TeXIR: Reliability- and Integrability-Aware Structure Image Fusion

Anonymous authors
Paper under double-blind review

Abstract

Visible and infrared image fusion involves two fundamental decisions: **what information to preserve** and **how that information should appear** in the output. Most existing methods map the two modalities directly to a fused image, coupling information selection with image formation. We introduce **TeXIR**, which breaks this coupling by first constructing an explicit **TeX**-inspired scene state. TeXIR builds this state according to state-specific **R**eliability and, for structure, enforces **I**ntegrability so that locally reliable contributions form a globally coherent representation. Once the scene state is fixed, a separate readout controls its visual presentation without revisiting the fusion decision. The same fused state can therefore support different appearance conditions while preserving the underlying scene content. Experiments on FMB, MSRS, and RoadScene show that TeXIR preserves complementary source information while achieving strong perceptual quality across diverse scenes. Different readouts further exhibit the intended separation between source fidelity and visual presentation from the same fused state. TeXIR thus formulates visible and infrared image fusion as reliable scene state construction followed by controllable image formation. Code will be made publicly available.

1 Introduction

Visible and infrared image fusion combines complementary observations of the same scene into a single image. Visible images largely encode reflected light shaped by illumination and surface reflectance, whereas infrared images encode thermal radiation shaped by temperature and material emissivity. The same object can therefore be informative in both modalities while carrying different physical meanings. Fusion must consequently answer two distinct questions: what information should be preserved, and how should that information appear in the output? Despite their architectural differences, many existing methods learn a direct mapping from the inputs to the fused image, leaving these two decisions implicit (Li & Wu, 2019; Ma et al., 2019; Xu et al., 2022; Ma et al., 2022; Zhao et al., 2023b). Such mappings couple information selection with visual presentation, so deciding what to preserve also constrains how that information appears in the output.

This coupling becomes harmful when one modality helps determine what to preserve but gives misleading cues for how it should appear. Figure 1 illustrates this case. A pedestrian in dark clothing remains clearly visible in the visible image but exhibits strong thermal contrast in infrared. Emphasizing the infrared response may reasonably improve target saliency. When information selection also determines appearance, however, greater infrared influence can brighten, recolor, or otherwise distort an already valid visible appearance. This coupling makes the influence assigned to a modality serve two roles at once, controlling both how much information enters the fusion and how strongly that modality shapes the output appearance. Consequently, stronger infrared support can unintentionally produce a more infrared-like appearance.

We address this coupling by dividing fusion into state construction and image formation. HADAR provides a useful precedent: it first decomposes thermal observations into temperature, emissivity, and texture, and only then renders these recovered attributes as TeX vision (Bao et al., 2023). We adopt this separation principle, rather than the original HADAR sensing model, to define explicit states for visible and infrared fusion. In our formulation, X captures scene structure, e carries region identity and visible color information, and T retains thermal state. Each modality contributes

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.

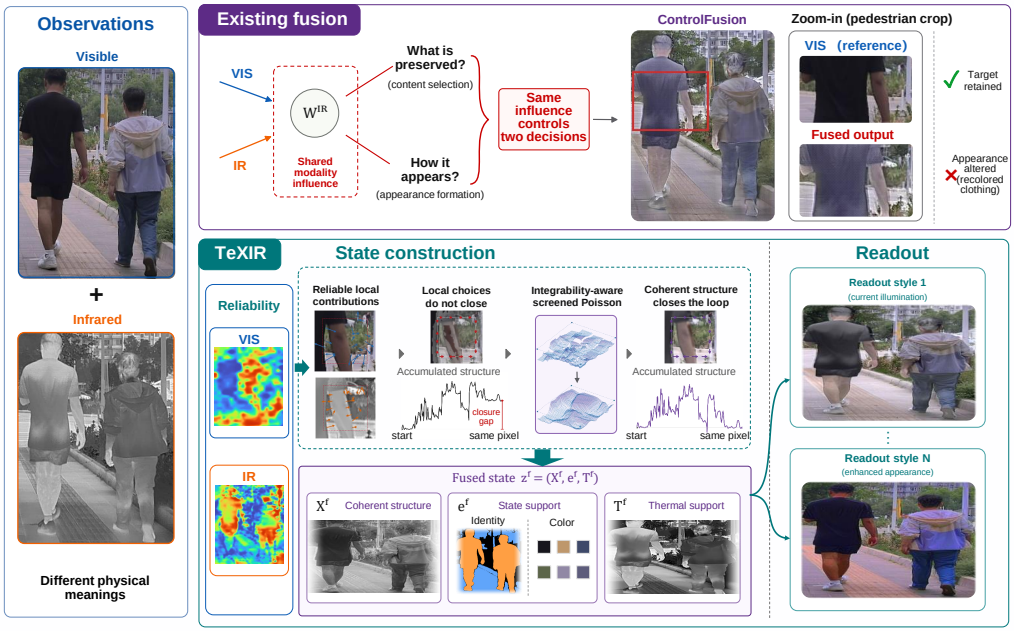

Figure 1: **TeXIR separates what is preserved from how it is presented.** Existing fusion can use the same modality influence to control both information selection and visual appearance. TeXIR instead uses state specific reliability to determine local contributions and integrability to reconcile these choices into a coherent structure before reconstruction. The resulting TeX inspired state is constructed once and reused by a separate readout to support multiple visual presentations without changing the preserved scene content.

only to the states it can support, while a separate readout determines how the fused states should appear. Infrared information can therefore change what the fused scene preserves without directly prescribing its visible brightness or color. Fusion then becomes a more precise question: which modality should contribute to each state, where, and by how much?

Each modality contributes differently across states and spatial locations. We describe this variation with reliability, which determines how strongly visible or infrared contributes to each state at each location. For structure, however, neighboring locations cannot be treated independently because they jointly determine the final structure. Reliability can justify each local contribution without ensuring that the contributions work together as a whole. We call this second requirement integrability: the locally weighted contributions must form one coherent structure. Reliability governs local contribution, while integrability governs how these contributions fit together. We build our framework, **TeXIR**, around these two requirements. TeXIR uses state-specific reliability to control local contributions to each state. For structure, it further reconciles these contributions through an integrability-aware screened Poisson reconstruction. A separate readout then renders the fused states into an image, keeping information selection separate from visual presentation.

Our contributions are **threefold**: (**1**) we formulate visible and infrared image fusion as TeX-inspired state construction with state-specific reliability, separating modality support from visual presentation; (**2**) we distinguish reliability from integrability in structure fusion. Reliability determines local modality contributions, while an integrability aware optimization adjusts these contributions to obtain a globally coherent structure; and (**3**) we introduce a controllable readout that reuses the same fused state across multiple output forms. TeXIR produces deterministic state views and natural RGB renderings under different appearance conditions without reconstructing the fused state. Together, TeXIR reframes visible and infrared fusion from direct image mixing to reliable state construction followed by controllable readout.

2 Related Work

Decomposition in Image Fusion. Decomposition methods separate source information into components that serve different roles during fusion. DIDFuse (Zhao et al., 2020) separates background

and detail features, whereas DeFusion (Liang et al., 2022) and C2RF (Tang et al., 2025b) decompose representations into common and unique components. CDDFuse (Zhao et al., 2023a) instead distinguishes shared low-frequency features from modality-specific high-frequency features. More recent methods move toward representations with more explicit roles. FreeFusion (Zhao et al., 2025) aligns complementary modal representations through cross reconstruction, URMIF (Liu et al., 2026) separates invariant structure from modality-specific texture, and ISFuse (Wang et al., 2026b) seeks stable cross-modal information under interventions. TeXIR further makes the fusion roles explicit in the state itself, with separate states for structure, region and color information, and thermal response before image formation.

HADAR (Bao et al., 2023) provides a related representation principle from thermal sensing. It decomposes hyperspectral thermal observations into temperature, emissivity, and texture before mapping these recovered attributes to TeX vision. HADAR further extends this representation to conventional thermal data through pseudo TeX vision and demonstrates fusion between TeX and RGB images. TeXIR transfers this representation principle to RGB and single-band infrared fusion. Because these inputs lack the spectral information required for physical TeX recovery, TeXIR treats X, e, and T as operational states and constructs them before a separate readout determines the output appearance.

Gradient Fusion and Integrability. Gradient-based fusion represents local structure through image derivatives rather than intensity alone. Poisson image editing (Pérez et al., 2003) provides a classical way to reconstruct an image from prescribed local gradients. GTF (Ma et al., 2016) adapts this idea to visible and infrared fusion by preserving infrared intensity while transferring visible gradients through a variational objective. Recent methods continue to use gradients as structural constraints. Wang et al. (2025) constrain fused images with Sobel and Laplacian gradients to retain strong source structures, while FRUFN Wang et al. (2026a) transfers gradient information across fusion tasks to exploit complementary structural details. These methods preserve or transfer local gradients as structural cues. For multimodal fusion, local reliability further determines which source should provide each derivative. These locally reliable choices together define the target gradient field, which must satisfy integrability to correspond to a single image. TeXIR therefore uses integrability to refine reliability based source selection before reconstruction.

Controllable Image Fusion. Generative methods increasingly combine image priors with restoration and fusion. DDFM (Zhao et al., 2023b), Deno-IF (Xu et al., 2025), DRFusion (Shi et al., 2026), and ReCoFuse (Zhang et al., 2026) use generative or restoration priors to model fused images under clean or degraded observations. Recent methods also introduce explicit control over the fusion result. Text-IF (Yi et al., 2024), Text-DiFuse (Zhang et al., 2024), ControlFusion (Tang et al., 2025a), and DPOFusion (Su et al., 2026) use semantic instructions, degradation prompts, source modulation, or preference optimization to steer fusion toward different requirements. TeXIR instead separates fusion from readout: it constructs a fused state once and renders the same state into different visual forms. Output control therefore changes the readout without changing the fused state.

3 Proposed Method

Overview. TeXIR separates state construction from image formation. Given visible and infrared images I_v and I_i, TeXIR first estimates the reliability of the modality information relevant to each state. These estimates control how that information contributes locally during state construction. For the structural state, reliability establishes local modality preferences, and integrability then refines them before reconstruction to obtain a globally coherent structure. A separate readout renders the same fused state under a controllable appearance condition. Figure 2 summarizes this pipeline.

