RETHINKING TASK-ORIENTED IMAGE FUSION AS PROBABILISTIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Although task-oriented image fusion has achieved promising downstream performance by incorporating task supervision, existing methods typically learn a direct mapping from multimodal inputs to a single fused image, with task objectives serving mainly as optimization signals. Such a formulation leaves the underlying conditional distribution of task-relevant fusion results unmodeled, making it difficult to capture the distributional structure and statistical regularities of multimodal integration for downstream tasks. To address this limitation, we rethink task-oriented image fusion as probabilistic inference, modeling the fused image as a conditional distribution jointly determined by source modalities and downstream task information. We introduce a latent fusion variable that captures both multimodal fusion cues and task-relevant semantics, which is sampled and decoded to generate fused images. During training, task annotations are used to construct a task-informed posterior, whose knowledge is transferred to a source-conditioned prior for label-free inference. We further align source modal and fused image task predictions to preserve task-relevant information. Experiments on infrared-visible fusion with semantic segmentation and object detection demonstrate consistent improvements in both fusion quality and downstream performance.

1 INTRODUCTION

Multimodal image fusion aims to integrate complementary information from multiple sensors to generate a more informative representation of the observed scene Ma et al. (2019). By exploiting the advantages of different imaging modalities, such as the thermal sensitivity of infrared images and the rich structural details of visible images, fusion techniques have been widely applied in vision scenarios, including surveillance, autonomous perception, and intelligent systems Zhang & Demiris (2023). With the rapid development of visual understanding, the role of image fusion has shifted from improving perceptual quality alone toward enhancing downstream vision tasks, where fused images are expected not only to preserve complementary multimodal information but also to facilitate subsequent semantic interpretation and decision-making Liu et al. (2024a).

Conventional image fusion methods, as shown in Figure. 1(a), mainly focus on integrating complementary information from source modalities to generate visually informative fused images, where the optimization objective is typically based on perceptual quality or information preservation. However, in practical perception systems, fused images often serve as intermediate representations for downstream analysis and decision-making. Therefore, recent studies have shifted towards task-oriented image fusion, as shown in Figure. 1(b), aiming to generate fused representations that better support downstream vision tasks Bai et al. (2025). Existing task-oriented fusion methods can be categorized into three groups based on how task information is incorporated. The first group introduces task objectives as additional optimization constraints by jointly optimizing task-specific and fusion losses Liu et al. (2022; 2023b). The second group designs task-fusion interaction architectures, where downstream networks provide semantic guidance during fusion feature extraction and reconstruction Yang et al. (2026). The third group combines both loss-level supervision and architecture-level interaction for joint optimization. Despite their effectiveness, existing methods still have limitations. Specifically, task information is usually introduced as optimization guidance after the fusion process is established, rather than as a prior condition for fusion generation. Moreover, existing frameworks lack the ability to model task-dependent fusion diversity and often require task-specific optimization, limiting their adaptability to different downstream requirements.

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.

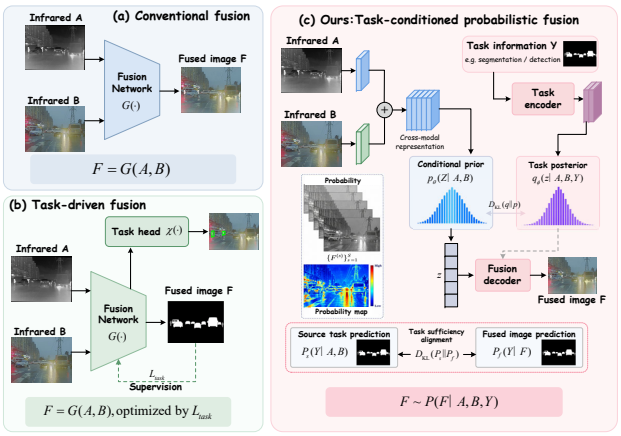


Figure 1: Illustration of conventional fusion, task-driven fusion, and the proposed task-conditioned probabilistic fusion framework.

To address these limitations, we rethink task-oriented image fusion from a probabilistic inference perspective and formulate the fused image as a task-conditioned distribution jointly determined by source modalities and downstream task information. Specifically, as shown in Figure. 1(c), we introduce a latent fusion variable to capture the hidden fusion strategies underlying multi-modal information fusion and task requirements, where the fused image is generated by sampling and decoding the latent representation. During training, task annotations are utilized to construct a task-informed posterior distribution, which learns task-aware latent representations from both source modalities and downstream objectives. Since task annotations are unavailable during inference, we further learn a source-conditioned prior distribution to approximate the posterior through variational inference, enabling label-free task-aware fusion. Furthermore, we introduce a task consistency constraint by aligning task predictions between source modalities and fused images, encouraging the preservation of task-relevant information during fusion. The main contributions of this work are summarized as follows:

- We propose a probabilistic formulation for task-oriented image fusion by modeling fused images as a task-conditioned distribution rather than a deterministic output, providing a new perspective for understanding how downstream task requirements influence fusion generation.
- We introduce a latent fusion variable and develop a posterior-prior variational inference framework, where task knowledge learned from annotations is transferred to a source-conditioned prior for label-free task-aware fusion.
- We propose a task consistency constraint by aligning task predictions between source modalities and fused images, which encourages the preservation of task-relevant information during the fusion process.

2 Related Work

With the advancement of deep learning, image fusion has evolved from handcrafted strategies to representation learning Liu et al. (2017). Early CNN-based methods extracted discriminative features for fusion reconstruction and achieved improved information preservation Li et al. (2020); Xu et al. (2020). Subsequently, Transformer-based approaches leveraged self-attention to model long-range dependencies and cross-modal interactions Tang et al. (2023b); Huang et al. (2025), while recent state space models and diffusion-based methods further improved global feature modeling and generative fusion capabilities Gu et al. (2022); Yu & Wang (2025); Xie et al. (2024); Zhu et al. (2025); Zhao et al. (2023); Tang et al. (2024). Despite these advances, most existing methods still formulate fusion as $\mathbf{F} = f_\theta(\mathbf{A}, \mathbf{B})$ or $P(\mathbf{F}|\mathbf{A}, \mathbf{B})$, focusing mainly on visual information integration. Consequently, downstream tasks are typically used only to evaluate the generated fused images, rather than being explicitly considered during the fusion generation process Liu et al. (2024b).

To enhance the applicability of image fusion methods in real-world scenarios, recent studies have shifted their attention toward incorporating downstream tasks into the fusion framework Guo et al. (2025); Shi et al. (2025). Early efforts mainly adopted a simple yet effective strategy by introducing downstream task losses as additional constraints and jointly optimizing them with fusion objectives to guide the fusion process Sun et al. (2022); Tang et al. (2023a). Beyond loss-based optimization, another line of research integrates downstream task networks with fusion networks into a unified framework. Through information exchange between the two networks, these methods provide se-

mantic guidance to the fusion process and encourage the fusion network to capture task-relevant information Liu et al. (2024c); Zhu et al. (2024); Wu et al. (2025). Despite these efforts, existing task-oriented fusion methods still treat downstream tasks mainly as external guidance rather than intrinsic factors involved in fusion generation. As a result, they cannot explicitly characterize the relationship between task requirements and fusion outcomes, limiting their ability to generate task-adaptive fused images Liu et al. (2024b); Yang et al. (2025).

