WHEN MOTION MISLEADS: CAUSAL-ROLE-AWARE MOTION-TURBULENCE DISENTANGLEMENT FOR TURBULENT INFRARED VIDEO SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

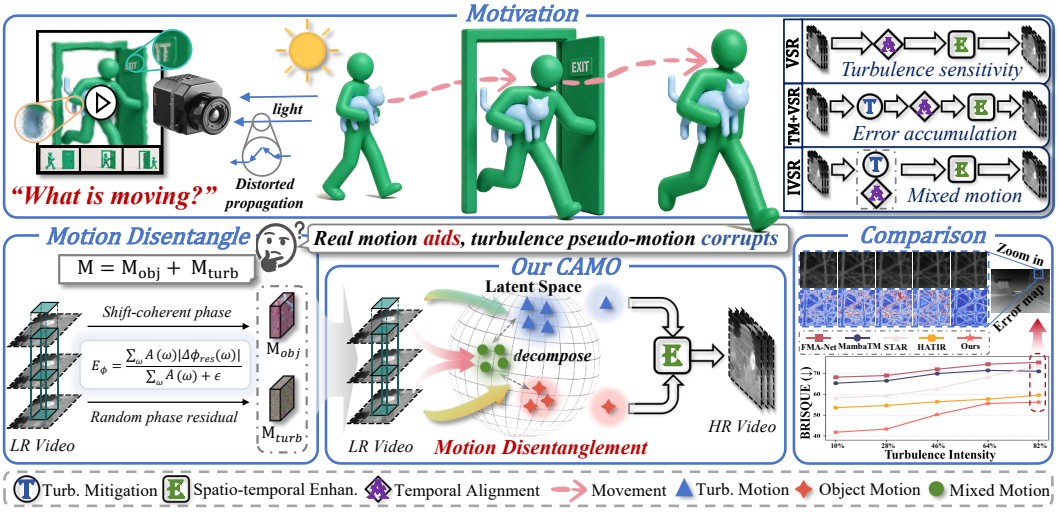

Figure 1: **Top:** Turbulence-induced pseudo-motion entangles with real object motion. Consequently, VSR is sensitive to turbulence, cascaded TM+VSR accumulates errors, and unified IVSR mistakes pseudo-motion for valid temporal correspondence. **Bottom:** Our proposed CAMO disentangles the two motions in latent space for robust spatio-temporal enhancement, delivering consistent gains across turbulence intensities.

ABSTRACT

Infrared turbulent video super-resolution is fundamentally challenging due to the entanglement of real object/camera motion and random turbulence-induced distortions. Existing turbulence mitigation and video super-resolution methods often conflate turbulence-induced pseudo-motion with real object/camera motion and model them as a unified temporal correspondence signal, leading to spurious alignments and unstable details. We propose **CAMO**, a **C**ausal-role-**A**ware **M**otion-turbulence disentanglement framework for **O**ne-step diffusion restoration, rooted in the physical principle that the real motion induces coherent phase transport in the infrared domain, while turbulence-induced refractive perturbations break such coherence. This coherence provides a reliability cue for factorizing mixed temporal dynamics into constructive object motion and destructive turbulence components. The turbulence component is used to suppress local non-rigid distortions, while the real motion component promotes temporal consistency and enhances motion-aware details during reconstruction. Furthermore, CAMO regulates the generative conditioning and velocity prediction based on motion reliability, enabling constructive motion cues to be selectively exploited while preventing the amplification of uncertain turbulence-induced pseudo-motion. This unified design enables end-to-end turbulence suppression, temporal information utilization, and high-resolution reconstruction. Extensive experiments under complex infrared turbulence demonstrate improved restoration fidelity, temporal stability, and robustness over existing methods. Anonymous code is available at CAMO.

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.