We formalize these stages with a reliability estimator $\mathcal{Q}$, a state constructor $\mathcal{C}$, and a readout $\mathcal{R}$:

$$(\mathbf{R}_v, \mathbf{R}_i) = \mathcal{Q}\left(I_v, I_i; \boldsymbol{\theta}\right), \tag{1}$$

$$z_f = \mathcal{C}\left(I_v, I_i, \mathbf{R}_v, \mathbf{R}_i\right) = \left(\boldsymbol{X}_f, e_f, \boldsymbol{T}_f\right), \tag{2}$$

$$I_q = \mathcal{R}\left(z_f, a_q; \varphi\right). \tag{3}$$

Here, $\mathbf{R}_v$ and $\mathbf{R}_i$ collect the reliability maps for the visible and infrared inputs across states, and $\boldsymbol{\theta}$ and φ denote the parameters of $\mathcal{Q}$ and $\mathcal{R}$. The fused state $z_f = (\boldsymbol{X}_f, e_f, \boldsymbol{T}_f)$ contains structure, region and color information, and thermal response. We write $e_f = (\boldsymbol{L}_f, C_v, C_f)$, where $\boldsymbol{L}_f$

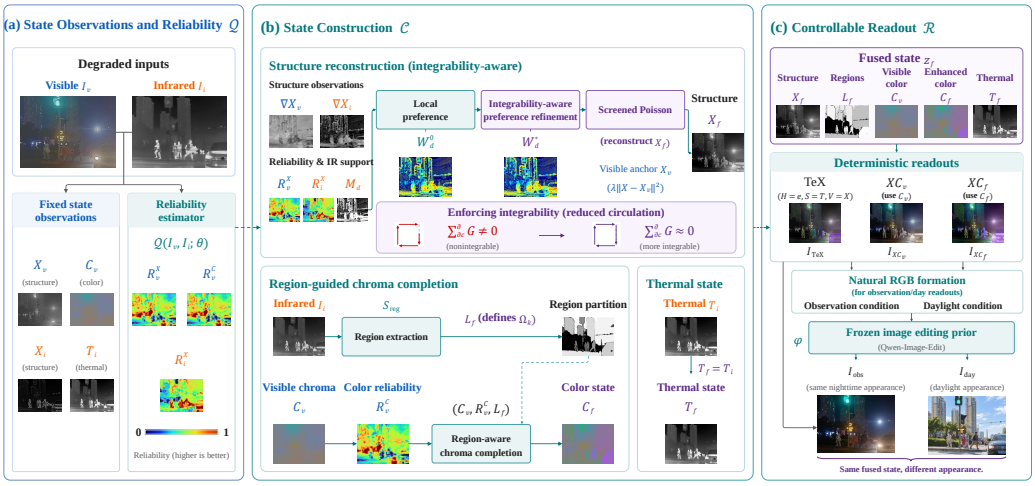

Figure 2: **Overview of TeXIR. (a)** TeXIR constructs state observations from the two inputs and estimates their state-specific reliability. **(b)** The state constructor uses reliability to establish local structure preferences, refines them for integrability before screened Poisson reconstruction, and constructs the remaining region, chroma, and thermal states to form the fused state z_f. **(c)** The readout uses deterministic image forms and a frozen image editing prior to render the constructed state under different readout conditions.

represents region identity and C_v and C_f carry color information. The condition a_q controls the appearance of the rendered image I_q without changing z_f.

3.1 State Observations and Reliability

TeX decomposition separates thermal radiation into texture X, emissivity e, and temperature T (Bao et al., 2023). TeXIR uses this separation to organize visible RGB and single band infrared observations into three operational states. In TeXIR, X represents spatial structure and T represents thermal response. We use $e = (L, C)$ to represent nonstructural region information: L encodes region identity, while C preserves visible chroma. Both modalities provide structural information, whereas the visible image additionally provides natural color and the infrared image provides thermal response. We therefore construct $\boldsymbol{X}_v$ and retain C_v from I_v, and construct $\boldsymbol{X}_i$ and $\boldsymbol{T}_i$ from I_i. The fixed observation operators are given in Appendix A.1; Section 3.3 constructs $\boldsymbol{L}_f$. Structure is thus the only state with direct observations from both modalities at this stage, which creates the local cross-modal choice addressed by reliability.

Reliability measures how faithfully an observation captures the corresponding state at each location. TeXIR estimates three reliability maps:

$$\mathbf{R}_v = (\boldsymbol{R}_v^X, \boldsymbol{R}_v^C), \qquad \mathbf{R}_i = \boldsymbol{R}_i^X. \tag{4}$$

The structure reliabilities $\boldsymbol{R}_v^X$ and $\boldsymbol{R}_i^X$ quantify the fidelity of the visible and infrared structure observations. Comparing them yields an initial local preference between the two structure observations. The chroma reliability $\boldsymbol{R}_v^C$ measures the reliability of visible color. The thermal response comes only from the infrared input, while the region state is constructed separately; neither requires modality selection.

The estimator $\mathcal{Q}$ in Eq. (1) predicts these reliability maps jointly from I_v and I_i. We train $\mathcal{Q}$ with spatially varying synthetic degradations and fine tune it on a small annotated set of real observations. Appendix A.2 describes the degradation models, reliability targets, and annotations. Local structure preferences can violate gradient integrability. Section 3.2 therefore adjusts these preferences before reconstructing $\boldsymbol{X}_f$.

3.2 Integrability and Structure Reconstruction

The local structure preferences from Section 3.1 encode which modality is favored at each location. For reconstruction, we apply these preferences to the horizontal and vertical derivatives. The

resulting gradient field may not correspond to any image, because integrability requires the signed changes around every grid cell to sum to zero. A Poisson solver can still reconstruct an image from a nonintegrable field, but the reconstructed gradients must then differ from the target derivatives. TeXIR therefore adjusts the directional preferences before Poisson reconstruction to reduce integrability violations in the target field.

To obtain these directional preferences, we combine the two structure reliabilities with a directional infrared support map $\boldsymbol{M}_d$, where $d \in \{x, y\}$. The reliability ratio sets the local balance between the two modalities, while $\boldsymbol{M}_d$ suppresses infrared derivatives dominated by background noise. We define the initial infrared preference as $\boldsymbol{W}_d^0 = \boldsymbol{M}_d \odot \boldsymbol{R}_i^X / (\boldsymbol{R}_i^X + \boldsymbol{R}_v^X + \epsilon)$. A value near zero favors the visible derivative, whereas a value near one favors the infrared derivative. Given a directional preference $\boldsymbol{W}_d$, we form the corresponding target derivative as $\boldsymbol{G}_d(\boldsymbol{W}_d) = (1 - \boldsymbol{W}_d) \odot \nabla_d \boldsymbol{X}_v + \boldsymbol{W}_d \odot \nabla_d \boldsymbol{X}_i$. Its construction is given in Appendix A.3.

Starting from $\boldsymbol{W}_d^0$, we optimize the horizontal and vertical preferences jointly to reduce local circulation while remaining close to the initial modality preferences:

$$
\begin{aligned}
(\boldsymbol{W}_x^\star, \boldsymbol{W}_y^\star) = \arg \min_{0 \le \boldsymbol{W}_x, \boldsymbol{W}_y \le 1} & \sum_{d \in \{x,y\}} \|\boldsymbol{W}_d - \boldsymbol{W}_d^0\|_2^2 \\
& + \mu \left\| \mathscr{C}\big(\boldsymbol{G}_x(\boldsymbol{W}_x), \boldsymbol{G}_y(\boldsymbol{W}_y)\big) \right\|_2^2 + \tau \sum_{d \in \{x,y\}} \mathrm{TV}(\boldsymbol{W}_d).
\end{aligned} \tag{5}
$$

Here, $\mathscr{C}$ computes the discrete circulation around each grid cell. The first term preserves the initial modality preferences, the circulation term promotes integrability, and the TV term suppresses abrupt spatial changes in the weights. This optimization reduces integrability violations at the preference level before reconstruction.

We then reconstruct the fused structure with screened Poisson reconstruction:

$$
\boldsymbol{X}_f = \arg\min_{\boldsymbol{X}} \sum_{d \in \{x,y\}} \|\nabla_d \boldsymbol{X} - \boldsymbol{G}_d(\boldsymbol{W}_d^\star)\|_2^2 + \kappa \|\boldsymbol{X} - \boldsymbol{X}_v\|_2^2 . \tag{6}
$$

The gradient term matches the derivatives of $\boldsymbol{X}_f$ to the adjusted targets. The screening term anchors the solution to the visible structure $\boldsymbol{X}_v$ and removes the additive ambiguity of gradient integration. The resulting $\boldsymbol{X}_f$ forms the structural component of the fused state z_f.

3.3 Fused State Construction

With the fused structure $\boldsymbol{X}_f$ reconstructed, we construct $e_f = (\boldsymbol{L}_f, C_v, C_f)$ and $\boldsymbol{T}_f$. The region state $\boldsymbol{L}_f$ provides the spatial support for color completion, C_f preserves and completes visible chroma, and $\boldsymbol{T}_f$ carries the infrared thermal response. We obtain the region state from the infrared image using a frozen segmentation model, $\boldsymbol{L}_f = \mathcal{S}_{\mathrm{reg}}(I_i)$. Let $\Omega_k = \{\boldsymbol{p} : \boldsymbol{L}_f(\boldsymbol{p}) = k\}$ denote region k. Each region defines the set of visible observations used to estimate its chroma reference.

For each region, we first aggregate its visible color reliability, $S_k^C = \sum_{\boldsymbol{p} \in \Omega_k} \boldsymbol{R}_v^C(\boldsymbol{p})$. When S_k^C reaches the reliability threshold τ_C, we estimate the regional chroma reference by weighting the visible chroma with its local reliability. Regions below this threshold use a preset chroma value C_0:

$$
\overline{C}_k = \begin{cases} \dfrac{\sum_{\boldsymbol{p} \in \Omega_k} \boldsymbol{R}_v^C(\boldsymbol{p})\, C_v(\boldsymbol{p})}{S_k^C + \epsilon}, & S_k^C \ge \tau_C, \\ C_0, & S_k^C < \tau_C. \end{cases} \tag{7}
$$

Here, C_0 is a fixed chroma value specified in advance.

We then combine the observed chroma at each location with the reference of its region:

$$
C_f(\boldsymbol{p}) = \boldsymbol{R}_v^C(\boldsymbol{p})\, C_v(\boldsymbol{p}) + \big[1 - \boldsymbol{R}_v^C(\boldsymbol{p})\big] \overline{C}_{k(\boldsymbol{p})}, \tag{8}
$$

where $k(\boldsymbol{p})$ denotes the region containing $\boldsymbol{p}$. High color reliability keeps $C_f(\boldsymbol{p})$ close to the observed chroma, whereas lower reliability gives more weight to the regional reference. In this construction, $\boldsymbol{L}_f$ determines the spatial support for chroma completion, while the visible input determines the chroma values.

The infrared input directly provides the thermal state, so we set $\boldsymbol{T}_f = \boldsymbol{T}_i$. The resulting fused state is $z_f = \big(\boldsymbol{X}_f, (\boldsymbol{L}_f, C_v, C_f), \boldsymbol{T}_f\big) = \big(\boldsymbol{X}_f, e_f, \boldsymbol{T}_f\big)$. The fused state specifies what TeXIR preserves; the readout determines how it appears.

3.4 Controllable Readout

The same fused state supports multiple visual readouts. TeXIR can read the constructed states out deterministically or use these readouts to form natural RGB images. The readout condition a_q selects the output form and supplies the appearance cues required by that form:

$$I_q = \mathcal{R}\left(z_f, a_q; \varphi\right), \qquad q \in \{\mathrm{TeX}, XC_v, XC_f, \mathrm{obs}, \mathrm{day}\}. \tag{9}$$

The TeX, XC_v, and XC_f outputs are deterministic readouts. The TeX form visualizes the fused TeX states. The XC_v form combines the fused structure $\boldsymbol{X}_f$ with the visible chroma C_v, while XC_f combines the same structure with the fused chroma C_f. These readouts provide complementary views of the constructed state and serve as image inputs for natural RGB formation. Appendix A.4 gives their exact mappings.

