Bridging Human Evaluation to Infrared and Visible Image Fusion

Supplementary Material

1. Human Feedback Dataset Construction

1.1. Data Cleaning Pipeline

To construct a human feedback IVIF dataset to support the modeling and evaluation of perceptual consistency, we collected 31,769 infrared-visible image pairs from 8 benchmark datasets (FMB [5], LLVIP [3], M^3FD [4], MFNet [2], RoadScene [6], SMOD [1], TNO , and VIFB). Since the dataset contained a large number of duplicate scenes, we used the CLIP model for data cleaning

Given that the CLIP model is primarily pre-trained on visible domain images, we employ the CLIP ViT-B/32 model to perform semantic analysis on visible images. We first extract CLIP feature vectors for all visible images and apply $L2$ normalization. Through extensive experiments, we determined an optimal cosine similarity threshold of 0.85, which effectively balances diversity preservation and redundancy removal—retaining 1,257 representative images. For each similar scene cluster identified through CLIP analysis, we implement a selection process that evaluates images based on multiple quality dimensions: visual quality metrics (sharpness and contrast), resolution, and information content (using file size as a proxy indicator). The selection process employs a weighted scoring system, with visual quality as the most important factor (50%), followed by resolution (30%), and information content contributing the remaining 20% weight. This approach ensures that the highest quality representative image is retained from each semantic cluster, ultimately filtering down to 900 representative visible-light images paired with their corresponding infrared images. Further expert screening was conducted on the CLIP-cleaned data, ultimately yielding 850 pairs of high-quality images.

1.2. Large Models and Human Feedback

We first invited four senior experts with over five years of research experience in infrared and visible image fusion (IVIF) to provide refined scores for 100 data triplets and annotate artifact regions using center coordinates and radii, yielding a high-quality seed dataset. Each triplet comprises a visible image, an infrared image, and a fused image, requiring approximately five minutes for annotation. The total annotation time per annotator was approximately 500 minutes (about 8 hours). The detailed dataset annotation guidelines are shown in Tab. 1.

Using the obtained seed dataset, we aligned the GPT-4o model (specifically version gpt-4o-2024-08-06 accessed via the OpenAI API) with expert preferences. This alignment is a prompt-driven process achieved through In-Context Learning. When constructing the input context, we embedded the evaluation guidelines from Tab. 1 as system instructions into each query to guide the model in understanding the specific meanings and scoring criteria for each evaluation dimension. During this alignment phase, GPT-4o receives visible, infrared, and fused image triplets as inputs to perform systematic multi-dimensional evaluations across thermal retention, texture preservation, artifacts, and sharpness. To ensure the model's outputs strictly adhere to human expert judgment, we iteratively optimized the system prompts and selected effective few-shot exemplars from the seed dataset. By minimizing the discrepancy between the model outputs and the expert labels in the seed dataset, we determined the optimal prompt configuration and exemplar combination. This discrepancy is measured by a composite error metric. The first component is the scoring discrepancy, derived by computing the normalized Mean Squared Error (MSE) between the model's predicted scores and ground-truth expert values across the four evaluation dimensions. The second component is the artifact heatmap discrepancy, obtained from the Binary Cross-Entropy (BCE) between the

```
{
  "scores": {
    "Thermal Retention": 4,
    "Texture Preservation": 3,
    "Artifacts": 2,
    "Sharpness": 3,
    "Overall Score": 3
  },
  "shapes": [
    {
      "label": "Artifacts",
      "points": [[390, 420], [430, 420]],
      "shape_type": "circle"
    },
    {
      "label": "Artifacts",
      "points": [[250, 170], [290, 170]],
      "shape_type": "circle"
    }
  ]
}
```

Listing 1. Example JSON output from GPT-4o

Table 1. Annotation Guidelines for Human Feedback Dataset

Category	Item	Description and Details
Evaluation Criteria (1-5 points)	**1. Thermal Retention**	Quantifies preservation accuracy of heat sources' intensity, distribution, and spatial positioning in fused images.
		• High-heat signatures (e.g., humans/vehicles) from infrared imagery • Structural thermal patterns unique to infrared imaging in low-light conditions (e.g., buildings) • Environmental texture retention
	2. Texture Preservation	Measures the fidelity of visible-light details and textures in fused results.
		• Surface/edge clarity of objects • Resistance to visible information occlusion by infrared data • Structural integrity of fine-grained patterns
	3. Artifacts	Identifies unnatural distortions including:
		• Bright spots • Ghost shadows • Physically implausible anomalies
	4. Sharpness	Evaluates overall visual coherence through:
		• Edge definition precision • Detail completeness • Perceptual clarity
Heatmap Annotation	**1. Artifact Marking**	**Label:** Artifacts
		Method: Encircle anomalous regions using circular annotations (center coordinates + radius)
		Note: For elongated artifacts, approximate coverage with minimally sized circles

generated artifact circular regions and the expert-annotated heatmaps.

The JSON format output by the GPT-4o model contains four fundamental quality scores and artifact heatmap annotations. Each artifact region is defined by two coordinate points: the first point specifies the circle center, and the second point lies on the circumference; together, they determine the precise location and extent of the artifact marking. The example JSON output is shown in Listing 1.