3 METHODOLOGY

3.1 BAYESIAN FORMULATION OF TASK-ORIENTED IMAGE FUSION

Existing fusion methods usually learn a mapping $\mathbf{F} = f_\theta(\mathbf{A}, \mathbf{B})$, where $\mathbf{A}$ and $\mathbf{B}$ denote source modalities and $\mathbf{F}$ represents the fused result. However, in task-oriented fusion, the desired fused representation is also affected by downstream task requirements. Therefore, a more flexible formulation is to model the conditional distribution $P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$, which aims to generate task-aware fusion results based on both source modalities and task objectives. To explain this formulation, we consider the complete process from source modalities to fused representation and task prediction, which can be modeled by the joint distribution $P(\mathbf{F}, \mathbf{Y}|\mathbf{A}, \mathbf{B})$. According to the probability chain rule, this joint distribution can be factorized into two equivalent forms. The first formulation:

$$P(\mathbf{F}, \mathbf{Y}|\mathbf{A}, \mathbf{B}) = P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(\mathbf{Y}|\mathbf{F}, \mathbf{A}, \mathbf{B}), \tag{1}$$

describes a forward generation process, where the fused representation $\mathbf{F}$ is first generated from the source modalities $(\mathbf{A}, \mathbf{B})$, and then the downstream prediction $\mathbf{Y}$ is obtained based on the fused representation. In contrast, the second formulation:

$$P(\mathbf{F}, \mathbf{Y}|\mathbf{A}, \mathbf{B}) = P(\mathbf{Y}|\mathbf{A}, \mathbf{B})P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}), \tag{2}$$

describes a task-guided inference process. Instead of generating the fusion result solely from the source modalities, it first considers the task condition $\mathbf{Y}$. The two factorizations of the joint distribution are equivalent:

$$P(\mathbf{Y}|\mathbf{A}, \mathbf{B})P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(\mathbf{Y}|\mathbf{F}, \mathbf{A}, \mathbf{B}). \tag{3}$$

Therefore,

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = \frac{P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(\mathbf{Y}|\mathbf{F}, \mathbf{A}, \mathbf{B})}{P(\mathbf{Y}|\mathbf{A}, \mathbf{B})}. \tag{4}$$

For fixed source modalities and task conditions, the denominator $P(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ is independent of the fusion representation $\mathbf{F}$. Moreover, we assume that the fused representation $\mathbf{F}$ preserves sufficient information from the source modalities for the downstream task. Therefore, the task-conditioned fusion distribution can be simplified as:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) \propto P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(\mathbf{Y}|\mathbf{F}). \tag{5}$$

This formulation provides the probabilistic foundation of task-oriented fusion, which can be interpreted as the combination of a fusion prior and a task likelihood. Specifically, $P(\mathbf{F}|\mathbf{A}, \mathbf{B})$ serves as the fusion prior, which preserves the complementary information from the source modalities. Meanwhile, $P(\mathbf{Y}|\mathbf{F})$ acts as the task likelihood, which encourages the fused representation to contain information beneficial for the downstream task.

3.2 LATENT VARIABLE BASED FUSION GENERATION

Having established the desired task-conditioned fusion distribution $P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$, the next question is how to effectively model and infer this distribution. Directly modeling the distribution of the high-dimensional fused representation $\mathbf{F}$ conditioned on the source modalities and the downstream task is challenging. To provide a structured representation of the fusion process, we introduce a latent variable $\mathbf{Z}$, which captures the underlying task-conditioned fusion strategy. By marginalizing over the latent variable Z, the target distribution can be written as:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = \int P(\mathbf{F}, \mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})\, d\mathbf{Z}. \tag{6}$$

According to the chain rule of probability, we have

$$P(\mathbf{F}, \mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = P(\mathbf{F}|\mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{Y})P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y}). \tag{7}$$

Therefore, Eq. (6) can be rewritten as:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = \int P(\mathbf{F}|\mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{Y})P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})\, d\mathbf{Z}. \tag{8}$$

We further assume that $\mathbf{Z}$ captures the task-relevant information from the source modalities required for generating the fused representation. In other words, once the latent fusion strategy $\mathbf{Z}$ is determined, the generation of $\mathbf{F}$ does not require direct access to the source modalities or the downstream task condition. This can be formulated as $\mathbf{F} \perp (\mathbf{A}, \mathbf{B}, \mathbf{Y}) \mid \mathbf{Z}$.

Accordingly, we obtain:

$$P(\mathbf{F}|\mathbf{Z}, \mathbf{A}, \mathbf{B}, \mathbf{Y}) = P(\mathbf{F}|\mathbf{Z}). \tag{9}$$

The target task-conditioned fusion distribution can therefore be expressed as:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = \int P(\mathbf{F}|\mathbf{Z})P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})\, d\mathbf{Z}. \tag{10}$$

This formulation decomposes the task-oriented fusion process into two stages. The first stage infers a latent fusion strategy from the source modalities and the downstream task $P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$, while the second stage generates the fused representation conditioned on the inferred latent variable $P(\mathbf{F}|\mathbf{Z})$. Thus, the overall fusion process can be interpreted as $(\mathbf{A}, \mathbf{B}, \mathbf{Y}) \rightarrow \mathbf{Z} \rightarrow \mathbf{F}$.

3.3 Variational Inference for Task-Conditioned Latent Fusion

Based on Eq. (10), we formulate a new objective for solving the task-conditioned fusion problem. Therefore, obtaining the desired fusion result requires inferring the task-conditioned latent posterior $P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$. However, this posterior is generally intractable. Since $P(\mathbf{Z}$ is an unobserved latent variable, its posterior cannot be obtained directly from the training data. Exact inference further requires marginalizing over the latent space to compute the posterior normalization, which is generally infeasible for a high-dimensional continuous $\mathbf{Z}$. To address this issue, we introduce a variational distribution $q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$ to approximate the true posterior:

$$q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) \approx P(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y}). \tag{11}$$

Specifically, the variational posterior is modeled as a Gaussian distribution:

$$q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\mu}_q, \mathrm{diag}(\boldsymbol{\sigma}_q^2)), \tag{12}$$

where $\boldsymbol{\mu}_q$ and $\boldsymbol{\sigma}_q^2$ are the mean and variance predicted by the task-informed posterior network. During training, the task condition $\mathbf{Y}$ provides explicit guidance for learning task-relevant fusion strategies. Therefore, the variational posterior is conditioned on $\mathbf{Y}$ to learn task-aware latent representations. However, $\mathbf{Y}$ is unavailable during inference, making the task-informed posterior unsuitable for test-time inference. To address this issue, we introduce a source-conditioned prior distribution $p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})$, which infers latent fusion representations using only the available source modalities. Specifically, the prior distribution is also parameterized as a Gaussian distribution:

$$p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}) = \mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p^2)), \tag{13}$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\sigma}_p^2$ are predicted by the source-conditioned prior network. After obtaining the mean and variance parameters, the latent variable $\mathbf{Z}$ is sampled using the reparameterization trick:

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{14}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the predicted mean and standard deviation of the latent distribution.