The observation readout uses a blend of XC_v and XC_f together with the visible image I_v. The blend combines visible and completed chroma on the same fused structure, while I_v provides the observed brightness and illumination. A frozen image editing prior harmonizes these inputs to produce I_{obs}. The daylight readout uses XC_v and XC_f. Both images share $\boldsymbol{X}_f$ but carry visible and fused chroma, respectively. A daylight instruction guides the same frozen prior to form I_{day} with a natural daytime appearance.

Thus, $\mathcal{R}$ produces multiple visual forms from the same fused state. The deterministic readouts expose the constructed state directly, while the natural RGB readouts build on these image forms to produce the requested appearance. Appendix B provides the readout definitions and image fusing instructions.

4 Experiments

4.1 Experimental Setup

Datasets and baselines. We evaluate TeXIR on the full test sets of FMB (Liu et al., 2023), MSRS (Tang et al., 2022), and RoadScene (Xu et al., 2020), which contain 280, 361, and 221 image pairs, respectively. We compare TeXIR with 14 fusion methods: CDDFuse (Zhao et al., 2023a), DDFM (Zhao et al., 2023b), TIMFusion (Liu et al., 2024), Text-IF (Yi et al., 2024), C2RF (Tang et al., 2025b), ControlFusion (Tang et al., 2025a), DCEvo (Liu et al., 2025), Deno-IF (Xu et al., 2025), LUT-Fuse (Yi et al., 2025), APTP (Deng et al., 2026), DPOFusion (Su et al., 2026), DRFusion (Shi et al., 2026), ISFuse (Wang et al., 2026b), and ReCoFuse (Zhang et al., 2026).

Evaluation metrics. We use six conventional fusion metrics: cross entropy (CE), visual information fidelity (VIF), entropy (EN), standard deviation (SD), spatial frequency (SF), and average gradient (AG). We also assess perceptual quality with CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), PAQ2PIQ (Ying et al., 2020), MANIQA (Yang et al., 2022), and TReS (Golestaneh et al., 2022). Lower CE is better; all other metrics favor higher values.

TeXIR readouts. For the main quantitative fusion comparison, we report three outputs derived from the same fused state: TeXIR$_{XC_v}$, TeXIR$_{XC_f}$, and TeXIR$_{\mathrm{day}}$. TeXIR$_{XC_v}$ combines the reconstructed structure $\boldsymbol{X}_f$ with visible chroma C_v, whereas TeXIR$_{XC_f}$ uses the fused chroma C_f. TeXIR$_{\mathrm{day}}$ renders the same fused state with a natural daytime appearance using the frozen image editing prior.

Implementation details. We implement TeXIR in `PyTorch`. The reliability estimator $\mathcal{Q}$ uses a pretrained ViT-Tiny/16 backbone with gated late fusion. We train $\mathcal{Q}$ on 448×448 crops for 100 epochs using AdamW with a batch size of 16, an initial learning rate of $3e-4$, and a weight decay of $1e-4$. Synthetic degradations are sampled online during training. We then fine tune the estimator on annotated real observations and use the resulting model in all experiments. We set $\mu = 1$, $\tau = 0.01$, and $\kappa = 1e-4$ for structure reconstruction. We instantiate $\mathcal{S}_{\mathrm{reg}}$ with frozen SAM3 and use Qwen-Image-Edit (Wu et al., 2025) as the frozen image editing prior for natural RGB readout. All readout

Table 1: Quantitative comparison on FMB, MSRS, and RoadScene. Best and second-best results are shown in **bold** and underlined, respectively.

	FMB						MSRS						RoadScene					
Method	CE↓	VIF↑	EN↑	SD↑	SF↑	AG↑	CE↓	VIF↑	EN↑	SD↑	SF↑	AG↑	CE↓	VIF↑	EN↑	SD↑	SF↑	AG↑
CDDFuse	1.82	0.49	6.78	37.03	14.56	4.31	0.99	0.51	6.70	43.37	11.55	3.75	0.87	0.41	7.47	56.21	18.83	6.75
DDFM	1.81	0.10	6.64	30.49	8.72	2.73	1.23	0.40	6.18	29.24	7.41	2.49	1.29	0.09	7.21	42.20	10.58	4.29
TIMFusion	2.07	0.38	6.28	26.30	10.90	3.23	1.76	0.36	5.76	40.32	9.39	2.80	0.55	0.37	6.83	37.27	8.99	3.39
Text-IF	1.75	0.52	6.74	34.36	14.35	4.42	1.02	**0.52**	6.73	44.59	11.88	3.88	0.81	0.43	7.38	49.99	16.05	6.47
C2RF	1.68	0.24	6.63	34.22	11.98	3.61	1.01	0.16	6.34	40.13	10.71	3.24	0.87	0.39	7.20	44.57	14.41	5.58
DCEvo	1.85	0.49	6.67	34.75	13.77	4.06	1.04	0.51	6.62	42.93	11.47	3.74	0.90	0.41	7.18	44.68	14.78	5.52
LUT-Fuse	1.75	0.43	6.55	33.42	15.27	3.75	1.26	0.47	6.57	42.19	12.66	3.81	0.78	0.40	6.89	45.12	19.47	5.05
ControlFusion	2.41	0.48	6.73	33.02	14.43	4.15	2.01	0.48	7.11	49.71	11.93	3.92	0.80	0.43	7.35	47.13	16.21	6.03
Deno-IF	2.05	0.40	5.90	30.78	9.93	2.97	1.34	0.45	6.66	41.09	9.66	3.11	0.90	0.38	7.34	47.02	12.84	5.05
APTP	1.91	0.44	6.59	33.04	11.94	3.53	**0.98**	0.47	6.57	41.36	10.19	3.24	1.06	0.42	7.46	52.66	13.81	5.20
DPOFusion	2.00	0.41	6.85	34.01	14.64	4.88	2.82	0.38	7.21	47.91	15.05	5.37	0.90	0.33	7.16	38.57	14.95	6.04
DRFusion	1.69	0.43	6.54	32.84	10.97	2.84	1.33	0.38	6.55	41.99	9.92	2.70	0.42	0.42	7.01	42.40	10.03	3.45
ISFuse	2.28	0.38	6.82	34.26	14.97	4.58	1.21	0.38	6.42	32.81	12.06	4.06	1.45	0.31	7.02	37.85	15.16	5.72
ReCoFuse	1.81	0.52	7.04	41.45	16.53	5.12	3.01	0.44	7.09	42.29	9.25	3.22	0.75	0.42	7.33	49.06	14.43	5.47
TeXIR$_{XC_v}$	**1.26**	**0.60**	7.29	50.21	19.75	**5.99**	1.14	0.49	6.89	47.01	11.34	3.76	1.25	0.38	7.43	54.77	14.37	5.32
TeXIR$_{XC_f}$	1.65	0.52	6.52	29.12	15.53	4.46	1.06	0.51	6.58	42.13	11.16	3.37	**0.35**	**0.45**	6.79	37.91	10.20	3.39
TeXIR$_{\text{day}}$	1.53	0.43	**7.43**	**64.50**	**20.00**	5.94	1.65	0.39	**7.49**	**59.72**	**21.91**	**7.65**	1.36	0.37	**7.48**	**60.10**	**21.06**	**7.13**

settings and prompts are fixed across the three benchmarks. Experiments are conducted on Hygon DCU K100. Appendix B provides the degradation settings, reliability supervision, optimization details, and readout instructions.

4.2 Comparison with Existing Methods

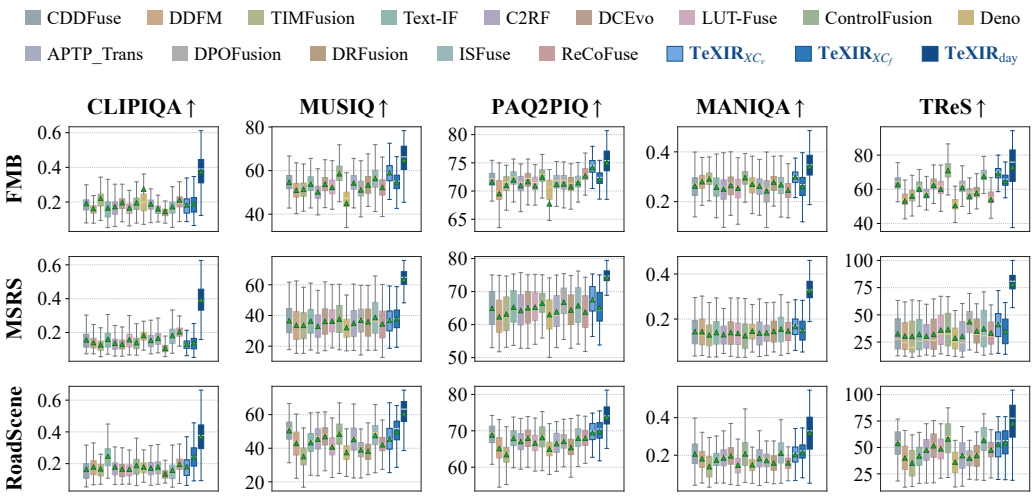

Figure 3: Perceptual quality across three benchmarks. TeXIR$_{\text{day}}$ achieves the highest score on all five perceptual metrics across FMB, MSRS, and RoadScene.

Quantitative comparison. Table 1 shows that TeXIR already performs strongly before the frozen image prior is applied. TeXIR$_{XC_v}$ outperforms all baselines on all six conventional metrics on FMB. Keeping the same reconstructed structure $\boldsymbol{X}_f$, replacing C_v with the completed chroma C_f improves CE and VIF on both MSRS and RoadScene, isolating the benefit of chroma completion. The daylight readout shows a different strength: it outperforms every baseline on EN, SD, SF, and AG across all three datasets and on every perceptual metric in Fig. 3. Together, these results support

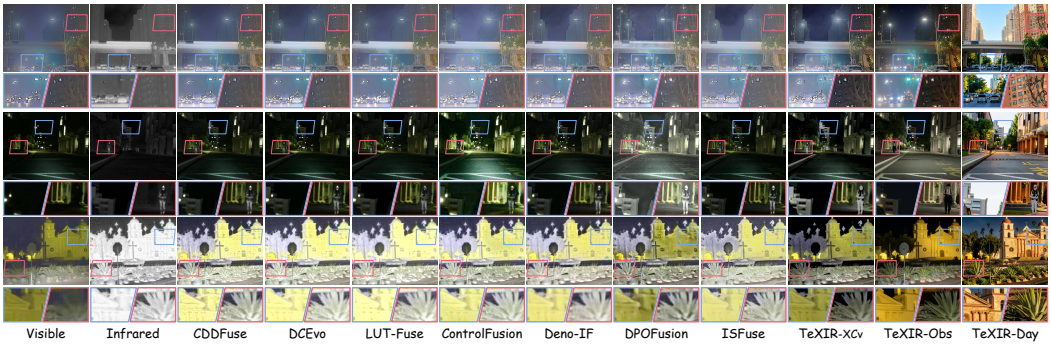

Figure 4: Qualitative comparison on FMB, MSRS, and RoadScene. TeXIR preserves infrared supported content while separating it from visual appearance. The observation and daylight readouts retain the same scene content under different appearance conditions.

the intended separation between state construction and image formation: the former determines what is preserved, while the latter controls how it is presented.