Then, we used the GPT-4o model to automatically annotate all 9,350 fused images. To ensure the accuracy of the automated annotations, we established a rigorous human quality control process. We invited 5 researchers with at least 3 years of research experience in the IVIF field. The review work was divided into two stages:

Sampling Inspection (15 hours). Each reviewing expert was randomly assigned 200 annotated images for detailed inspection, achieving a total review coverage rate of 10.7% (1,000/9,350). Experts needed to evaluate: (1) whether the scores conformed to the evaluation guideline standards; (2) whether the artifact region annotations were complete and accurate; (3) whether there were systematic biases or error patterns. Through this stage, we identified the following main issues: 8% of images had scoring deviations; 12% of images had incomplete artifact annotations, missing artifacts in certain regions.

Comprehensive Correction and Consistency Calibration (150 hours). Based on the problem patterns identified in the first stage, the review team conducted a systematic review of all 9,350 images. The specific work included: (1) Scoring standard calibration: for images with scoring deviations, experts re-evaluated the scores according to the evaluation guidelines to ensure scoring consistency; (2) Artifact annotation correction: for missed artifact regions, experts supplemented annotations, particularly targeting small artifacts and blurry edge regions, and removed false artifact annotations by comparing the infrared and visible source images; (3) Extreme case discussion: for images where the

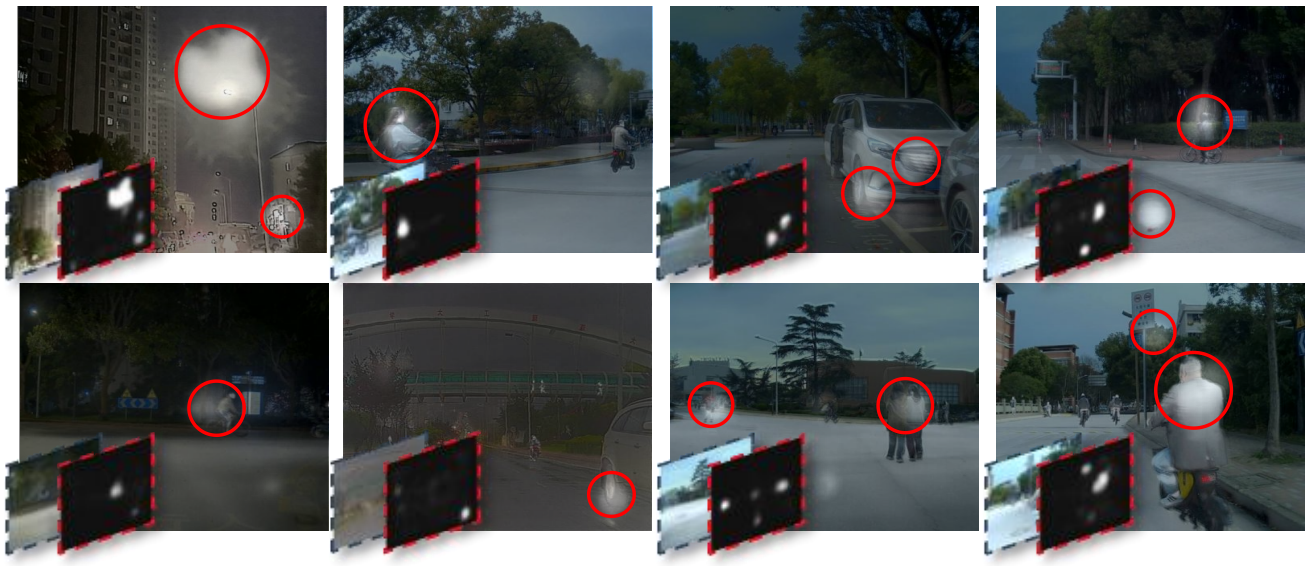

Figure 1. Heatmap visualization generated by the reward model on fused images, with artifact regions highlighted by red circles.

5 experts disagreed (3%), collective discussions were organized to reach consensus.

Through this rigorous quality control process, the annotation quality of the final dataset was significantly improved. The final human feedback IVIF dataset was used to train the fusion-oriented reward model.

2. Heatmap Visualization

To intuitively demonstrate the artifact identification capability of our trained fusion-oriented reward model, we conducted heatmap visualization analysis on images generated by different fusion methods, as shown in Fig. 1. The red circles highlight artifact regions identified by the model, including bright spots, ghosting, and structural distortions. This visualization validates the effectiveness of the reward model in quantifying human perceptual preferences, providing reliable guidance signals for subsequent RLHF optimization.

3. Policy Optimization Strategy Ablation

To validate the effectiveness of our proposed optimization strategy, we compare it with Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO).

For PPO, we implemented a value network as the critical model with a value loss coefficient of 0.5, using standard hyperparameters including a clipping parameter ϵ of 0.2 and GAE parameter λ of 0.95. The KL penalty coefficient β was set to 0.01, with early stopping when KL divergence exceeded 0.015 to prevent excessive policy deviation. For DPO, we followed the official formulation with $\beta = 0.1$, using the Bradley-Terry model for preference learning while keeping the reference model fixed during training, which is consistent with the standard DPO implementation.

References

[1] Zizhao Chen, Yeqiang Qian, Xiaoxiao Yang, Chunxiang Wang, and Ming Yang. Amfd: Distillation via adaptive multimodal fusion for multispectral pedestrian detection. *arXiv preprint arXiv:2405.12944*, 2024. 1

[2] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 1

[3] 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*, pages 3496–3504, 2021. 1

[4] 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*, pages 5802–5811, 2022. 1

[5] 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 the IEEE/CVF international conference on computer vision*, pages 8115–8124, 2023. 1

[6] 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*, 2020. 1