1 INTRODUCTION

Infrared video super-resolution (IVSR) is essential for long-range perception tasks under adverse environments, such as surveillance, autonomous driving, and target tracking Zou et al. (2026b;a); Liu et al. (2022). Unlike conventional video super-resolution (VSR), IVSR in open-path settings faces joint degradation from limited infrared resolution and weak thermal textures, together with turbulence-induced geometric distortions, spatially varying blur, and temporal intensity fluctuations Zhang et al. (2024b;a); Zou et al. (2026b). At the same time, real object or camera motion also produces frame-to-frame variations, which are indispensable for exploiting cross-frame complementary information. Consequently, as shown in Figure 1 (top), the temporal variations observed in infrared turbulent videos are entangled mixtures of real scene motion and turbulence-induced distortions. Although both factors may appear as similar displacements in the image space, they have opposite causal effects on high-resolution reconstruction: *real motion provides useful temporal cues, whereas turbulence-induced pseudo-motion corrupts correspondence and should be suppressed*.

To recover details under these compound degradations, a natural strategy is to exploit temporal information, as in conventional video restoration. Sliding-window, recurrent, and transformer-based VSR methods model temporal dependencies through motion compensation, feature propagation, deformable alignment, or attention Wang et al. (2019); Chan et al. (2021; 2022); Liang et al. (2022; 2024); Zhou et al. (2024b), while diffusion- and flow-based models incorporate temporal conditions, motion guidance, or aligned features Zhou et al. (2024a); Rota et al. (2024); Yang et al. (2024); Chen et al. (2026b); Wang et al. (2026); Xu et al. (2026). However, these approaches generally treat inter-frame variations as informative motion cues, an assumption that becomes unreliable when turbulence also produces apparent displacement.

To address this limitation, recent turbulence mitigation (TM) methods Saha et al. (2024); Zhang et al. (2024a;b); Jaiswal et al. (2023); Zhang et al. (2025) and turbulence-aware IVSR Zou et al. (2026b) incorporate physical priors, specialized alignment, or unified restoration. HATIR Zou et al. (2026b) unifies turbulence mitigation with super-resolution, yet still mistakes pseudo-motion for valid temporal correspondence via optical-flow guidance. Consequently, beneficial real motion remains entangled with detrimental turbulence-induced pseudo-motion, motivating an explicit separation of their opposing temporal roles.

While prior studies have significantly advanced turbulence mitigation and infrared video super-resolution, three fundamental challenges remain for turbulent IVSR. **(1) Insufficient coupling between turbulence suppression and super-resolution.** These two processes are inherently coupled. Suppressing local distortions may attenuate useful thermal structures, whereas residual turbulence can be amplified into false details during high-resolution reconstruction Zou et al. (2026b). Existing cascaded or weakly coupled frameworks (e.g., HATIR) still lack a fully end-to-end formulation that jointly optimizes temporal stabilization, turbulence suppression, and detail generation. **(2) Causal ambiguity of temporal variations.** Existing methods commonly rely on optical flow Yang et al. (2024); Xue et al. (2019); Chan et al. (2021), deformable alignment Tian et al. (2020); Wang et al. (2019); Liang et al. (2022), or holistic motion guidance Zou et al. (2026b); Zhuang et al. (2026); Xu et al. (2026), implicitly representing real motion and turbulence-induced pseudo-motion as a single mixed correspondence field. This overlooks their opposite reconstruction effects: real motion supplies complementary details, whereas pseudo-motion corrupts correspondence. **(3) Generative amplification of unreliable temporal cues.** Generative restoration models often condition reconstruction on estimated motion or aligned features without explicitly modeling their reliability. Under severe turbulence, unreliable temporal cues can bias the generative trajectory, causing ghosting, structural drift, and hallucinated details.

To this end, we formulate turbulent infrared video super-resolution as a causal-role-aware generative restoration problem and propose **CAMO**, an end-to-end motion-turbulence disentangled framework for turbulent IVSR. As shown in Figure 1 (bottom), rather than pursuing a more robust holistic correspondence field (e.g., optical flow or deformation fields), CAMO explicitly assigns different causal roles to different temporal factors. The real object/camera motion is treated as a constructive cue for cross-frame information aggregation, while turbulence-induced pseudo-motion is regarded as a destructive disturbance to be suppressed. Specifically, CAMO introduces *Latent Motion-Turbulence Disentanglement* (LMTD) to factorize mixed temporal dynamics into motion-related and turbulence-related latent components. To ground this factorization in infrared imaging