Qualitative comparison. Figure 4 shows the visual consequence of selection and appearance being coupled. Existing methods often preserve thermally salient targets by transferring infrared contrast into the fused image, which can alter visible brightness, color, and background appearance. TeXIR instead preserves these targets and structures while keeping their presentation under separate control. Across the three scenes, $\text{TeXIR}_{\text{obs}}$ maintains the observed illumination and color relationships, whereas $\text{TeXIR}_{\text{day}}$ changes the scene appearance substantially without changing object locations or structural content. These results visually support the central design of TeXIR: infrared information can determine what is preserved without determining how it must appear. Additional qualitative examples are provided in the supplementary material (Sec. F).

4.3 Controllable Readout

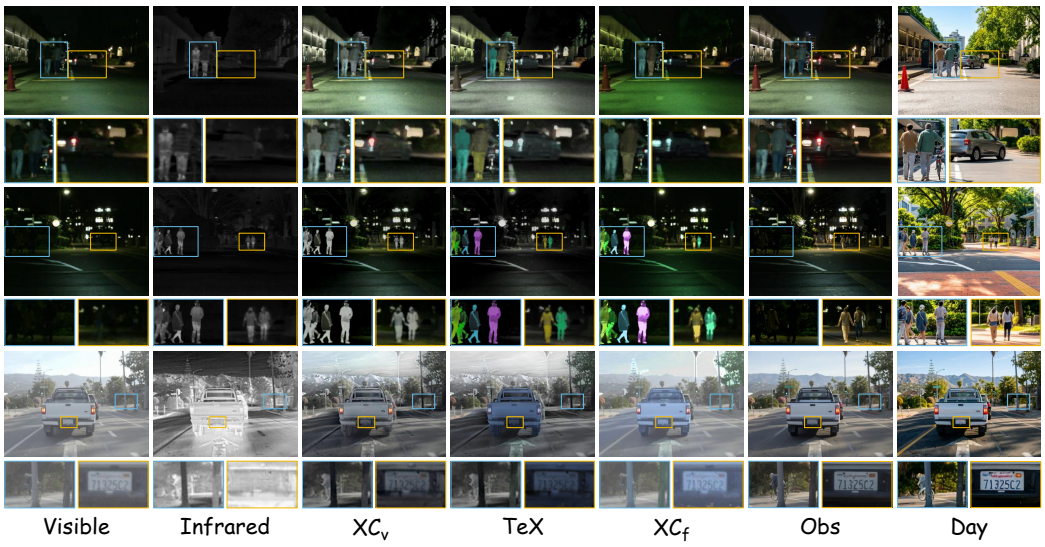

Figure 5: Controllable readout from one fused state. TeXIR reuses the same z_f to produce deterministic state views and natural RGB outputs. Object locations and structural details remain consistent as chroma and illumination change.

For each input pair, TeXIR constructs z_f once and reuses it for all readouts in Fig. 5. In the two nighttime scenes, I_{obs} retains the observed illumination, whereas I_{day} changes the scene to a daylight appearance; the pedestrians, vehicles, and local structures remain consistent across both outputs. The deterministic readouts expose the same constructed structure under visible chroma, fused chroma, and TeX state visualization. The TeX readout visualizes the fused state in HSV: a fixed mapping assigns the region identity $\boldsymbol{L}_f$ to hue, while $\boldsymbol{T}_f$ and $\boldsymbol{X}_f$ determine saturation and value,

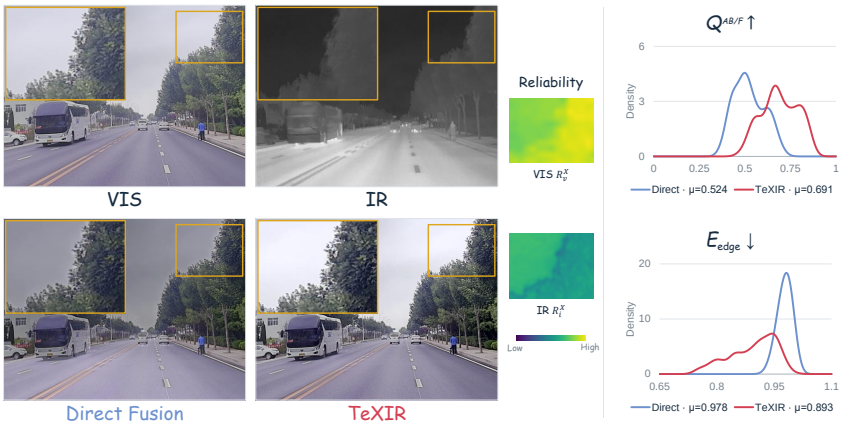

Figure 6: **Effect of integrability on structure fusion.** Direct Fusion and TeXIR use the same source structures and reliability maps. Direct blending introduces an artificial band along the tree canopy, whereas TeXIR reconciles the directional preferences before reconstruction and preserves a clean boundary. Right: distributions of $Q^{AB/F} \uparrow$ and $E_{\text{edge}} \downarrow$ on FMB.

respectively. This design adapts the original TeX convention $H = e$, $S = T$, and $V = X$, where hue distinguishes materials, saturation conveys temperature, and brightness carries texture, allowing the three states to occupy distinct visual roles in one image. As shown in Fig. 5, the TeX view makes thermal targets and fused structure explicit while preserving the spatial content shared by the other readouts. For the daytime example, the observation and daylight readouts remain closer in appearance, while the cyclist and license plate details persist across the readouts. These results show that TeXIR can change how a fused scene is presented without reconstructing what the fusion preserves.

4.4 Integrability in Structure Fusion

We isolate the role of integrability by comparing TeXIR with Direct Fusion using the same visible structure $\boldsymbol{X}_v$, infrared structure $\boldsymbol{X}_i$, and reliability maps. Direct Fusion blends the two structures at each location according to their relative reliability. TeXIR instead applies the same local preferences to directional derivatives, adjusts them for integrability, and then performs screened Poisson reconstruction. We evaluate edge preservation with $Q^{AB/F}$ and edge transfer error with E_{edge}.

Figure 6 shows the limitation of local reliability alone. When the blending preference changes spatially, direct fusion can introduce artificial gradients where the two source structures differ. This effect produces the visible band along the tree canopy. TeXIR avoids directly mixing the source structures and instead integrates adjusted directional derivatives, preserving a clean canopy boundary. Across FMB, mean $Q^{AB/F}$ increases from 0.524 to 0.691, while E_{edge} decreases from 0.978 to 0.893. Local reliability determines which source to trust, while integrability ensures that these local choices form a coherent structure.

5 Conclusion

We presented TeXIR, a visible and infrared image fusion framework built around an explicit TeX inspired scene state. TeXIR addresses the coupling between information selection and visual presentation by separating fused state construction from subsequent image formation. State specific reliability controls local modality contributions, integrability reconciles local structural choices before reconstruction, and a separate readout determines appearance after the fused state has been constructed. Experiments on FMB, MSRS, and RoadScene validate these contributions. The deterministic readouts preserve complementary source information, while the integrability study shows that reliable local choices alone cannot guarantee coherent structure. The controllable readout study further shows that the same fused state can support different illumination and color appearances without changing scene content. The overall results also demonstrate strong fusion performance across diverse scenes. State construction therefore determines what the fused scene preserves, while readout determines how that scene appears.

AI USE STATEMENT

In this work, we used generative AI tools for none of the tasks with required disclosure. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. The authors checked all AI-assisted edits to ensure that they preserved the intended technical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not raise any ethical concerns.

REPRODUCIBILITY STATEMENT

We provide the experimental settings and implementation details in Section 4.1 and Appendix B, with additional method details in Appendix A. Ablation studies and parameter analyses are reported in Appendix G. We will publicly release the source code upon acceptance to facilitate reproduction of the reported results.

REFERENCES

Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619 (7971):743–748, 2023.

Yanglin Deng, Tianyang Xu, Chunyang Cheng, Hui Li, Xiaojun Wu, and Josef Kittler. Beyond Strict Pairing: Arbitrarily Paired Training for High-Performance Infrared and Visible Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12334–12343, June 2026.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3209–3218, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Hui Li and Xiao-Jun Wu. DenseFuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. doi: 10.1109/TIP.2018.2887342.

Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *European Conference on Computer Vision*, 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 the IEEE/CVF International Conference on Computer Vision*, pp. 8115–8124, 2023.

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, Zengxi Zhang, Jiahao Zhang, Zhiying Jiang, Long Ma, Xin Fan, and Risheng Liu. Universal representation for real-world misaligned infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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, 2024.

Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information fusion*, 31:100–109, 2016.

Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. doi: 10.1016/j.inffus.2018.09.004.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. doi: 10.1109/JAS.2022.105686.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pp. 313–318, 2003.

Yu Shi, Yu Liu, Zhong-Cheng Wu, Juan Cheng, Huafeng Li, and Xun Chen. Degradation-Robust Fusion: An Efficient Degradation-Aware Diffusion Framework for Multimodal Image Fusion in Arbitrary Degradation Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026.

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, June 2026.

Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23199–23209, 2025.

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, 2022. doi: 10.1016/j.inffus.2021.12.004.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. ControlFusion: A controllable image fusion network with language-vision degradation prompts. In *Advances in Neural Information Processing Systems*, 2025a.

Linfeng Tang, Qinglong Yan, Xinyu Xiang, Leyuan Fang, and Jiayi Ma. C2RF: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International Journal of Computer Vision*, 133(8):5262–5280, 2025b. doi: 10.1007/s11263-025-02427-1.

Jianyi Wang, Kelvin C. K. 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. doi: 10.1609/aaai.v37i2.25353.

Song Wang, Xie Han, Liqun Kuang, Boying Wang, Zhongyu Chen, Zherui Qiao, Fan Yang, Xiaoxia Liu, Bingyu Zhang, and Zhixun Wang. The source image is the best attention for infrared and visible image fusion. In *2025 IEEE/CVF International Conference on Computer Vision*, pp. 13513–13522. IEEE, 2025.

Wu Wang, Liang-Jian Deng, Qi Cao, and Gemine Vivone. A general image fusion approach exploiting gradient transfer learning and fusion rule unfolding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026a.

Xue Wang, Zheng Guan, Wenhua Qian, Chengchao Wang, and Runzhuo Ma. Multi-modal image fusion via intervention-stable feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33827–33837, June 2026b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

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, 2022. doi: 10.1109/TPAMI.2020.3012548.

Han Xu, Yuyang Li, Yunfei Deng, Jiayi Ma, and Guangcan Liu. Deno-IF: Unsupervised noisy visible and infrared image fusion method. In *Advances in Neural Information Processing Systems*, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1190–1199, 2022.