As discussed above, we introduce a task posterior $q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, \mathbf{Y})$ during training and a source-conditioned prior $p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})$ for inference. The key to bridging the training and inference stages is to encourage the source-conditioned prior to capture the task-relevant information contained in the task-informed posterior. We achieve this through variational inference by minimizing the divergence between the two distributions. Specifically, we define the source-conditioned generative model as:

$$p_\theta(\mathbf{F}, \mathbf{Z}|\mathbf{A}, \mathbf{B}) = p(\mathbf{F}|\mathbf{Z})p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}). \tag{15}$$

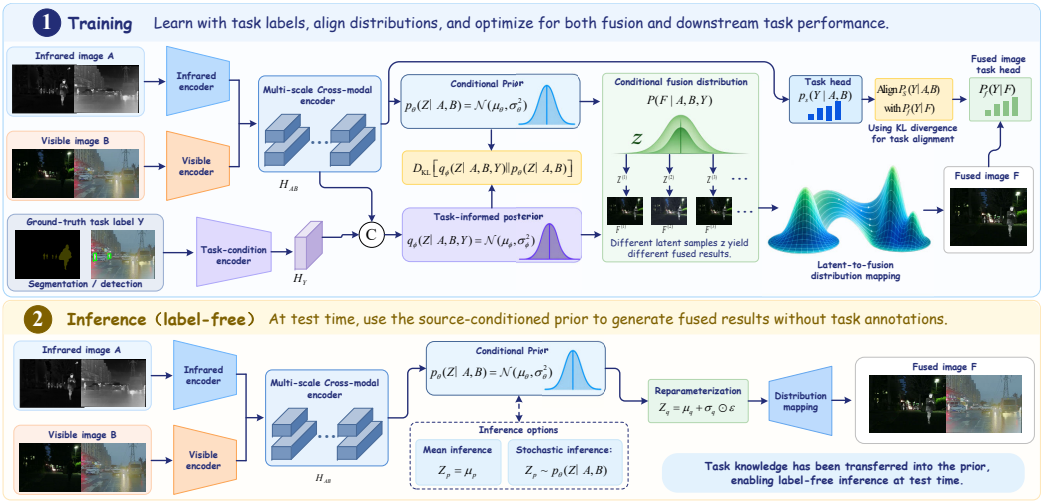


Figure 2: Overview of the proposed framework.

The corresponding marginal likelihood is

$$p_\theta(\mathbf{F}|\mathbf{A},\mathbf{B}) = \int p(\mathbf{F}|\mathbf{Z})p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})\, d\mathbf{Z}. \tag{16}$$

By introducing the task-informed variational posterior, we rewrite the marginal likelihood as:

$$\log p_\theta(\mathbf{F}|\mathbf{A},\mathbf{B}) = \log \int q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y}) \frac{p(\mathbf{F}|\mathbf{Z})p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})}\, d\mathbf{Z}. \tag{17}$$

By Jensen's inequality, we obtain:

$$\begin{aligned} \log p_\theta(\mathbf{F}|\mathbf{A},\mathbf{B}) &\geq \mathbb{E}_{q_\phi}\left[\log \frac{p(\mathbf{F}|\mathbf{Z})p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})}\right] \\ &= \mathbb{E}_{q_\phi}\left[\log p(\mathbf{F}|\mathbf{Z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})\|p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})\right). \end{aligned} \tag{18}$$

Therefore, the variational lower bound is

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})}\left[\log p(\mathbf{F}|\mathbf{Z})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})\|p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})\right). \tag{19}$$

Accordingly, we minimize the negative ELBO:

$$\mathcal{L} = -\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})}\left[\log p(\mathbf{F}|\mathbf{Z})\right] + D_{\mathrm{KL}}\left(q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})\|p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})\right). \tag{20}$$

The above objective consists of two complementary terms. The first term $-\mathbb{E}_{q_\phi}\left[\log p(\mathbf{F}|\mathbf{Z})\right]$ encourages the latent representation $\mathbf{Z}$ to preserve sufficient information for generating the desired fused representation $\mathbf{F}$. The second term $D_{\mathrm{KL}}\left(q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})\|p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})\right)$, encourages the prior to approximate the task posterior. In this way, task-relevant information learned from $\mathbf{Y}$ during training can be transferred to the source-conditioned prior.

3.4 Latent-to-Fusion Generation

Based on the latent variable formulation, the fused representation is generated from the latent fusion variable through the conditional distribution $P(\mathbf{F}|\mathbf{Z})$. Here,$\mathbf{Z}$ serves as a compact representation of the fusion strategy, which encodes the source information and task-related knowledge learned during training. Different latent variables correspond to different possible fusion results, indicating that the latent space captures the uncertainty and diversity of task-oriented fusion. Specifically, the latent variable $\mathbf{Z}$ is mapped to the fused representation through a conditional distribution $p_\theta(\mathbf{F}|\mathbf{Z})$, which characterizes the stochastic relationship between the latent fusion strategy and the generated fused image. In practice, this distribution is parameterized by a neural network that transforms $\mathbf{Z}$ into the fused image $\mathbf{F}$.

As illustrated in Figure. 2, the proposed framework consists of a training stage and an inference stage. During training, the source modalities are first encoded by modality-specific encoders and a multi-scale cross-modal encoder to obtain the source feature representation. Together with the task condition $\mathbf{Y}$, the task-condition encoder infers the task-informed latent distribution $q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},\mathbf{Y})$, while the conditional prior estimates $p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})$ from the source modalities alone. The latent variable $\mathbf{Z}$ is then sampled and transformed into the fused image through the latent-to-fusion mapping. During inference, since the task condition $\mathbf{Y}$ is unavailable, only the source-conditioned prior $p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B})$ is used to generate the latent representation, which is subsequently mapped to the final fused image. Through the distribution alignment between the posterior and prior, the task knowledge learned during training is transferred to the inference stage without requiring task annotations.

3.5 Task Consistency Constraint

To further preserve task-relevant information during the fusion process, we introduce a task consistency constraint. Specifically, let $T(\cdot)$ denote the downstream task network. Given the source modalities, the task prediction distribution is defined as $P_s(\mathbf{Y}|\mathbf{A},\mathbf{B}) = T(\mathbf{A},\mathbf{B})$, while the prediction distribution obtained from the fused image $\mathbf{F}$ is given by $P_f(\mathbf{Y}|\mathbf{F}) = T(\mathbf{F})$. Since the fused image is expected to retain the task-relevant information contained in the source modalities, we encourage the task prediction distributions of the source modalities and the fused image to remain consistent. To this end, we measure their discrepancy using the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\mathrm{tc}} = D_{\mathrm{KL}}\left(P_s(\mathbf{Y}|\mathbf{A},\mathbf{B}) \,\|\, P_f(\mathbf{Y}|\mathbf{F})\right). \tag{21}$$

Minimizing $\mathcal{L}_{\mathrm{tc}}$ encourages the fused image to preserve task-relevant information from the source modalities, thereby improving its suitability for downstream vision tasks.

3.6 Overall Training Objective

To jointly preserve source information and task-relevant semantics, we optimize the proposed framework with a unified objective consisting of fusion fidelity, task supervision, latent distribution alignment, and task distribution consistency. The overall loss is defined as:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{fus}}\mathcal{L}_{\mathrm{fus}} + \lambda_s\mathcal{L}^s_{\mathrm{task}} + \lambda_f\mathcal{L}^f_{\mathrm{task}} + \beta\mathcal{L}_{\mathrm{prior}} + \delta\mathcal{L}_{\mathrm{cons}}, \tag{22}$$

where λ_{fus}, λ_s, λ_f, β, and δ are weighting coefficients. The fusion loss $\mathcal{L}_{\mathrm{fus}}$ constrains the fused image $\mathbf{F}$ to preserve complementary visual information from the source images $\mathbf{A}$ and $\mathbf{B}$. It can be optimize using commonly adopted intensity and gradient constraints:

$$\mathcal{L}_{\mathrm{fus}} = \lambda_{\mathrm{int}}\mathcal{L}_{\mathrm{int}} + \lambda_{\mathrm{grad}}\mathcal{L}_{\mathrm{grad}}. \tag{23}$$

To explicitly incorporate downstream semantics, we introduce two task prediction pathways. The source pathway estimates $P_s(\mathbf{Y} \mid \mathbf{A},\mathbf{B})$, while the fusion pathway predicts $P_f(\mathbf{Y} \mid \mathbf{F})$. Their corresponding supervised losses are denoted by $\mathcal{L}^s_{\mathrm{task}}$ and $\mathcal{L}^f_{\mathrm{task}}$, respectively. These losses are defined according to the corresponding downstream task objectives, where the same task-specific loss functions are applied to the predictions obtained from the source modalities and the fused image. To bridge the discrepancy between training and inference, we align the task posterior with the source-conditioned prior:

$$\mathcal{L}_{\mathrm{prior}} = D_{\mathrm{KL}}\left[q_\phi(\mathbf{Z} \mid \mathbf{A},\mathbf{B},\mathbf{Y}) \,\|\, p_\theta(\mathbf{Z} \mid \mathbf{A},\mathbf{B})\right]. \tag{24}$$