physics, CAMO further derives a *Thermo-Radiative Motion Coherence Prior* (TRMCP), based on the observation that real motion induces coherent phase transport in the infrared domain, whereas turbulence-induced refractive perturbations break such coherence Kuglin (1975); Reddy & Chatterji (1996); Chan (2022); Chimitt & Chan (2020). TRMCP provides a reliability cue for LMTD, guiding the turbulence component to suppress local non-rigid distortions and the motion component to enhance temporally consistent details. Finally, CAMO develops a *Motion Reliability-Guided Diffusion* (MRGD) to regulate generative conditioning and velocity prediction according to motion reliability, preventing uncertain pseudo-motion from being amplified during high-resolution generation. By unifying turbulence suppression, temporal information utilization, and high-resolution generation, CAMO provides a causal-role-aware and reliability-guided end-to-end solution for turbulent IVSR.

In summary, (1) we propose CAMO, a causal-role-aware one-step diffusion for turbulent IVSR that jointly models turbulence suppression, temporal information utilization, and high-resolution reconstruction in an end-to-end manner; (2) we introduce Latent Motion-Turbulence Disentanglement (LMTD), which factorizes mixed temporal dynamics into role-specific motion and turbulence components, with a Thermo-Radiative Motion Coherence Prior (TRMCP) providing physical reliability guidance for turbulent infrared videos; (3) we design Motion Reliability-Guided Diffusion (MRGD), which regulates generative conditioning and velocity prediction according to motion reliability, enabling constructive temporal cues to be selectively exploited while suppressing turbulence-contaminated pseudo-motion.

2 Related Work

2.1 Video Super-Resolution

Existing VSR methods mainly follow restoration-oriented and generative paradigms. Restoration-oriented methods aggregate multi-frame observations through deformable alignment Wang et al. (2019), recurrent propagation Chan et al. (2021; 2022), and parallel or recurrent Transformer modeling Liang et al. (2024; 2022). Although effective in reconstruction fidelity, these methods depend on reliable temporal correspondence and tend to smooth uncertain high-frequency details. Generative methods instead adapt image diffusion priors using temporal conditioning, motion guidance, or latent propagation Zhou et al. (2024a); Rota et al. (2024); Yang et al. (2024), while another paradigm exploits pretrained video diffusion models for joint spatio-temporal generation Xie et al. (2025); Chen et al. (2026a); Bai et al. (2025). One-step and distilled diffusion formulations further reduce sampling costs Chen et al. (2026b); Zhuang et al. (2026); Lv et al. (2026). Infrared-specific VSR additionally incorporates thermal dynamics and temperature-field consistency Zhou et al. (2026). Nevertheless, these paradigms largely treat inter-frame variations as reliable scene correspondence, an assumption violated when turbulence-induced pseudo-motion is entangled with real motion.

2.2 Atmospheric Turbulence Mitigation

Atmospheric turbulence mitigation (TM) addresses coupled degradations such as spatially varying blur, geometric distortion, local jitter, and temporal flickering. Classical methods reconstruct a latent image through registration, lucky-region selection, and multi-frame fusion Anantrasirichai et al. (2013), but often rely on accurate alignment and static scenes. Learning-based methods replace these hand-crafted pipelines with data-driven restoration. Early CNN- and GAN-based approaches learn distortion correction and deblurring from single images or short sequences Yasarla & Patel (2020); Jin et al. (2021), while physics-informed methods incorporate turbulence representations, physical degradation models, Transformer architectures, or simulation-guided refinement Mao et al. (2022); Zhang et al. (2024b); Jaiswal et al. (2023). For dynamic videos, recent methods integrate classical TM operations into learnable temporal aggregation, separate dynamic foregrounds from static backgrounds, regularize neural video representations, or employ state-space models for efficient long-range modeling Zhang et al. (2024a); Saha et al. (2024); Cai et al. (2023); Zhang et al. (2025). Beyond standalone TM, turbulence-aware IVSR jointly models turbulence suppression and structural detail recovery using heat-aware diffusion priors Zou et al. (2026b). Nevertheless, these methods mainly suppress or regularize turbulence-induced variations, without explicitly separating beneficial real motion from detrimental pseudo-motion in a unified IVSR framework.