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. 27026–27035, 2024.

Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. LUT-Fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14559–14568, 2025.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3585, 2020.

Hao Zhang, Lei Cao, and Jaiyi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

Hao Zhang, Shuhan Yang, Linfeng Tang, Xunpeng Yi, and Jiayi Ma. ReCoFuse: Ultra-Robust Image Fusion via Restorative Multi-Modal Diffusion Reciprocal Coupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33838–33847, June 2026.

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.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pp. 970–976. ijcai.org, 2020.

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 the IEEE/CVF International Conference on Computer Vision*, pp. 8082–8093, 2023b.

A Additional Method Details

1. State observations. We construct the visible and infrared state observations with fixed operators. For the visible image $I_v = (I_v^R, I_v^G, I_v^B)$, we derive structure from luminance and retain chroma separately. For the infrared image I_i, we normalize its response and derive the structural and thermal states:

$$\begin{aligned} \boldsymbol{Y}_v &= 0.299 I_v^R + 0.587 I_v^G + 0.114 I_v^B, \\ \boldsymbol{X}_v &= \boldsymbol{Y}_v + \lambda\big(\boldsymbol{Y}_v - \mathcal{G}_{\rho_v}(\boldsymbol{Y}_v)\big), \\ C_v &= \left[I_v^R - I_v^B,\ 2I_v^G - I_v^R - I_v^B\right]^\top, \\ \boldsymbol{J}_i &= \frac{I_i - \mathrm{perc}_1(I_i)}{\mathrm{perc}_{99}(I_i) - \mathrm{perc}_1(I_i) + \epsilon}, \\ \boldsymbol{X}_i &= \alpha_1 \boldsymbol{J}_i + \alpha_2\big(\boldsymbol{J}_i - \mathcal{G}_{\rho_i}(\boldsymbol{J}_i)\big), \\ \boldsymbol{T}_i &= \boldsymbol{J}_i. \end{aligned} \tag{10}$$

Here, $\mathcal{G}_\rho$ denotes Gaussian smoothing, and $\mathrm{perc}_p(I_i)$ denotes the p-th percentile of the infrared intensities. The thermal state $\boldsymbol{T}_i$ represents relative thermal response.

2. Reliability supervision. We train the reliability estimator with spatially varying synthetic degradations. Visible degradations include blur, noise, color distortion, ISP artifacts, low illumination, overexposure, occlusion, and modality failure. Infrared degradations include blur, response drift, saturation, nonuniformity residuals, range compression, noise, defective pixels, thermal blooming, and modality failure.

Let $\boldsymbol{d}_m(\boldsymbol{p})$ collect the degradation strengths of modality m at location $\boldsymbol{p}$. Since TeXIR predicts only visible structure, visible chroma, and infrared structure reliability, we define the supervised index set $\mathcal{I}_R = \{(v, X), (v, C), (i, X)\}$. For each $(m, s) \in \mathcal{I}_R$, we construct the reliability target as

$$\boldsymbol{R}_m^{s,\mathrm{tar}}(\boldsymbol{p}) = \exp\left[-(\boldsymbol{a}_m^s)^\top \boldsymbol{d}_m(\boldsymbol{p})\right], \qquad (m, s) \in \mathcal{I}_R, \tag{11}$$

where $\boldsymbol{a}_m^s$ specifies the sensitivity of state s to each degradation in modality m. This mapping allows the same degradation to affect the supervised states differently.

We train the three reliability maps with a weighted L_1 loss:

$$\mathcal{L}_R = \sum_{(m,s)\in\mathcal{I}_R} \frac{\sum_{\boldsymbol{p}} \left[1 + \beta\left(1 - \boldsymbol{R}_m^{s,\mathrm{tar}}(\boldsymbol{p})\right)\right] \left|\boldsymbol{R}_m^s(\boldsymbol{p}) - \boldsymbol{R}_m^{s,\mathrm{tar}}(\boldsymbol{p})\right|}{\sum_{\boldsymbol{p}} \left[1 + \beta\left(1 - \boldsymbol{R}_m^{s,\mathrm{tar}}(\boldsymbol{p})\right)\right]}. \tag{12}$$

The weighting emphasizes strongly degraded regions. After synthetic training, we manually annotate $\boldsymbol{R}_v^X$, $\boldsymbol{R}_v^C$, and $\boldsymbol{R}_i^X$ for 100 visible and infrared pairs from FMB and use these annotations to fine tune the estimator on real observations.

3. Directional infrared support. The support map $\boldsymbol{M}_d$ in Sec. 3.2 suppresses isolated infrared fluctuations before they enter the directional structure preferences. For each direction $d \in \{x, y\}$, we smooth the infrared structure observation and identify derivatives with sufficient magnitude:

$$\boldsymbol{B}_{i,d} = \mathbb{1}\left[|\nabla_d \mathcal{G}_{\rho_i}(\boldsymbol{X}_i)| > \tau_i\right], \qquad d \in \{x, y\}. \tag{13}$$

Here, ρ_i is the Gaussian smoothing scale and τ_i is the minimum directional gradient magnitude required for infrared structural support. We remove connected components smaller than $n_{\min}$ from each $\boldsymbol{B}_{i,d}$ and use the remaining components to form $\boldsymbol{M}_d$. The resulting support maps specify where infrared derivatives may enter the initial directional preferences. Reliability determines their relative contribution, while the subsequent integrability optimization jointly adjusts the horizontal and vertical preferences.

4. Controllable readout. The readout in Section 3.4 operates on the fixed fused state z_f and does not modify state construction. We first construct three deterministic forms of this state and then use them to produce natural RGB outputs.

The XC_v and XC_f forms combine the same fused structure with visible and completed chroma, respectively:

$$\begin{aligned} I_{XC_v} &= \mathcal{C}_{\mathrm{rgb}}(\boldsymbol{X}_f, C_v), \\ I_{XC_f} &= \mathcal{C}_{\mathrm{rgb}}(\boldsymbol{X}_f, C_f). \end{aligned} \tag{14}$$

They therefore contain the same structural content and differ only in chroma.

The TeX form visualizes the structural, region, and thermal components of the fused state:

$$I_{\mathrm{TeX}} = \mathrm{HSV2RGB}\big(h(\boldsymbol{L}_f), \boldsymbol{T}_f, \boldsymbol{X}_f\big), \tag{15}$$

where $h(\cdot)$ assigns each region index a fixed hue. The hue only distinguishes spatial regions; it does not encode semantic class, material identity, or natural object color. The saturation visualizes relative thermal response rather than calibrated temperature. Thus, the TeX form provides an interpretable state visualization, whereas XC_v and XC_f provide natural-color forms for subsequent image formation.

For the observation readout, we mix I_{XC_v} and I_{XC_f} in YCbCr space with fixed weights 0.65 and 0.35. Because both forms share $\boldsymbol{X}_f$, this operation keeps the constructed structure fixed while combining observed and completed chroma. We then provide the mixed image together with I_v to the frozen image editing prior. The mixed image defines the fused scene content, whereas I_v serves only as an appearance reference for the observed illumination, exposure, and reliable visible colors. The prior therefore changes the presentation without revisiting the fusion decision.

For the daylight readout, we provide I_{XC_v} and I_{XC_f} directly to the same frozen prior. The two inputs again share the same fused structure: XC_v supplies the observed visible-color tendency, while XC_f supplies completed chroma where visible color is weak or missing. A fixed daylight instruction changes illumination, exposure, white balance, and local appearance while preserving the scene structure, object identities, and spatial layout.

We use fixed instructions for both natural RGB readouts. The observation instruction preserves the constructed scene while matching its presentation to the current visible-light condition. The daylight instruction preserves the same scene while rendering it under a natural daylight condition. Neither instruction permits adding, removing, moving, resizing, or reinterpreting scene objects.

B Implementation Details

Observation and reliability settings. We set $\rho_v = \rho_i = 1.2$, $\lambda = 1e-4$, and $(\alpha_1, \alpha_2) = (0.01, 0.03)$. For reliability training, we sample easy, medium, and hard cases with probabilities 0.20, 0.70, and 0.10 from severity ranges $[0.05, 0.35]$, $[0.35, 0.75]$, and $[0.75, 1.0]$, respectively. We set $\beta = 3$ in Eq. equation 12. After synthetic training, we manually annotate $\boldsymbol{R}_v^X$, $\boldsymbol{R}_v^C$, and $\boldsymbol{R}_i^X$ for 100 visible and infrared pairs from FMB and use these annotations to fine tune the estimator on real observations.

Structure optimization. For directional infrared support, we use a Gaussian smoothing scale of 1.0, a gradient threshold of 0.10, and a minimum component size of eight pixels. We initialize the directional weights with $\boldsymbol{W}_d^0$ and project each update to $[0, 1]$. The optimization runs for at most 1000 iterations and stops when the relative change falls below $1e-4$. We solve the screened Poisson system with preconditioned conjugate gradients using at most 200 iterations and a relative tolerance of $1e-5$.

Readout. The region model and image editing prior remain frozen in all experiments. We use the same SAM3 model across the three datasets and Qwen-Image-Edit (Wu et al., 2025) for natural RGB readout. For I_{obs}, we set the I_{XC_v} mixture coefficient to $\gamma = 0.65$. Qwen-Image-Edit uses 40 inference steps, `true_cfg_scale`$= 4.0$, and `guidance_scale`$= 1.0$. We keep all readout settings and prompts fixed across FMB, MSRS, and RoadScene.

For the observation readout, we use:

> Image 1: authoritative fusion target and output canvas. It alone defines the framing, viewpoint, complete scene content, exact geometry, positions, silhouettes, identities, layout, visibility, and recoverable detail, including completed, small, distant, peripheral, seated, and low-visibility elements.

Image 2: visible-light appearance reference only. It provides the current illumination, exposure, and partially reliable color evidence.

Priority: Any conflict in content, geometry, or visibility must follow Image 1. Missing or weak evidence in Image 2 never cancels content in Image 1.

Task: Convert Image 1 into a clear, realistic visible-light image under the current illumination conditions of Image 2, without changing or omitting any scene content from Image 1.

Keep from Image 1: every depicted object and structure, even if faint, only outlined, or absent or unclear in Image 2.

Allow from Image 1: naturalize its rendering only, while keeping its content and geometry fixed.

Do not from Image 1: treat any depicted structure as noise; add, delete, merge, move, resize, reshape, replace, reinterpret, or change object identity.

Keep from Image 2: the current illumination, exposure, and clearly supported visible-light colors.

Allow from Image 2: use denoised appearance evidence, locally harmonize illumination and color, and infer plausible colors only where its color evidence is weak or missing.

Do not from Image 2: use it as the output canvas or to define, remove, reshape, reposition, resize, merge, or hide scene content; globally relight the scene, change the global exposure, or change the time of day.

For the daylight readout, we use:

Image 1: authoritative scene canvas with visible chroma. It fixes the framing, viewpoint, scene content, geometry, object positions, identities, boundaries, and structural detail.

Image 2: aligned fused-chroma reference. It represents the same scene and the same structure as Image 1, but provides completed chroma where the visible color information is weak or missing.