This term transfers the task-aware fusion knowledge encoded by the posterior, which has access to $\mathbf{Y}$ during training, into the prior that depends only on $\mathbf{A}$ and $\mathbf{B}$, enabling task-aware fusion without task annotations at inference time. Finally, we introduce a task distribution consistency loss to explicitly preserve the task information available in the original modalities:

$$\mathcal{L}_{\mathrm{cons}} = D_{\mathrm{KL}}\left[\mathrm{sg}(P_s(\mathbf{Y} \mid \mathbf{A},\mathbf{B})) \,\|\, P_f(\mathbf{Y} \mid \mathbf{F})\right], \tag{25}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. By encouraging the task distribution inferred from the fused image to approach that inferred from the complete source modalities, this constraint reduces the loss of task-relevant information during fusion. Together, these objectives enable the model to learn fused representations that are both visually informative and explicitly aligned with downstream task requirements.

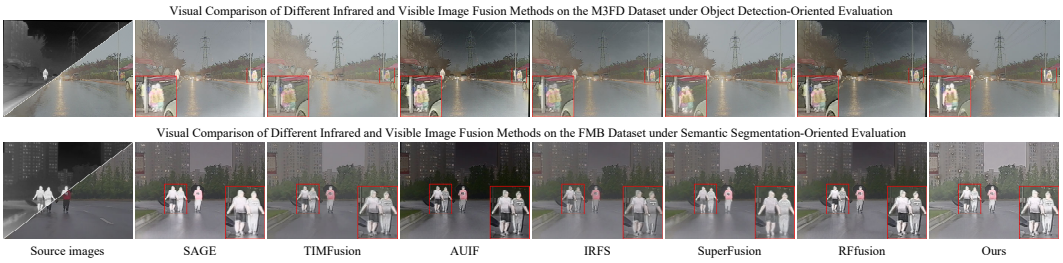


Figure 3: Qualitative comparison of fusion results on the M3FD and FMB datasets. Red boxes indicate local regions for comparison.

Table 1: Quantitative comparison of different fusion methods on the M3FD and FMB datasets. The best and second-best results are highlighted in bold and underline, respectively.

	M3FD							FMB						
Metric	SAGE	TIMFusion	AUIF	IRFS	SuperFusion	RFusion	Ours	SAGE	TIMFusion	AUIF	IRFS	SuperFusion	RFusion	Ours
AG↑	4.4778	3.5873	**5.0220**	3.6435	3.5235	<u>4.5475</u>	3.8341	3.7957	3.0558	3.9255	2.9952	3.0146	3.3243	**4.1152**
EI↑	46.5637	37.3941	**52.4367**	38.1473	36.4635	<u>47.3173</u>	40.1550	39.6008	31.8751	<u>41.1227</u>	31.3413	31.3854	34.7222	**43.3180**
MI↑	2.1834	<u>2.5942</u>	2.0453	1.9552	2.2090	2.2664	**2.6855**	<u>2.3767</u>	**2.4035**	2.2339	2.1391	2.3289	2.3189	2.2051
CE↓	1.7679	2.9447	1.2944	1.8895	1.6782	<u>1.3416</u>	**1.2400**	2.0515	3.8021	<u>1.4888</u>	2.1143	1.7012	**1.3957**	2.2161
SF↑	13.5204	11.0189	**14.9348**	11.2892	10.6189	<u>13.8093</u>	12.2266	12.6344	10.3252	**13.3655**	10.3829	10.2267	11.4861	<u>13.2495</u>
SD↑	9.2503	<u>9.6207</u>	9.3051	8.9058	9.2664	9.1362	**9.9098**	9.0431	8.8203	8.5913	8.3451	<u>9.1545</u>	8.3972	**9.2669**

4 EXPERIMENTS

4.1 DATASETS

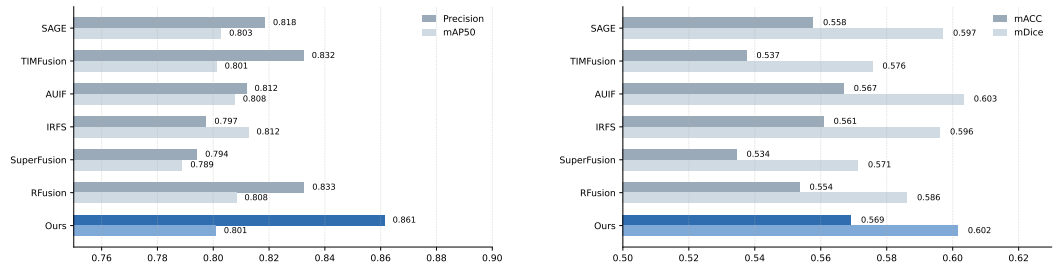
We conduct experiments on two widely used infrared and visible image fusion benchmarks, M3FD Liu et al. (2022) and FMB Liu et al. (2023a), to evaluate the proposed task-conditioned probabilistic fusion framework. M3FD provides paired infrared-visible images with object detection annotations for evaluating task-oriented fusion, while FMB contains aligned image pairs with pixel-level segmentation annotations for semantic segmentation evaluation. During training, task annotations are utilized as task conditions, whereas only source modalities are required during inference.

4.2 IMPLEMENTATION DETAILS

All experiments are implemented in PyTorch and trained in an end-to-end manner. The input images are resized to 768×1024 for M3FD, with corresponding resolutions adopted for FMB. The proposed framework is optimized using AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The batch size is set to 1 due to the high-resolution inputs. The base feature and latent channel dimensions are both set to 32. The training objective consists of fusion reconstruction, latent distribution alignment, and task consistency losses, with weights $\lambda_{\text{fus}} = 1$, $\lambda_{\text{s}} = 1$, $\lambda_{\text{f}} = 1$, $\lambda_{\text{prior}} = 0.01$, and $\lambda_{\text{cons}} = 1$. The latent KL and task consistency losses are gradually introduced using a 10-epoch warm-up strategy for stable optimization.

4.3 COMPARISON METHODS AND EVALUATION METRICS

We compare our method with six representative infrared and visible image fusion approaches, including SAGE , TIMFusion , AUIF , IRFS , SuperFusion , and RFfusion . These methods cover recent structure-aware, attention-based, and learning-based fusion strategies. For quantitative evaluation, we adopt six widely used metrics, including average gradient (AG), edge intensity (EI), mutual information (MI), cross-entropy (CE), spatial frequency (SF), and standard deviation (SD). AG and EI evaluate structural and edge preservation, MI measures the transferred information from source modalities, CE reflects distribution consistency, SF assesses spatial detail richness, and SD evaluates contrast variation of fused images.



(a) Object detection performance on M3FD. (b) Semantic segmentation performance on FMB.

Figure 4: Downstream task performance evaluation of fused images generated by different infrared and visible image fusion methods.



(a) Downstream object detection performance on M3FD (b) Downstream semantic segmentation performance on FMB

Figure 5: Visualization of downstream task performance using fused images.

4.4 Task-Oriented Fusion Evaluation

Figure. 3 presents qualitative comparisons between our method and existing image fusion approaches on the M3FD and FMB datasets. Different methods show varying abilities in preserving complementary information from two modalities. Compared with other approaches, our method achieves more balanced fusion results by maintaining both infrared thermal responses and visible texture details. In the highlighted pedestrian regions, our method preserves clearer target brightness and richer structural details, while methods such as AUIF and RFfusion tend to produce darker results with weaker thermal contrast. These results demonstrate that the proposed task-conditioned fusion framework effectively preserves task-relevant information for downstream vision tasks.

