using
vague quantifiers.
6. Avoid unstable details such as readable text, small symbols, license plates, facial details, clothing details,
and color-specific descriptions.
7. Do not use uncertain words such as might, maybe, possibly, seems, appears to.
8. Do not use vague quantifiers such as some, several, a few, many.
9. If an object is too small or ambiguous, omit it.
10. Keep wording consistent with the reference caption style.

What to include:

- Scene type and environment when evident: urban Road, rural Road, parking area, indoor corridor, tunnel,
intersection, or similar setting.
- Major semantic objects using standardized vocabulary.
- Reliable object quantities.
- Spatial positions: left, right, center, upper, lower, foreground, background.
- Object-scene relations: on the Road, beside the Sidewalk, near a Building, near Vegetation, beside a Pole.
- Object-object relations: beside, behind, in front of, adjacent to, between.
- Observable states only when evident: parked Car, standing Person, walking Person, stationary Bus.

What not to include:

- Image quality terms: clear, blurry, sharp, high-resolution, noisy, good, poor.
- Modality terms: infrared, thermal, visible, heat, sensor, camera, fused image.
- Subjective terms: beautiful, natural, well-lit, high-quality.
- Non-standard object names: pedestrian, human, people, vehicle, tree, plant, wall, structure, lamp, street light,
street, pavement.
- Speculative details: readable text, emotion, intent, identity, exact speed, or object function not visible.

Style requirements:

- Continuous prose only.
- One paragraph.
- Direct factual tone.
- No bullet points.

Output the caption text only. Do not output JSON.

Good examples:

"Urban Road scene with one Car in the center region and one Person near the left Sidewalk. Buildings line the
Background, with Pole structures along the Road edge. The Road occupies the foreground and Sidewalk areas extend
beside the Buildings."

"Outdoor Road area with two Cars and one Bus positioned across the center and right regions. Vegetation borders
the right side, while Buildings occupy the Background under open Sky."

"Urban intersection with one Motorcycle on the right Road area and one Car near the center. A Traffic Light is
positioned above the Road near a Pole, with Buildings and Sidewalk regions in the Background."

Bad examples:

- "This infrared image shows thermal signatures of pedestrians."
  Reason: Mentions image type, thermal information, and non-standard vocabulary.

- "A clear image with several vehicles on the street."
  Reason: Uses quality word, vague count, and non-standard terms.

- "The fused image has good visible-light details and natural colors."
  Reason: Mentions fusion, modality-specific appearance, and subjective quality.

- "There appears to be a person near the building."
  Reason: Uses uncertainty wording and does not follow standardized capitalization.

Now observe the provided image and generate the caption.
```

Figure 9: Single-image evaluation prompt for image captioning.

Prompt for VQA Evaluation

```
You are evaluating a single image for a machine-centric image understanding task.

Answer the following Yes-or-No question using only the visual evidence in the image.

Question: {question}

Instructions:
1. Answer only "Yes" or "No".
2. Do not explain your reasoning.
3. Do not add any other words.
4. If the visual evidence is insufficient to verify the statement, answer "No".

Answer:
```

Figure 10: Single-image evaluation prompt for visual question answering.