Task: Naturalize Image 1 into a vivid, realistic daylight-oriented image.

Keep: objects, positions, shapes, identities, layout, and supported completed content from Image 2; material boundaries, color tendency, texture, and structure from Image 1

Allow: vivid natural material color, local illumination coordination, and color-transition harmonization

Do not: reinterpret the scene or redesign objects; add, delete, move, change object identity, alter material boundaries, or create new targets

C SEMANTIC SEGMENTATION

Table 2: Semantic segmentation on FMB using CLIPer (Sun et al., 2025). Best and second-best results are shown in bold and underlined, respectively.

Method	Venue	$\mathrm{mIoU_S}$	$\mathrm{mAcc_S}$	Road		Building		Sky		Car		Truck		Bus		Veg.		Pole	
				IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc
CDDFuse (Zhao et al., 2023a)	CVPR'23	0.510	0.632	0.560	0.811	0.575	0.700	0.679	0.692	0.631	0.884	0.450	0.499	0.622	0.837	0.465	0.474	0.100	0.160
DDFM (Zhao et al., 2023b)	ICCV'23	0.472	0.588	0.536	0.724	0.585	0.709	0.692	0.704	0.555	0.786	0.269	0.343	0.565	0.797	0.485	0.497	0.089	0.144
TIMFusion (Liu et al., 2024)	TPAMI'24	0.511	0.627	0.560	0.798	0.555	0.671	0.607	0.621	0.633	0.896	0.487	0.551	0.676	0.839	0.472	0.482	0.100	0.157
Text-IF (Yi et al., 2024)	CVPR'24	0.513	0.637	0.565	0.807	0.579	0.706	0.675	0.687	0.641	0.888	0.439	0.518	0.643	0.864	0.471	0.481	0.095	0.147
C2RF (Tang et al., 2025b)	IJCV'25	0.482	0.605	0.534	0.781	0.573	0.703	0.607	0.615	0.607	0.865	0.410	0.485	0.607	0.820	0.443	0.450	0.080	0.124
DCEvo (Liu et al., 2025)	CVPR'25	0.492	0.623	0.550	0.806	0.568	0.698	0.688	0.702	0.628	0.888	0.385	0.475	0.567	0.793	0.454	0.463	0.099	0.157
LUT-Fuse (Yi et al., 2025)	ICCV'25	0.491	0.615	0.512	0.756	0.566	0.695	0.667	0.679	0.605	0.882	0.379	0.414	0.621	0.852	0.476	0.487	0.100	0.157
Deno-IF (Xu et al., 2025)	NeurIPS'25	0.432	0.524	0.486	0.663	0.484	0.572	0.558	0.566	0.610	0.831	0.424	0.466	0.392	0.532	0.410	0.416	0.092	0.146
ControlFusion (Tang et al., 2025a)	NeurIPS'25	0.506	0.632	0.563	0.804	0.576	0.706	0.693	0.707	0.629	0.887	0.412	0.473	0.619	0.856	0.462	0.472	0.096	0.152
APTP (Deng et al., 2026)	CVPR'26	0.505	0.630	0.542	0.805	0.573	0.703	0.684	0.694	0.623	0.885	0.444	0.500	0.617	0.837	0.462	0.472	0.092	0.144
DPOFusion (Su et al., 2026)	CVPR'26	0.484	0.616	0.556	0.808	0.574	0.700	0.646	0.659	0.635	0.865	0.338	0.458	0.584	0.837	0.443	0.451	0.096	0.155
DRFusion (Shi et al., 2026)	CVPR'26	0.458	0.591	0.530	0.790	0.555	0.681	0.658	0.678	0.604	0.902	0.283	0.351	0.505	0.730	0.423	0.431	**0.105**	**0.169**
ISFuse (Wang et al., 2026b)	CVPR'26	0.507	0.618	0.538	0.769	0.577	0.703	0.634	0.641	0.633	0.875	0.456	0.503	0.668	0.838	0.450	0.459	0.096	0.152
ReCoFuse (Zhang et al., 2026)	CVPR'26	0.513	0.635	0.572	0.811	0.571	0.695	0.696	0.712	0.637	0.894	0.381	0.440	0.679	0.897	0.465	0.476	0.099	0.155
TeXIR (TeX)	Ours	0.531	0.665	0.600	0.835	0.576	0.718	**0.735**	**0.749**	0.653	0.902	0.460	0.608	0.686	0.888	0.442	0.454	0.098	0.164
TeXIR (XC_v)	Ours	0.527	0.655	0.601	0.834	0.567	0.692	0.701	0.718	0.648	0.899	0.472	0.567	0.658	0.894	0.463	0.473	0.101	0.163
TeXIR (XC_f)	Ours	0.536	0.661	0.610	0.833	0.579	0.706	0.707	0.718	0.648	0.894	0.477	0.571	0.682	0.907	0.480	0.491	0.102	0.166
TeXIR (obs)	Ours	0.553	0.688	0.617	0.850	0.581	0.719	0.716	0.727	0.664	0.904	**0.504**	**0.635**	**0.704**	**0.950**	0.541	0.556	0.098	0.165
TeXIR (day)	Ours	**0.560**	**0.689**	**0.638**	**0.855**	**0.595**	**0.726**	0.734	0.745	**0.668**	**0.920**	0.483	0.593	0.703	0.930	**0.562**	**0.578**	0.100	0.167

We evaluate whether TeXIR outputs retain information useful for semantic segmentation. On the FMB test set, we apply CLIPer (Sun et al., 2025) to every fused image. The FMB category names listed in Table 2 serve as text prompts, and we compare the predicted masks with the pixel annotations. CLIPer is frozen and kept identical across all methods, and the five TeXIR readouts are evaluated separately.

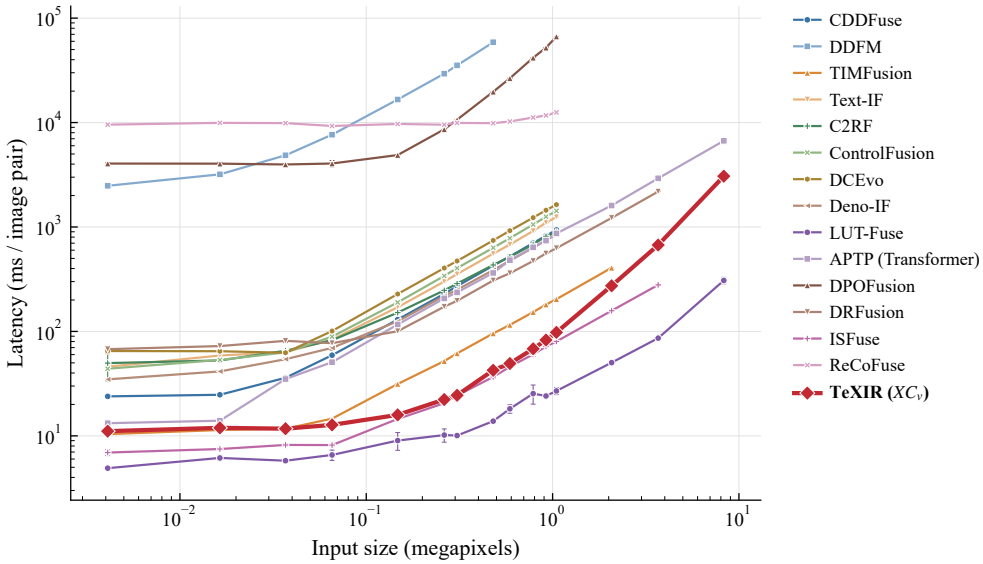

Figure 7: Inference latency across input resolutions.

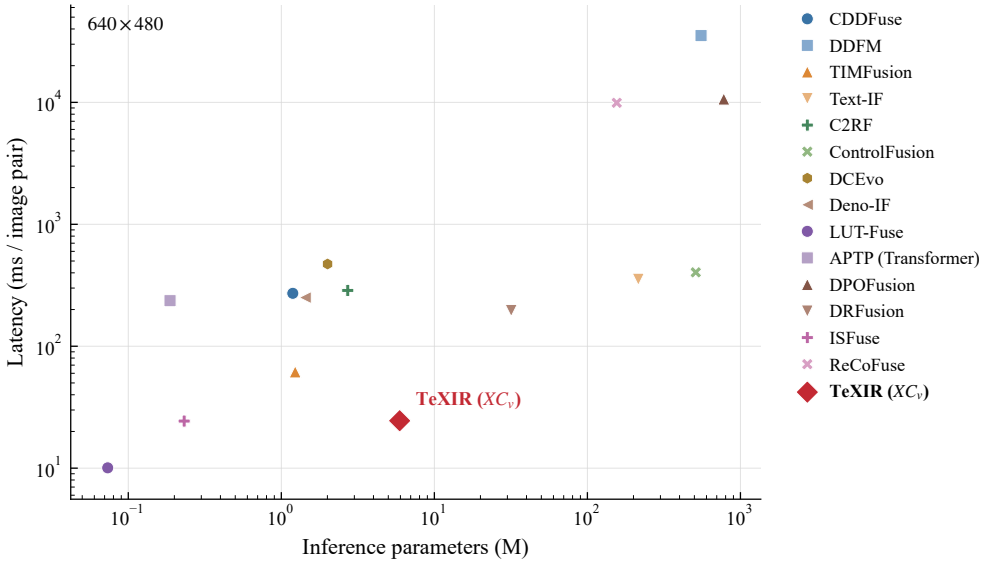

Figure 8: Model size and inference latency at 640×480. Each point represents one fusion method.

Table 2 shows that all three deterministic TeXIR readouts outperform the compared fusion methods in both mean IoU and mean accuracy. Their mIoU ranges from 0.527 to 0.536, above the best baseline score of 0.513, and their mAcc also exceeds the best baseline score of 0.637. This result shows that the fused state itself preserves strong semantic information. The two natural RGB readouts further improve segmentation performance while reusing the same state, reaching mIoU values from 0.553 to 0.560 and mAcc values from 0.688 to 0.689. These results support the intended separation: state construction preserves useful scene information, while readout improves its accessibility through visual presentation.