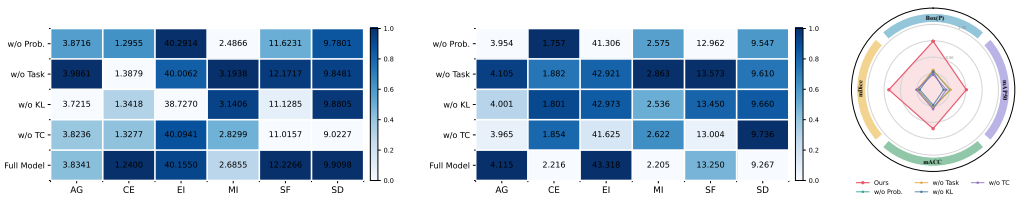
Table 1 presents the quantitative comparison of different fusion methods on the M3FD and FMB datasets. The proposed method achieves competitive performance on both datasets, demonstrating its effectiveness in preserving complementary information from infrared and visible modalities. On the M3FD dataset, our method obtains the best results in MI, CE, and SD, indicating richer information content and improved contrast preservation. On the FMB dataset, it achieves the highest AG, EI, and SD, showing better capability in maintaining structural details. These results verify that the proposed task-conditioned fusion strategy can effectively preserve task-relevant information under different downstream scenarios.

4.5 Downstream Task Performance Evaluation

To further evaluate the effectiveness of the proposed task-conditioned fusion framework, we conduct downstream object detection and semantic segmentation experiments on the M3FD and FMB datasets. As shown in Figure. 4, our method achieves competitive performance on both tasks, obtaining the highest Precision on M3FD detection and the best mACC on FMB segmentation. The qualitative results in Figure. 5 further demonstrate that the generated fused images preserve task-relevant information, enabling more accurate downstream predictions.

4.6 Ablation Study

We conduct ablation studies to evaluate the effectiveness of each component in the proposed framework. As shown in Fig. 6(a)-(b), removing any component results in performance degradation on both fusion quality and downstream evaluation, demonstrating the contribution of each proposed design. Specifically, w/o Prob. removes the probabilistic modeling module, leading to degraded perfor-



(a) (b) (c)

Figure 6: Ablation study of the proposed method. (a)-(b) Fusion quality evaluation on M3FD and FMB datasets, respectively. (c) Downstream task performance evaluation. w/o Prob., w/o Task, w/o KL, and w/o TC indicate removing probabilistic modeling, task guidance, prior-posterior KL constraint, and task consistency constraint, respectively.

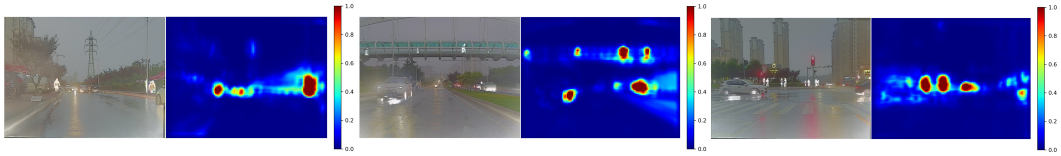


Figure 7: Visualization of latent uncertainty maps generated by the proposed framework.

mance and verifying the importance of learning uncertainty-aware latent representations. w/o Task eliminates the task condition $\mathbf{Y}$, which demonstrates the necessity of task-guided fusion optimization for generating task-relevant representations. w/o KL removes the prior-posterior distribution alignment constraint, indicating that the latent distribution regularization is beneficial for learning stable fusion representations. w/o TC discards the task consistency constraint between source and fused images, resulting in inferior downstream performance and validating its role in preserving task-related information. Furthermore, Fig. 6(c) shows that the full model consistently achieves better detection and segmentation performance, confirming that the proposed components complement each other and jointly enhance both fusion quality and task-oriented representation capability.

4.7 Latent Uncertainty Visualization

To investigate the interpretability of the task-conditioned probabilistic representation, we visualize the latent uncertainty maps generated by our framework. Specifically, the uncertainty is derived from the variance of the learned latent distribution, i.e., $\sigma^2 = \exp(\log \sigma^2)$, which reflects the uncertainty of latent representations during the fusion process. As shown in Figure. 7, higher uncertainty responses are mainly concentrated on task-relevant regions, such as pedestrians and vehicles, indicating that the model assigns larger uncertainty to regions requiring more adaptive fusion of infrared and visible information. In contrast, background regions with consistent information exhibit lower uncertainty. These observations demonstrate that the introduced task condition $\mathbf{Y}$ guides the latent representation to focus on task-relevant regions and enables the proposed framework to adaptively model uncertainty according to both multimodal characteristics and downstream task requirements.

5 Conclusion

In this paper, we propose a probabilistic and task-oriented infrared and visible image fusion framework, which effectively models the uncertainty of fused representations and incorporates task-related constraints to generate more informative fusion results. Extensive experiments on multiple datasets demonstrate that the proposed method achieves competitive fusion quality while providing more effective fused images for downstream vision tasks, including object detection and semantic segmentation. Ablation studies further verify the effectiveness of each designed component and highlight the importance of probabilistic modeling and task consistency constraints. In the future, we will explore more adaptive, generalizable, and efficient task-aware fusion strategies for broader multi-modal perception scenarios.

AI USE STATEMENT

In this work, we used large language models to improve the grammar and clarity of the manuscript and to assist with code revision. The authors designed the overall code structure and wrote the main implementation. These language models were not used to generate experimental data. All AI-assisted content was reviewed by the authors, who take full responsibility for the entire work.

REPRODUCIBILITY STATEMENT

The main paper describes the proposed method, datasets, and evaluation metrics. The appendix provides detailed mathematical derivations and additional experiment results. We will release the code upon acceptance to support reproducibility.

REFERENCES

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7457–7468. IEEE, 2025.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Lin Guo, Xiaoqing Luo, Yue Liu, Zhancheng Zhang, and Xiaojun Wu. Sam-guided multi-level collaborative transformer for infrared and visible image fusion. *Pattern Recognition*, 162:111391, 2025.

Junjia Huang, Pengxiang Yan, Jiyang Liu, Jie Wu, Zhao Wang, Yitong Wang, Liang Lin, and Guanbin Li. Dreamfuse: Adaptive image fusion with diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Hui Li, Xiao-Jun Wu, and Tariq Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12):9645–9656, 2020.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5792–5801. IEEE, 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 *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 8081–8090. IEEE, 2023a.

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.

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.

Xiaowen Liu, Hongtao Huo, Jing Li, Shan Pang, and Bowen Zheng. A semantic-driven coupled network for infrared and visible image fusion. *Information Fusion*, 108:102352, 2024c.

Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

Zhu Liu, Jinyuan Liu, Guanyao Wu, Long Ma, Xin Fan, and Risheng Liu. Bi-level dynamic learning for jointly multi-modality image fusion and beyond. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1240–1248, 2023b.

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

Yu Shi, Yu Liu, Juan Cheng, Huafeng Li, and Xun Chen. Semantic-guided diffusion sampling: A generalized strategy for enhancing object segmentation oriented multimodal image fusion. *IEEE Journal of Selected Topics in Signal Processing*, 19(8):1603–1616, 2025.

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, 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, 2023a.

Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8546–8555, 2024.

Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17882–17891, 2025. doi: 10.1109/CVPR52734.2025.01666.

Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1):37, 2024.

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.

Zengyi Yang, Yafei Zhang, Huafeng Li, and Yu Liu. Instruction-driven fusion of infrared–visible images: Tailoring for diverse downstream tasks. *Information Fusion*, 121:103148, 2025.

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.

Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4484–4496. IEEE, 2025.

Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8048–8059. IEEE, 2023.

Chunyu Zhu, Shangqi Deng, Xuan Song, Yachao Li, and Qi Wang. Mamba collaborative implicit neural representation for hyperspectral and multispectral remote sensing image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7099–7108, 2024. doi: 10.1109/CVPR52733.2024.00678.

A Appendix

A.1 Bayesian Derivation of Task-Conditioned Fusion Formulation

The proposed formulation is derived from a Bayesian perspective by considering image fusion and downstream perception as two correlated random processes. Let $\mathbf{A}$ and $\mathbf{B}$ denote the source modalities, $\mathbf{F}$ denote the fused image, and Y represent the downstream task variable. Instead of independently optimizing fusion and task prediction, we model their dependency through the joint conditional distribution:

$$P(\mathbf{F}, Y|\mathbf{A}, \mathbf{B}). \tag{26}$$

According to the probability chain rule, the joint distribution can be factorized in two equivalent ways. The first factorization corresponds to the conventional fusion paradigm, where the fused image is generated from source modalities and subsequently used for downstream prediction:

$$P(\mathbf{F}, Y|\mathbf{A}, \mathbf{B}) = P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(Y|\mathbf{F}, \mathbf{A}, \mathbf{B}). \tag{27}$$

This formulation indicates that the fusion process is independent of the downstream task during generation, and the task information is only considered after obtaining the fused result.

Alternatively, the same joint distribution can be factorized by first considering the task condition:

$$P(\mathbf{F}, Y|\mathbf{A}, \mathbf{B}) = P(Y|\mathbf{A}, \mathbf{B})P(\mathbf{F}|\mathbf{A}, \mathbf{B}, Y). \tag{28}$$

This factorization provides a different interpretation of task-oriented fusion, where the downstream objective serves as additional evidence for inferring the desired fused representation.

Since both factorization forms describe the same joint distribution, we obtain:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, Y) = \frac{P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(Y|\mathbf{F}, \mathbf{A}, \mathbf{B})}{P(Y|\mathbf{A}, \mathbf{B})}. \tag{29}$$

For a given pair of source modalities and task condition, the denominator is a normalization term independent of $\mathbf{F}$. Therefore, the task-conditioned fusion distribution can be expressed as:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, Y) \propto P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(Y|\mathbf{F}, \mathbf{A}, \mathbf{B}). \tag{30}$$

Furthermore, since the fused image is expected to preserve the task-relevant information from the source modalities, we approximate:

$$P(Y|\mathbf{F}, \mathbf{A}, \mathbf{B}) \approx P(Y|\mathbf{F}), \tag{31}$$

leading to the final formulation:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, Y) \propto P(\mathbf{F}|\mathbf{A}, \mathbf{B})P(Y|\mathbf{F}). \tag{32}$$

This derivation provides a Bayesian interpretation of task-oriented image fusion. The first term represents the fusion prior induced by multimodal observations, while the second term acts as task-dependent evidence that evaluates whether a fused result satisfies downstream perception requirements. Therefore, instead of using task supervision merely as an optimization objective after fusion generation, the proposed formulation incorporates task information directly into the probabilistic generation process of fused images.

A.2 Motivation and Derivation of Latent Fusion Variable

Although the task-conditioned formulation $P(\mathbf{F}|\mathbf{A},\mathbf{B},Y)$ provides a principled way to describe task-oriented fusion, directly learning this distribution in the image space is challenging. The main difficulty originates from the fact that the fused image is not only a high-dimensional visual signal, but also the consequence of a complex decision process that determines which information from different modalities should be preserved or emphasized.

In particular, given the same pair of source modalities $\mathbf{A}$ and $\mathbf{B}$, there may exist multiple reasonable fusion solutions depending on different downstream requirements. For example, a fusion strategy beneficial for object detection may prefer enhanced target saliency, while a segmentation-oriented strategy may require more accurate semantic boundaries. These different solutions cannot be explicitly represented by a single deterministic mapping:

$$\mathbf{F} = G(\mathbf{A},\mathbf{B}). \tag{33}$$

Therefore, instead of directly modeling the output distribution in the image space, we introduce a latent fusion variable $\mathbf{Z}$ to characterize the hidden decision process underlying fusion generation. Different from conventional latent features that only encode visual representations, $\mathbf{Z}$ is designed to represent latent fusion strategies, which determine how multimodal information should be integrated under specific task requirements.

Formally, the task-conditioned fusion distribution can be decomposed by introducing $\mathbf{Z}$:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B},Y) = \int P(\mathbf{F},\mathbf{Z}|\mathbf{A},\mathbf{B},Y)d\mathbf{Z}. \tag{34}$$

Using the chain rule of probability, we obtain:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B},Y) = \int P(\mathbf{F}|\mathbf{Z},\mathbf{A},\mathbf{B},Y)P(\mathbf{Z}|\mathbf{A},\mathbf{B},Y)d\mathbf{Z}. \tag{35}$$

The introduction of $\mathbf{Z}$ provides a hierarchical representation of the fusion process. Specifically, the latent distribution $P(\mathbf{Z}|\mathbf{A},\mathbf{B},Y)$ describes the probability of different fusion strategies under the given source modalities and task conditions, while $P(\mathbf{F}|\mathbf{Z},\mathbf{A},\mathbf{B},Y)$ describes how a selected fusion strategy is transformed into the final visual output.

This decomposition separates two fundamentally different uncertainties in image fusion. The first is the *strategy uncertainty*, indicating that multiple fusion decisions may be suitable for the same input depending on task requirements. The second is the *generation uncertainty*, describing the variability of fused images produced from different latent strategies. Such a decomposition enables the model to capture the distributional structure of fusion solutions rather than only approximating their average output.

Furthermore, since $\mathbf{Z}$ is assumed to contain the essential information required for fusion generation, the fused image generation process is conditionally independent of the original variables given the latent fusion strategy:

$$\mathbf{F} \perp (\mathbf{A},\mathbf{B},Y)|\mathbf{Z}. \tag{36}$$

Accordingly, the generation process can be simplified as:

$$P(\mathbf{F}|\mathbf{Z},\mathbf{A},\mathbf{B},Y) = P(\mathbf{F}|\mathbf{Z}), \tag{37}$$

and the task-conditioned fusion distribution becomes:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B},Y) = \int P(\mathbf{F}|\mathbf{Z})P(\mathbf{Z}|\mathbf{A},\mathbf{B},Y)d\mathbf{Z}. \tag{38}$$

This formulation reveals the essential role of the latent fusion variable. Instead of learning a direct correspondence between multimodal inputs and a single fused image, the model first infers a distribution over possible fusion strategies and subsequently generates corresponding fusion outcomes. Consequently, different samples from $P(\mathbf{Z}|\mathbf{A},\mathbf{B},Y)$ can produce diverse but task-consistent fused results.

Therefore, $\mathbf{Z}$ serves as an intermediate probabilistic representation that bridges multimodal evidence and downstream task requirements. It enables the proposed framework to model the intrinsic diversity of task-oriented fusion solutions and provides the foundation for posterior learning during training and prior-based inference without task annotations.

Algorithm 1 Task-Oriented Probabilistic Image Fusion

Require: Multimodal images $\mathbf{A}$ and $\mathbf{B}$, task annotation Y during training, fusion encoder $E(\cdot)$, posterior network $Q_\phi(\cdot)$, prior network $P_\theta(\cdot)$, fusion decoder $D_\psi(\cdot)$, task predictor $T(\cdot)$.
Ensure: Fused image $\mathbf{F}$.