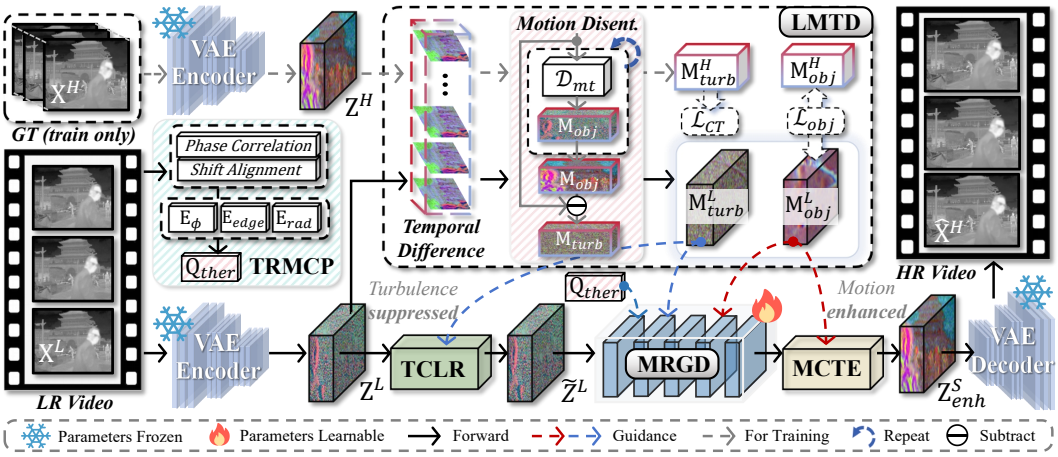

Figure 2: **Overall pipeline of CAMO.** A frozen VAE encodes the turbulent LR video into latent space, where LMTD disentangles temporal variations into object-motion and turbulence components under TRMCP guidance. The turbulence component rectifies the LR latent through TCLR, while the object-motion component guides MCTE for temporally consistent enhancement. MRGD further modulates the DiT features and velocity according to motion reliability. The GT branch is used only for training supervision.

3 METHOD

3.1 LATENT MOTION-TURBULENCE DISENTANGLEMENT

Temporal variations in degraded videos mainly come from real object/camera motion and turbulence-induced non-rigid deformation, which are often mixed during restoration. We introduce Latent Motion-Turbulence Disentanglement (LMTD) to separate latent temporal variations into object-motion and turbulence components, with the Thermo-Radiative Motion Coherence Prior (TRMCP) providing physical supervision for the disentanglement. The turbulence component is then used for latent rectification, while the object-motion component supports temporal enhancement.

As shown in Figure 2, given a clean video $\mathbf{X}^H$ and its turbulent low-quality counterpart $\mathbf{X}^L$, we encode them into the latent space using a pre-trained VAE encoder $\mathcal{E}(\cdot)$, obtaining $\mathbf{Z}^H = \mathcal{E}(\mathbf{X}^H)$ and $\mathbf{Z}^L = \mathcal{E}(\mathbf{X}^L)$, where $\mathbf{Z} \in \mathbb{R}^{B\times C\times F\times H_z\times W_z}$, with B, C, F, H_z, and W_z denoting the batch size, latent channels, number of frames, latent height, and latent width. For a latent sequence $\mathbf{Z}$, we compute the first-order temporal difference $\Delta\mathbf{Z}_t = \mathbf{Z}_{t+1} - \mathbf{Z}_t$ to capture inter-frame variations, and pad it along the temporal dimension to obtain $\mathbf{M} \in \mathbb{R}^{B\times C\times F\times H_z\times W_z}$. Since $\mathbf{M}$ contains both real motion and turbulence-induced variations, we decompose it into two components:

$$\mathbf{M} = \mathbf{M}_{\text{obj}} + \mathbf{M}_{\text{turb}}, \tag{1}$$

where $\mathbf{M}_{\text{obj}}$ represents object/camera motion and $\mathbf{M}_{\text{turb}}$ captures turbulence-induced perturbations. To disentangle $\mathbf{M}$, we design a lightweight motion–turbulence decoupler $\mathcal{D}_{\text{mt}}(\cdot)$. It first applies Haar decomposition along the temporal dimension to split $\mathbf{M}$ into a low-frequency component $\mathbf{M}_{\text{lp}}$, which captures smooth motion trends, and a high-frequency component $\mathbf{M}_{\text{hp}}$, which contains local temporal variations. A gating network then predicts an adaptive weight $\mathbf{G}$ to retain useful high-frequency information, and the object-motion component is estimated by

$$\mathbf{M}_{\text{base}} = \mathbf{M}_{\text{lp}} + \mathbf{G} \odot \mathbf{M}_{\text{hp}}, \quad \mathbf{M}_{\text{obj}} = \mathcal{F}_s(\mathbf{M}_{\text{base}}) + \mathbf{M}_{\text{base}}, \quad \mathbf{M}_{\text{turb}} = \mathbf{M} - \mathbf{M}_{\text{obj}}. \tag{2}$$