D Computational Efficiency

We evaluate runtime across input resolutions and compare model size with latency at 640×480. All measurements use an NVIDIA RTX GPU with batch size 1. For TeXIR, we report the XC_v readout. Figure 7 shows that TeXIR maintains low latency as the input resolution increases. At 640×480, TeXIR requires 24.49 ms per image pair, close to ISFuse at 24.34 ms and substantially faster than

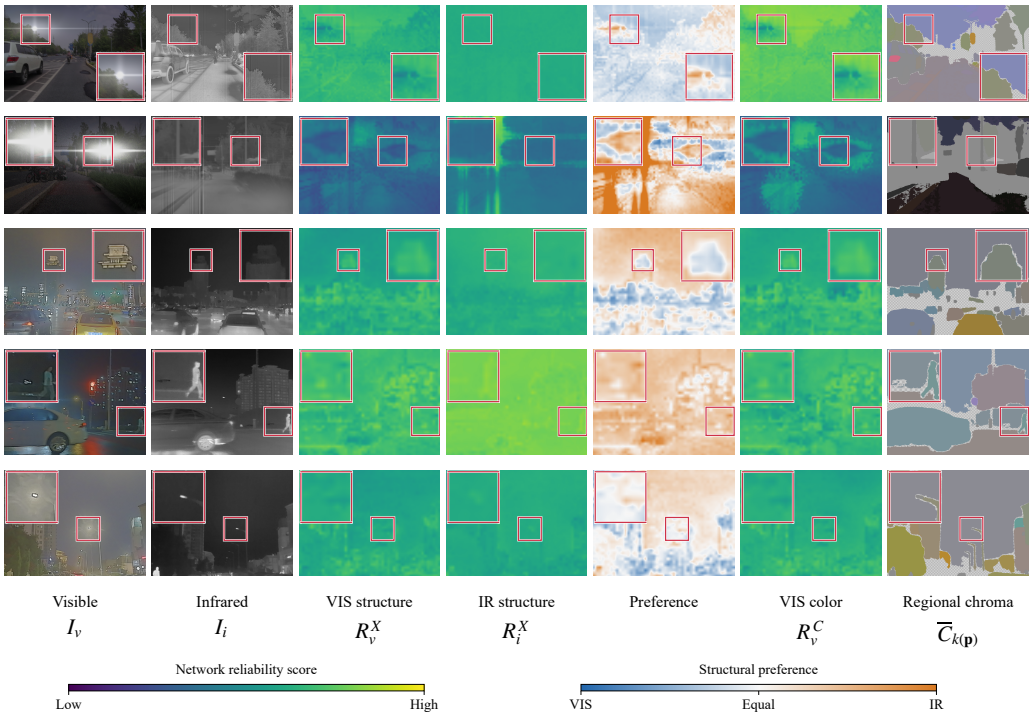

Figure 9: State specific reliability for scene state construction. The first two rows show synthetic degradations with the pretrained estimator, and the remaining rows show FMB examples with the fine tuned estimator. Columns show the input pair, structure reliabilities, structural preference, visible chroma reliability, and regional chroma reference $\overline{C}_{k(\boldsymbol{p})}$. Blue and orange indicate preference for visible and infrared structure, respectively, while white indicates comparable support.

most competing methods. At 1920×1080, TeXIR remains executable at 273.52 ms, whereas nine compared methods exceed the memory budget. Only LUT-Fuse and ISFuse remain consistently faster across the tested resolutions. LUT-Fuse achieves its particularly low latency through lookup table inference, a design explicitly aimed at efficient deployment. In contrast, TeXIR performs learned state prediction and reconstruction while retaining competitive runtime across resolutions.

Figure 8 shows that inference time depends on more than parameter count. TeXIR contains 5.934M parameters, yet several smaller models, including CDDFuse, C2RF, DCEvo, and Deno-IF, require much longer inference time. Together with the fusion results reported in the main experiments, these measurements show that TeXIR retains strong fusion performance while keeping inference cost low.

E Reliability for Scene State Construction

Figure 9 shows how reliability controls local state construction. In the synthetic examples, glare and blooming reduce visible structure reliability, and the structural preference shifts toward infrared where infrared retains clearer structure. The real FMB examples show the same local adaptation: visible structure receives stronger support around the illuminated tower, whereas infrared dominates around the pedestrian in the dark region. Notably, the tower example is not used during fine tuning, yet the estimator produces the same reliability driven behavior.

Structure and chroma follow separate reliability decisions. A region may favor infrared for structure while its chroma reference is still estimated only from visible observations with reliable color support. The last two columns make this distinction explicit: $\boldsymbol{R}_v^C$ determines which visible colors contribute to $\overline{C}_k$, independently of the structural preference. TeXIR can therefore recover structure from infrared without interpreting infrared intensity as visible color.

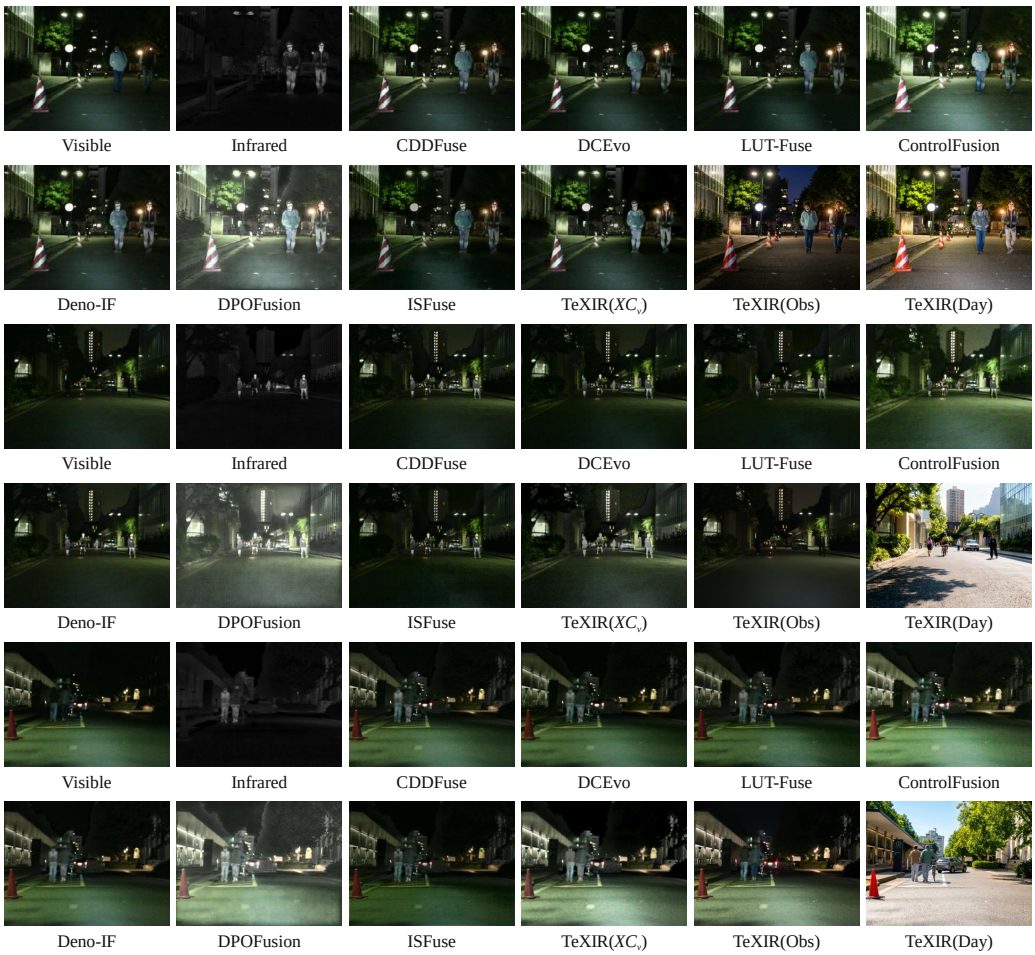

Figure 10: Qualitative comparisons on MSRS. MSRS is unseen during training. Three nighttime scenes are shown.

F Generalization to Unseen Datasets

We evaluate TeXIR on MSRS and RoadScene, neither of which is used for training. Figures 10 and 11 show nighttime pedestrian scenes from MSRS and urban scenes under different illumination conditions from RoadScene. Across both datasets, TeXIR preserves salient targets and scene structure while producing the intended differences among its readouts. Pedestrians and vehicles remain visible under challenging illumination, while the day readout reveals more of the surrounding scene without changing the underlying content. These results show that the learned state construction and readout behavior transfer to unseen datasets and illumination conditions.

G Ablation Study

Structure fusion design. Table 3 examines the roles of reliability and integrability in structure fusion. Removing reliability causes a substantial drop across all metrics, particularly in VIF and SF, showing that local modality contributions must adapt to the input rather than remain uniform. Removing integrability also consistently degrades performance. Together, these results show that the two components serve complementary roles: reliability determines which modality provides stronger local structural support, while integrability reconciles these local decisions before reconstruction.



Figure 11: Qualitative comparisons on RoadScene. Examples cover both daytime and nighttime scenes.

Table 3: **Ablation and parameter sensitivity on FMB.** $\star$ denotes the default setting.

Study	Setting	CE↓	VIF↑	EN↑	SD↑	SF↑	AG↑
Design	w/o integrability	1.296	0.572	7.181	47.707	18.205	5.351
	w/o reliability	1.319	0.501	7.212	46.031	15.920	4.992
	Full	**1.262**	**0.595**	**7.289**	**50.208**	**19.752**	**5.992**
α_2	0	1.263	0.587	7.273	49.724	19.295	5.838
	0.003	1.265	0.592	7.277	49.890	19.485	5.878
	0.0075	1.265	0.595	7.280	50.022	19.591	5.913
	0.015	1.264	0.596	7.284	50.136	19.674	5.948
	0.03*	1.262	0.595	7.289	50.208	19.752	5.992
	0.06	1.256	0.586	7.292	50.086	19.805	6.047
	0.12	1.252	0.569	7.291	49.800	19.890	6.144
	0.3	1.251	0.540	7.269	49.078	20.015	6.293
	1	1.251	0.510	7.222	48.305	19.971	6.355
λ	0	1.273	0.598	7.337	49.877	19.694	5.969
	10^{-6}	1.272	0.598	7.330	49.783	19.699	5.970
	10^{-5}	1.285	0.598	7.299	49.455	19.723	5.980
	$10^{-4\star}$	1.262	0.595	7.289	50.208	19.752	5.992
	10^{-3}	1.254	0.593	7.300	51.548	19.767	5.998
	10^{-2}	1.258	0.597	7.291	52.503	19.828	6.014
	10^{-1}	1.279	0.642	7.269	53.476	20.252	6.137
	10^{0}	1.330	0.733	7.249	54.855	22.054	6.671
	10^{1}	1.490	0.763	7.131	55.803	24.273	7.232

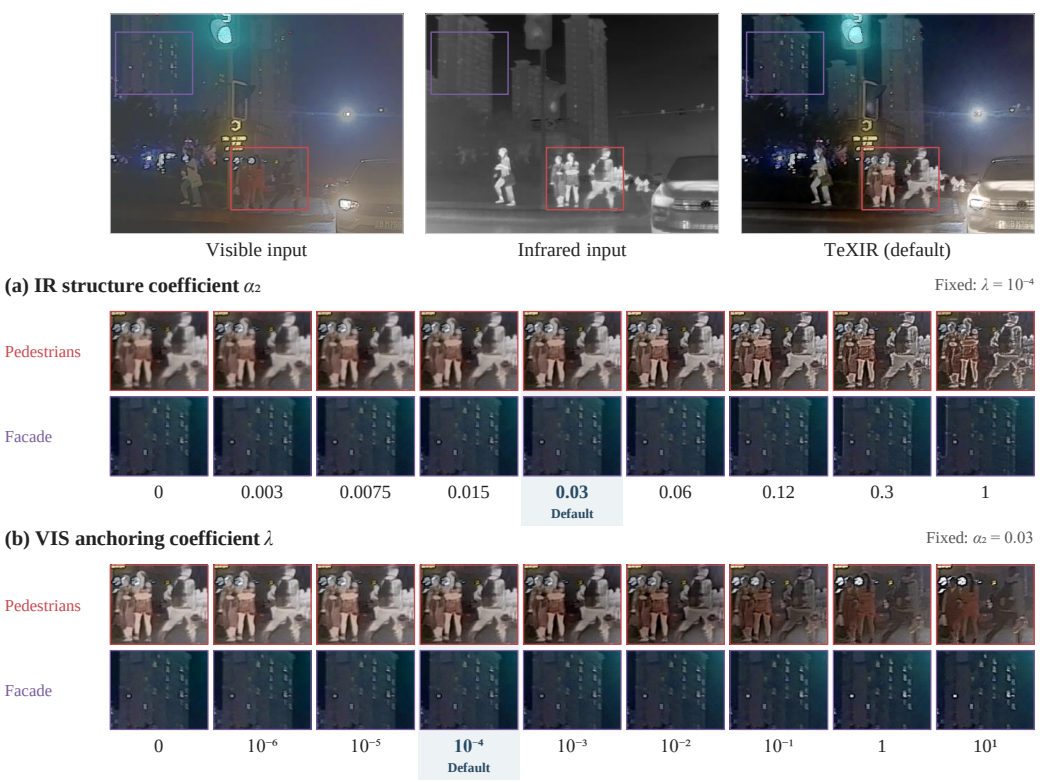

Figure 12: Visual analysis of the structure fusion parameters. α_2 controls infrared structure strength, while λ controls visible anchoring.