1: **Training Phase:**
2: Extract modality features:

$$\mathbf{H}_A = E_A(\mathbf{A}), \quad \mathbf{H}_B = E_B(\mathbf{B})$$

3: Generate multimodal representation:

$$\mathbf{H}_{AB} = E_c(\mathbf{H}_A, \mathbf{H}_B)$$

4: Encode task information:

$$\mathbf{H}_Y = E_Y(Y)$$

5: Obtain task-informed posterior distribution:

$$q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y) = \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$$

6: Obtain source-conditioned prior distribution:

$$p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$$

7: Sample latent fusion representation using reparameterization:

$$\mathbf{Z} = \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$$

8: Decode latent representation:

$$\mathbf{F} = D_\psi(\mathbf{Z})$$

9: Obtain task predictions from source modalities and fused image:

$$P_s(Y|\mathbf{A}, \mathbf{B}) = T(\mathbf{H}_{AB})$$
$$P_f(Y|\mathbf{F}) = T(\mathbf{F})$$

10: Optimize the overall objective:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 D_{KL}(q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)||p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})) + \lambda_2 D_{KL}(P_s(Y|\mathbf{A}, \mathbf{B})||P_f(Y|\mathbf{F}))$$

11: **Inference Phase:**
12: Extract multimodal representation:

$$\mathbf{H}_{AB} = E_c(E_A(\mathbf{A}), E_B(\mathbf{B}))$$

13: Predict task-free latent distribution:

$$p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$$

14: Generate latent representation:

$$\mathbf{Z} = \boldsymbol{\mu}_p + \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}$$

15: Generate fused image:

$$\mathbf{F} = D_\psi(\mathbf{Z})$$

16: **return** $\mathbf{F}$

A.3 Variational Inference and ELBO Derivation

The proposed framework aims to learn a task-aware latent fusion distribution while allowing task-free inference. According to the latent variable formulation, the task-free generative process is defined as:

$$p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}) = \int p_\theta(\mathbf{F}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}) d\mathbf{Z}, \tag{39}$$

where $\mathbf{Z}$ represents the latent fusion strategy. Since $\mathbf{Z}$ is unobserved, directly maximizing the marginal likelihood is intractable. Therefore, we introduce a variational posterior distribution:

$$q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y), \tag{40}$$

which utilizes task annotations during training to infer task-dependent latent fusion strategies.

Starting from the conditional log-likelihood:

$$\log p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}), \tag{41}$$

we substitute the marginal distribution of $\mathbf{F}$:

$$\log p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}) = \log \int p_\theta(\mathbf{F}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B}) d\mathbf{Z}. \tag{42}$$

To introduce the variational posterior, we multiply and divide the integrand by $q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)$:

$$\begin{aligned} \log p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}) = \log \int & \frac{p_\theta(\mathbf{F}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \\ & \times q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y) d\mathbf{Z}. \end{aligned} \tag{43}$$

The above expression can be rewritten as an expectation with respect to the variational posterior:

$$\log p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}) = \log \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \left[\frac{p_\theta(\mathbf{F}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \right]. \tag{44}$$

Since the logarithm is a concave function, applying Jensen's inequality yields:

$$\log p_\theta(\mathbf{F}|\mathbf{A}, \mathbf{B}) \geq \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \left[\log \frac{p_\theta(\mathbf{F}|\mathbf{Z}) p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \right]. \tag{45}$$

Therefore, the evidence lower bound (ELBO) is:

$$\begin{aligned} \mathcal{L}_{ELBO} = & \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} [\log p_\theta(\mathbf{F}|\mathbf{Z})] \\ & + \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \left[\log \frac{p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \right]. \end{aligned} \tag{46}$$

The second expectation term corresponds to the negative KL divergence between the task-informed posterior and the source-conditioned prior:

$$\begin{aligned} & \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \left[\log \frac{p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})}{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} \right] \\ & = -D_{KL}(q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y) || p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})). \end{aligned} \tag{47}$$

Thus, the final variational objective becomes:

$$\begin{aligned} \mathcal{L}_{ELBO} = & \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)} [\log p_\theta(\mathbf{F}|\mathbf{Z})] \\ & - D_{KL}(q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y) || p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})). \end{aligned} \tag{48}$$

The first term optimizes the reconstruction of fused images from latent fusion strategies, while the second term performs posterior-to-prior knowledge transfer. Specifically, the posterior $q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y)$ acts as a task-aware teacher during training, whereas the prior $p_\theta(\mathbf{Z}|\mathbf{A}, \mathbf{B})$ absorbs such task knowledge and enables label-free inference. Therefore, the proposed variational objective

does not merely regularize the latent space, but explicitly transfers downstream task information from the training-time posterior to the inference-time prior.

The overall optimization and inference procedure is summarized in Algorithm 1.

B Relationship with Existing Fusion Paradigms

In this section, we provide a formulation-level comparison between the proposed framework and existing image fusion paradigms. Unlike conventional comparisons based on network architectures, we analyze the fundamental probabilistic assumptions underlying different fusion strategies.

B.1 Deterministic Fusion Formulation

Most existing image fusion methods implicitly formulate fusion as a deterministic mapping from source modalities to a single fused image:

$$\mathbf{F} = G_\theta(\mathbf{A}, \mathbf{B}), \tag{49}$$

where $\mathbf{A}$ and $\mathbf{B}$ represent the source modalities and $\mathbf{F}$ is the generated fused image. Although modern deep networks significantly improve the capacity of this mapping, the formulation itself remains deterministic.

From a probabilistic perspective, such a formulation can be interpreted as learning a degenerate conditional distribution:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}) \approx \delta\left(\mathbf{F} - G_\theta(\mathbf{A}, \mathbf{B})\right), \tag{50}$$

where $\delta(\cdot)$ denotes a Dirac distribution. This indicates that given the same source modalities, the model always produces a single fusion solution. Consequently, the potential diversity of fusion strategies and the underlying distribution of possible fusion outcomes are not explicitly modeled.

B.2 Task-Oriented Fusion Formulation

Recent task-oriented fusion methods incorporate downstream supervision to improve the utility of fused images for perception tasks. Typically, these methods optimize the fusion network with an additional task objective:

$$\mathcal{L} = \mathcal{L}_{fusion} + \lambda \mathcal{L}_{task}, \tag{51}$$

where $\mathcal{L}_{task}$ encourages the fused representation to preserve task-relevant information.

However, the fusion generation process remains unchanged:

$$\mathbf{F} = G_\theta(\mathbf{A}, \mathbf{B}). \tag{52}$$

In other words, task information is introduced as an optimization constraint rather than a variable that directly determines the fusion distribution. Therefore, these approaches still learn a single task-optimized fusion output instead of explicitly modeling:

$$P(\mathbf{F}|\mathbf{A}, \mathbf{B}, Y), \tag{53}$$

where Y represents downstream task information.

Table 2: Comparison of different image fusion formulations.

Method	Task Condition	Distribution Modeling	Inference Type
Deterministic Fusion	×	×	Single output
Task-Oriented Fusion	✓	×	Single output
Probabilistic Fusion	×	✓	Distribution modeling
Ours	✓	✓	Task-conditioned inference

B.3 Probabilistic Fusion Formulation

Another line of research explores probabilistic generative models for image synthesis and fusion. These methods aim to model the distribution of possible outputs:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B}), \tag{54}$$

rather than a single deterministic prediction. By introducing latent variables, such models can represent multiple possible fusion outcomes.

However, without explicit task conditioning, the learned distribution only describes the general variability of fused images. It does not characterize how downstream objectives should influence the selection of fusion strategies.

B.4 Proposed Task-Conditioned Probabilistic Formulation

Different from the above paradigms, we formulate task-oriented image fusion as learning a task-conditioned conditional distribution:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B},Y). \tag{55}$$

In this formulation, the downstream task variable Y is no longer treated as an external optimization signal, but becomes an explicit condition that determines the fusion distribution.