Here, $\mathcal{F}_s(\cdot)$ is a smooth motion estimator and $\odot$ denotes element-wise multiplication. We stack multiple decouplers progressively, where the estimated object-motion feature is further refined in the next stage and the separated turbulence residuals are accumulated. This progressive decomposition preserves stable motion while gradually removing unstable turbulence perturbations. As illustrated in Figure 3 (a), LMTD separates the mixed temporal variation into object-motion and turbulence components, with further visualization and analysis provided in Appendix A.1.

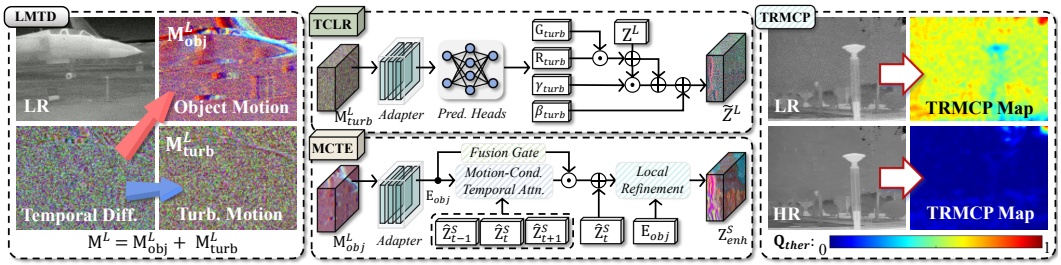

Figure 3: (a) Visualization of the mixed latent temporal-difference feature and its disentangled object-motion and turbulence components. (b) Architectures of the proposed Turbulence-Conditioned Latent Rectifier (TCLR) and Motion-Conditioned Temporal Enhancement (MCTE). (c) TRMCP maps computed from the LR input and HR reference, illustrating their responses to turbulence-induced perturbations, where warmer colors indicate stronger perturbations.

After disentanglement, the turbulence component is used to rectify the LR latent before DiT restoration. As shown in the upper part of Figure 3 (b), a lightweight adapter maps $\mathbf{M}^L_{\text{turb}}$ to the turbulence condition $\mathbf{E}_{\text{turb}} = \mathcal{A}_{\text{turb}}(\mathbf{M}^L_{\text{turb}})$, from which Turbulence-Conditioned Latent Rectifier (TCLR) predicts a correction residual $\mathbf{R}_{\text{turb}}$, a gating map $\mathbf{G}_{\text{turb}}$, and affine parameters $\boldsymbol{\gamma}_{\text{turb}}$ and $\boldsymbol{\beta}_{\text{turb}}$. The rectified latent is given by

$$\widetilde{\mathbf{Z}}^L = (1 + \boldsymbol{\gamma}_{\text{turb}}) \odot \left(\mathbf{Z}^L + \mathbf{G}_{\text{turb}} \odot \mathbf{R}_{\text{turb}}\right) + \boldsymbol{\beta}_{\text{turb}}. \tag{3}$$

This suppresses local turbulence perturbations and provides a more stable condition for the DiT, which predicts the velocity $\mathbf{v}_\theta = \Phi_{\text{DiT}}(\widetilde{\mathbf{Z}}^L, t)$. The predicted velocity is then applied to the rectified LR latent $\widetilde{\mathbf{Z}}^L$ to obtain a preliminary restored latent $\widehat{\mathbf{Z}}^S$, which is subsequently refined by Motion-Conditioned Temporal Enhancement (MCTE) to improve temporal consistency.