Parameter analysis. Table 3 and Fig. 12 further examine the IR structure coefficient α_2 and the VIS high-frequency enhancement coefficient λ. Increasing α_2 progressively strengthens the infrared structure observation. SF and AG increase accordingly, whereas VIF peaks at moderate values and declines as α_2 becomes large. The visual results show the same transition: moderate values recover clearer pedestrian and background details, while large values introduce exaggerated structural responses. We use $\alpha_2 = 0.03$ to retain informative infrared structure without excessive enhancement.

The coefficient λ controls the high-frequency enhancement of the visible structure observation. Small values leave more freedom for complementary infrared structure, whereas large values constrain the reconstruction more strongly toward the visible observation. At large values of λ, VIF, SD, SF, and AG increase, while CE and EN deteriorate; the visual results also show weaker infrared supported pedestrian structure. We therefore use $\lambda = 10^{-4}$ to retain a visible anchor without suppressing complementary infrared structure. The opposing metric trends also show that selecting either parameter by a single fusion metric would favor an extreme operating point rather than a balanced reconstruction.

State Construction and Natural Image Rendering. We further study how the readout affects the use of a generative prior. All methods use the same Qwen-Image-Edit-2511 backend, the same prompt, and the same inference settings. Each method also provides two conditioning images. For every baseline, we duplicate its fused image and use the two copies as input. TeXIR instead provides two complementary readouts from the same fused state. One emphasizes scene structure, while the other carries the intended visual appearance. This setting keeps the generation process unchanged and isolates how the readout presents fused information to the generative prior.

Figure 13 shows that the final result depends strongly on this guidance. In the first scene, baseline results carry bright pedestrian responses into the daylight image and produce unnatural local appearance. TeXIR preserves the pedestrian structure while rendering the people with a more coherent daylight appearance. The second scene reveals a different failure mode. Several baseline results become more photographic after generation but weaken the license plate patterns already present in their fused images. TeXIR retains a clearer plate region while producing a natural daylight result.

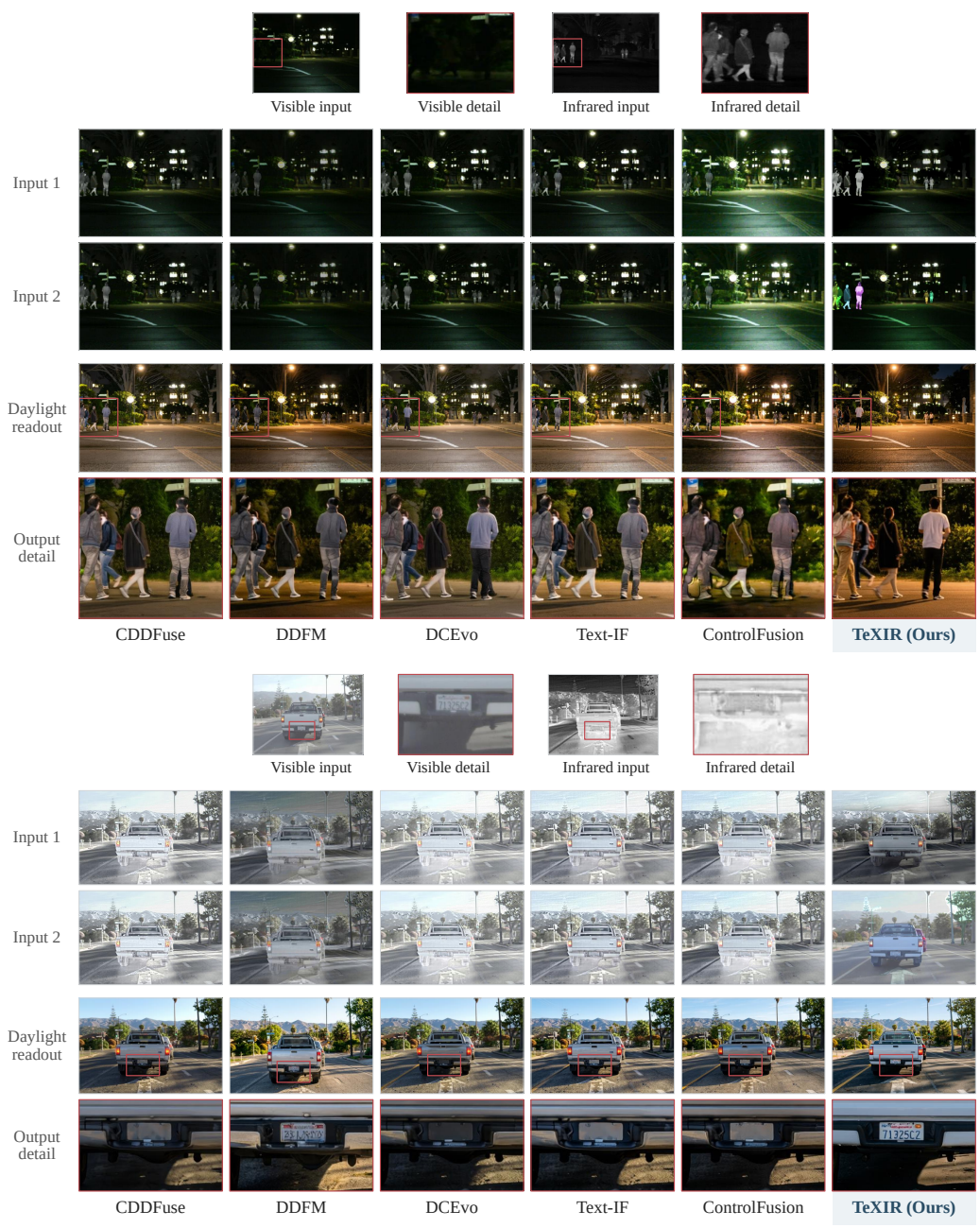

Figure 13: Complementary readouts for generative image rendering.

These examples explain the role of the two TeXIR readouts. A conventional fused image combines structural information and visual appearance in a single representation, so the generative model must infer which properties to preserve and which properties to change. TeXIR separates these roles before generation. The structure readout guides content preservation, and the appearance readout guides visual rendering. This division allows the generative prior to focus on image formation while the fused state continues to govern the scene content.

Effect of Integrability on Structure Reconstruction. We use a synthetic example to isolate the effect of integrability refinement. We keep the structure observations, initial directional preferences, TV term, and screened Poisson reconstruction fixed. The initial directional preferences are constant and have zero TV, but the target derivatives can still violate integrability. We therefore change only

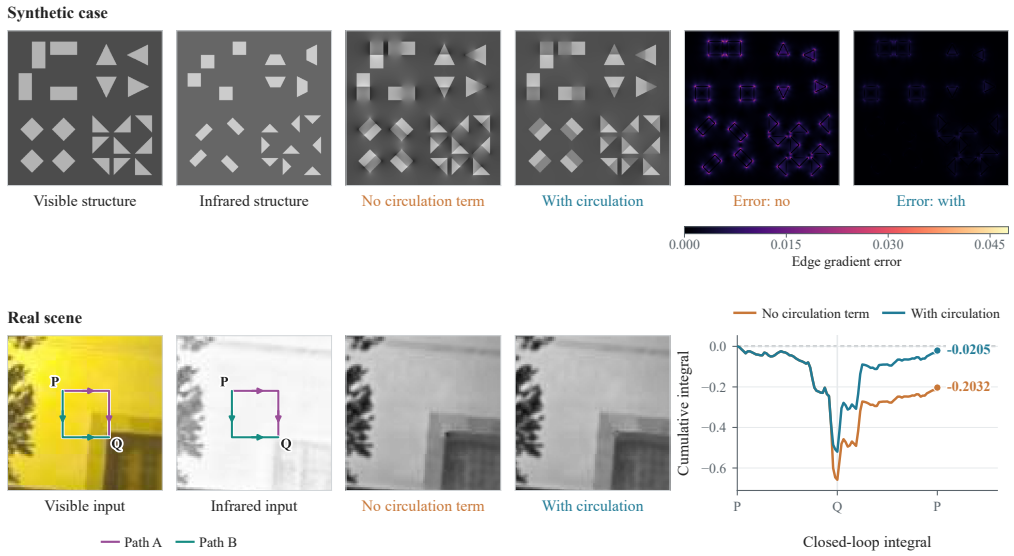

Figure 14: Effect of integrability refinement on structure reconstruction.

the circulation term while keeping the same Poisson solver. Both settings use the same Poisson solver. Without refinement, stronger local circulation forces the reconstruction to compromise between the target derivatives, producing errors near structure boundaries and gradients in flat regions. Integrability refinement reduces these violations before reconstruction.

The synthetic example makes the benefit visually clear. Without integrability refinement, screened Poisson reconstruction introduces shading and streaks around structure boundaries, together with unwanted gradients in flat regions. After refinement, the background becomes flatter and the reconstructed boundaries better follow the input structures. The error maps show the same reduction around edges present in both structure observations. This result shows that reducing integrability violations before reconstruction directly reduces structural distortions in the reconstructed image. The real example in Fig. 14 provides an additional real image observation. The marked wall corner defines two paths between the same pair of points. For target derivatives that better satisfy integrability, the accumulated signed changes along these two paths should be closer. We therefore measure the target derivatives before screened Poisson reconstruction. The difference between the two path accumulations decreases from 0.2032 to 0.0205 after integrability refinement. The corresponding reconstruction also shows a cleaner transition around the corner.

Integrability refinement reduces violations in the target derivatives before reconstruction. This allows screened Poisson reconstruction to match the horizontal and vertical targets more closely and recover the fused structure with fewer artifacts.