To make this distribution tractable, we introduce a latent fusion variable $\mathbf{Z}$:

$$P(\mathbf{F}|\mathbf{A},\mathbf{B},Y) = \int P(\mathbf{F}|\mathbf{Z})P(\mathbf{Z}|\mathbf{A},\mathbf{B},Y)d\mathbf{Z}. \tag{56}$$

The latent variable $\mathbf{Z}$ captures task-dependent fusion strategies, allowing the model to represent different possible fusion outcomes under different task requirements. During training, the task-informed posterior:

$$q_\phi(\mathbf{Z}|\mathbf{A},\mathbf{B},Y) \tag{57}$$

learns how downstream objectives influence fusion strategies. During inference, the learned knowledge is transferred into the task-free prior:

$$p_\theta(\mathbf{Z}|\mathbf{A},\mathbf{B}), \tag{58}$$

enabling task-aware fusion without requiring task annotations.

The differences between existing formulations and the proposed framework are summarized in Table 2.

Overall, the proposed framework differs from existing fusion paradigms at the formulation level. Instead of learning a deterministic correspondence between multimodal inputs and a single fused image, we model fusion as probabilistic inference conditioned on downstream task requirements. This formulation enables the explicit representation of task-dependent fusion strategies and provides a principled way to transfer task knowledge from training-time supervision to label-free inference.

C Task Sufficiency Constraint Analysis

The proposed framework introduces a task sufficiency constraint to ensure that the fused image preserves information that is relevant to downstream perception. Different from directly optimizing task losses on fused images, our objective focuses on maintaining the consistency between task prediction distributions obtained from source modalities and the generated fused image.

Given the source modalities $\mathbf{A}$ and $\mathbf{B}$, a task prediction network can directly estimate the downstream prediction distribution:

$$P_s(Y|\mathbf{A},\mathbf{B}), \tag{59}$$

where the subscript s denotes the prediction distribution obtained from source modalities. Since the source images contain complementary information from multiple modalities, this distribution can be regarded as the task-related knowledge available before fusion.

Meanwhile, the fused image $\mathbf{F}$ is expected to provide sufficient information for the same downstream task. Therefore, a task predictor is applied to the fused result to obtain:

$$P_f(Y|\mathbf{F}), \tag{60}$$

where the subscript f denotes the prediction distribution inferred from the fused image.

Ideally, if the fusion process successfully preserves task-relevant information from the source modalities, the downstream prediction distributions before and after fusion should be consistent:

$$P_f(Y|\mathbf{F}) \approx P_s(Y|\mathbf{A},\mathbf{B}). \tag{61}$$

To explicitly enforce this property, we minimize the Kullback–Leibler divergence between the two predictive distributions:

$$\mathcal{L}_{task} = D_{KL}(P_s(Y|\mathbf{A},\mathbf{B})||P_f(Y|\mathbf{F})). \tag{62}$$

This objective can be interpreted as a distribution-level knowledge transfer process. The source modalities provide a richer observation space containing complementary information, and their prediction distribution serves as a soft teacher that guides the task prediction of the fused image.

Unlike conventional task-guided fusion methods that directly optimize the fused image with hard task labels, our formulation does not force the fused image to reproduce a specific prediction. Instead, it encourages the fused representation to maintain the same probabilistic decision behavior as the original multimodal input. This property is particularly important for dense perception tasks, where multiple predictions can be uncertain or ambiguous.

C.1 Implementation for Semantic Segmentation and Object Detection

In practical implementations, the task sufficiency constraint is applied according to the specific downstream perception task. For semantic segmentation, the task prediction network outputs a categorical probability distribution for each pixel. Given an image, the segmentation model produces:

$$P_s(Y_{seg}|\mathbf{A},\mathbf{B}) \in \mathbb{R}^{H\times W\times C}, \tag{63}$$

where H and W denote the spatial resolution and C represents the number of semantic categories.

Similarly, the fused image generates:

$$P_f(Y_{seg}|\mathbf{F}) \in \mathbb{R}^{H\times W\times C}. \tag{64}$$

The KL divergence is calculated at the pixel level:

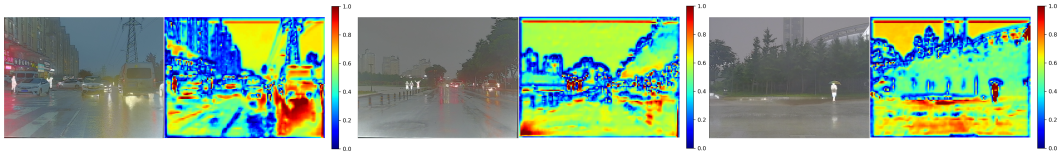


Figure 8: Visualization of task-guided detector activation maps on fused images. The left part of each sample shows the fused image, while the right part presents the corresponding detector activation map extracted from the downstream detection network. Warmer colors indicate stronger semantic responses. The activation maps demonstrate that the proposed task-conditioned fusion framework encourages the fused representations to focus on task-relevant objects, such as pedestrians and vehicles, rather than irrelevant background regions.

$$\mathcal{L}_{seg} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} D_{KL}(P_s^{i,j} || P_f^{i,j}), \tag{65}$$

where $P_s^{i,j}$ and $P_f^{i,j}$ denote the class probability distributions of the pixel at location (i, j).

For object detection, the task network predicts both classification confidence and localization information. The classification branch produces category distributions:

$$P_s(Y_{det}^{cls} | \mathbf{A}, \mathbf{B}), \tag{66}$$

and:

$$P_f(Y_{det}^{cls} | \mathbf{F}). \tag{67}$$

The task consistency loss is applied to the classification probability distributions:

$$\mathcal{L}_{det} = D_{KL}(P_s(Y_{det}^{cls} | \mathbf{A}, \mathbf{B}) || P_f(Y_{det}^{cls} | \mathbf{F})). \tag{68}$$

For both segmentation and detection tasks, the source-modal prediction distribution provides task-aware guidance, while the fused-image prediction distribution is optimized to approach this target distribution. Consequently, the fusion network is encouraged to retain the information required by downstream perception rather than only optimizing visual reconstruction quality.

Overall, the task sufficiency constraint establishes a direct connection between fusion generation and downstream decision making. By aligning prediction distributions instead of only minimizing task errors, the proposed method ensures that the learned fusion distribution remains consistent with the task requirements.

C.2 Task-guided Representation Visualization

To further investigate whether the proposed task-conditioned probabilistic fusion framework learns representations aligned with downstream perception requirements, we visualize the activation responses of the downstream detector on the generated fused images.

Specifically, after obtaining the fused image $\mathbf{F}$, it is fed into the detection network, and the final-stage feature activations are averaged across channels to generate a spatial activation map. The activation intensity reflects the degree to which different regions contribute to downstream object recognition.

As shown in Figure. 8, the detector responses are not uniformly distributed over the fused images. Instead, high activation responses are mainly concentrated on semantically meaningful regions, including pedestrians, vehicles, and other potential targets. In contrast, background structures, such as buildings, roads, and electric towers, generally exhibit weaker responses with cooler colors.

This phenomenon indicates that the learned fusion representation preserves and emphasizes visual patterns that are beneficial for downstream perception. From the perspective of the proposed probabilistic formulation, the task condition Y provides additional guidance for learning the latent fusion distribution:

$$q_\phi(\mathbf{Z}|\mathbf{A}, \mathbf{B}, Y), \tag{69}$$

which encourages the generated fused images to contain task-discriminative information. Consequently, the downstream detector can extract stronger responses from regions associated with target objects.

Different from conventional fusion approaches that mainly optimize visual consistency between modalities, the proposed framework incorporates task information into the fusion generation process. The activation visualization provides qualitative evidence that the learned task-conditioned fusion distribution produces representations that are better aligned with downstream detection objectives.