MCTE propagates reliable information across adjacent frames based on the disentangled object-motion component. As shown in the lower part of Figure 3(b), the object-motion feature is first mapped to a latent condition $\mathbf{E}_{\text{obj}} = \mathcal{A}_{\text{obj}}(\mathbf{M}^L_{\text{obj}})$, which modulates local temporal attention. For the current latent $\widehat{\mathbf{Z}}^S_t$, we use its adjacent frames $\widehat{\mathbf{Z}}^S_r$, $r \in \{t-1, t+1\}$, as temporal references. The query, key, and value features are computed as $\mathbf{Q} = W_q \widehat{\mathbf{Z}}^S_t$, $\mathbf{K}_r = W_k \widehat{\mathbf{Z}}^S_r$, and $\mathbf{V}_r = W_v \widehat{\mathbf{Z}}^S_r$. The motion condition $\mathbf{E}_{\text{obj}}$ predicts scale-shift parameters to modulate the normalized query and key features, yielding $\mathbf{Q}'$ and $\mathbf{K}'_r$. The temporal context is then obtained by scaled dot-product attention as $\mathbf{C}_r = \text{Softmax}(\mathbf{Q}'\mathbf{K}'^{\top}_r/\sqrt{d})\mathbf{V}_r$, where d denotes the feature dimension. The temporal contexts from adjacent frames are averaged and fused into the current SR latent through a motion-conditioned gate:

$$\mathbf{C} = \text{Avg}_{r\in\{t-1,t+1\}}(\mathbf{C}_r), \quad \mathbf{Z}^S_{\text{prop}} = \widehat{\mathbf{Z}}^S_t + \sigma(\mathcal{F}_p(\mathbf{E}_{\text{obj}})) \odot \mathbf{C}, \quad \mathbf{Z}^S_{\text{enh}} = \mathcal{F}_{\text{ref}}\left(\mathbf{Z}^S_{\text{prop}}, \mathbf{E}_{\text{obj}}\right). \tag{4}$$

Here, $\sigma(\mathcal{F}_p(\mathbf{E}_{\text{obj}}))$ controls temporal fusion, and $\mathcal{F}_{\text{ref}}(\cdot)$ performs lightweight local refinement.

Overall, LMTD separates mixed temporal variations into object-motion and turbulence components, which are then used at different restoration stages. TCLR suppresses turbulence-induced perturbations before DiT restoration, while MCTE exploits object-motion cues to enhance temporal consistency and local details.

3.2 Motion Reliability-Guided Diffusion

Disentangled motion features may still be unreliable under strong turbulence. We therefore use TRMCP to estimate the local turbulence strength $\mathbf{Q}_{\text{ther}}$ and combine it with the motion uncertainty $\mathbf{U}_{\text{mot}}$ derived from the disentangled features to obtain a spatio-temporal reliability map $\boldsymbol{\rho}$. The reliability map modulates DiT features and scales the predicted velocity field, reducing the influence of unreliable motion cues during restoration.

As illustrated in Figure 3 (c), given adjacent LR frames $\mathbf{X}^L_f$ and $\mathbf{X}^L_{f+1}$, TRMCP first estimates the local translation $(\Delta y, \Delta x)$ by phase correlation and compensates for this motion. It then measures

the residual perturbations from phase, radiative, and edge coherence, denoted as E_ϕ, E_{rad}, and E_{edge}. These terms are fused into a patch-level perturbation score:

$$Q_{\text{patch}} = \frac{w_\phi E_\phi + w_{\text{rad}} E_{\text{rad}} + w_{\text{edge}} E_{\text{edge}}}{w_\phi + w_{\text{rad}} + w_{\text{edge}} + \epsilon}, \tag{5}$$

where w_ϕ, w_{rad}, and w_{edge} control the contributions of the three residuals, and ϵ is a small constant for numerical stability. Further details of TRMCP are provided in Appendix A.2, with additional turbulence-strength analysis and TRMCP visualizations presented in Appendix A.3 and A.4.

Besides the image-domain prior $\mathbf{Q}_{\text{ther}}$, we estimate motion uncertainty from the disentangled latent features. We measure the second-order temporal variation of object motion as $\mathbf{R}_m = \text{Mean}_C \left|\Delta_f^2 \mathbf{M}_{\text{obj}}^L\right|$, where $\Delta_f^2 \mathbf{M}_{\text{obj}}^L = \mathbf{M}_{\text{obj},f+1}^L - 2\mathbf{M}_{\text{obj},f}^L + \mathbf{M}_{\text{obj},f-1}^L$, and compute the feature conflict as $\mathbf{R}_{\text{int}} = 1 - \cos(\mathbf{M}_{\text{obj}}^L, \mathbf{M}_{\text{turb}}^L)$. These two cues are concatenated and fed into a lightweight 3D convolutional network to predict the motion uncertainty, $\mathbf{U}_{\text{mot}} = \sigma(\mathcal{H}_u([\mathbf{R}_m, \mathbf{R}_{\text{int}}]))$, where $\mathcal{H}_u(\cdot)$ denotes the uncertainty prediction network, $[\cdot]$ denotes channel-wise concatenation, and $\sigma(\cdot)$ is the Sigmoid function.

Finally, we combine the TRMCP perturbation prior and motion uncertainty as $\mathbf{U} = (1-\lambda_s)\mathbf{U}_{\text{mot}} + \lambda_s \mathbf{Q}_{\text{ther}}$, where λ_s balances their contributions. The motion reliability map is then defined as $\boldsymbol{\rho} = 1 - \mathbf{U}$. A larger $\boldsymbol{\rho}$ indicates more reliable motion, while a smaller value corresponds to stronger turbulence or greater motion–turbulence ambiguity.

In the DiT, the reliability map $\boldsymbol{\rho}$ is mapped to the token grid and flattened into a token-level representation $\boldsymbol{\rho}_{\text{tok}}$. We then introduce an AdaLN-style modulation after the normalization layer of each Transformer block:

$$\mathbf{h}' = \mathbf{h} \odot (1 + \boldsymbol{\rho}_{\text{tok}} \odot \Delta\mathbf{s}) + \boldsymbol{\rho}_{\text{tok}} \odot \Delta\mathbf{b}, \tag{6}$$

where $\mathbf{h}$ denotes the DiT token features, and $\Delta\mathbf{s}$ and $\Delta\mathbf{b}$ are scale-shift parameters predicted from $\boldsymbol{\rho}_{\text{tok}}$ by a lightweight MLP. By weighting the modulation with motion reliability, reliable regions receive stronger feature updates, while the influence of low-reliability, turbulence-contaminated regions is adaptively reduced. This helps prevent unreliable temporal cues from dominating feature propagation in the DiT. In addition, we scale the DiT velocity field with the reliability map as $\widetilde{\mathbf{v}}_\theta = \boldsymbol{\rho} \odot \mathbf{v}_\theta$, and reconstruct the latent by

$$\widehat{\mathbf{Z}}^S = \sqrt{\bar{\alpha}_t}\, \widetilde{\mathbf{Z}}^L - \sqrt{1-\bar{\alpha}_t}\, \widetilde{\mathbf{v}}_\theta, \tag{7}$$

where $\widetilde{\mathbf{Z}}^L$ is the turbulence-rectified LR latent, $\bar{\alpha}_t$ denotes the cumulative noise coefficient at timestep t, and $\widetilde{\mathbf{v}}_\theta$ is the reliability-scaled velocity prediction. This reduces unreliable updates in low-reliability regions and suppresses turbulence-induced temporal errors.

Overall, TRMCP and latent motion uncertainty jointly estimate the spatio-temporal reliability map $\boldsymbol{\rho}$. By modulating DiT features and velocity updates with $\boldsymbol{\rho}$, the model preserves reliable motion cues while suppressing turbulence-contaminated temporal information.

3.3 Loss Function

During training, the final enhanced latent $\mathbf{Z}_{\text{enh}}^S$ is supervised by the high-quality latent $\mathbf{Z}^H$. To reduce the memory cost of pixel-space supervision for long high-resolution videos, we impose the reconstruction constraint directly in the VAE latent space: $\mathcal{L}_{\text{rec}} = \|\mathbf{Z}_{\text{enh}}^S - \mathbf{Z}^H\|_2^2$. To regularize motion–turbulence disentanglement, we apply the shared decoupler $\mathcal{D}_{\text{mt}}(\cdot)$ to the high- and low-quality motion representations $\mathbf{M}^H$ and $\mathbf{M}^L$:

$$(\mathbf{M}_{\text{obj}}^H, \mathbf{M}_{\text{turb}}^H) = \mathcal{D}_{\text{mt}}(\mathbf{M}^H), \quad (\mathbf{M}_{\text{obj}}^L, \mathbf{M}_{\text{turb}}^L) = \mathcal{D}_{\text{mt}}(\mathbf{M}^L). \tag{8}$$

Since $\mathbf{X}^H$ and $\mathbf{X}^L$ share the same scene motion, we enforce object-motion consistency with $\mathcal{L}_{\text{obj}} = \|\mathbf{M}_{\text{obj}}^H - \mathbf{M}_{\text{obj}}^L\|_1$. Moreover, as the high-quality video is nearly free of turbulence, its turbulence component should be minimized, leading to $\mathcal{L}_{\text{CT}} = \|\mathbf{M}_{\text{turb}}^H\|_1$. To align the disentangled turbulence with the TRMCP prior, we compute the turbulence ratio $\mathbf{R}_{\text{turb}}^L$ from the relative magnitudes of $\mathbf{M}_{\text{turb}}^L$ and $\mathbf{M}_{\text{obj}}^L$. We then define $\mathcal{L}_{\text{ratio}} = \|\mathbf{R}_{\text{turb}}^L - \mathbf{Q}_{\text{ther}}\|_1$ to align the latent turbulence distribution with the image-domain perturbation prior. The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{motion}} \left(\mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{CT}} + \mathcal{L}_{\text{ratio}}\right), \tag{9}$$

where λ_{motion} balances the latent reconstruction loss and the motion–turbulence disentanglement regularization.

Table 1: Quantitative comparison on FLIR-IVSR and $\mathrm{M}^3\mathrm{FD}$. The best is in **bold**, while the second is underlined.

Methods			PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	BRISQUE↓	DOVER↑	FasterVQA↑
FLIR-IVSR										
Turb-Seg-Res	CVPR'24	TM + VSR	32.8455	0.9208	0.2066	0.1654	29.3413	68.3326	0.1182	0.2445
DATUM	CVPR'24	TM + VSR	35.4914	0.9361	0.1974	0.1499	27.6554	64.8126	0.0672	0.1620
MambaTM	CVPR'25	TM + VSR	35.9627	0.9314	0.1840	0.1423	28.5586	63.7148	0.0843	0.1786
FMA-Net	CVPR'24	VSR	35.5266	0.9369	0.1669	0.1369	29.9899	65.1855	0.0955	0.2695
STAR	ICCV'25	VSR	32.6164	0.9172	0.2094	0.1408	29.1100	56.6060	0.0743	0.1838
DLoRAL	NeurIPS'25	VSR	28.0540	0.8123	0.3830	0.2414	35.6405	48.9566	0.0875	0.2040
DGAF-VSR	CVPR'26	VSR	31.5517	0.8812	0.2506	0.1788	30.2780	61.9984	0.0644	0.2994
HATIR	AAAI'26	IVSR	33.6613	0.9120	0.1431	0.1394	27.8147	47.6045	0.0704	0.2143
Ours	-	IVSR	**36.7342**	**0.9477**	**0.1160**	**0.1304**	**35.7926**	**38.0304**	**0.1521**	**0.3692**
$\mathrm{M}^3\mathrm{FD}$										
Turb-Seg-Res	CVPR'24	TM + VSR	31.9689	0.8906	0.3438	0.2781	25.2415	75.0399	0.0985	0.1718
DATUM	CVPR'24	TM + VSR	34.2716	0.9179	0.2562	0.1889	30.1538	67.6024	0.0669	0.1651
MambaTM	CVPR'25	TM + VSR	34.3441	0.9150	0.2364	**0.1749**	31.4487	66.4243	0.0832	0.1933
FMA-Net	CVPR'24	VSR	33.1064	0.9051	0.2722	0.2182	27.8835	68.9096	0.0562	0.2002
STAR	ICCV'25	VSR	31.4200	0.8956	0.2668	0.1785	29.0865	59.2008	0.0622	0.1659
DLoRAL	NeurIPS'25	VSR	28.4490	0.8091	0.3632	0.2513	**32.9205**	**41.0341**	0.0624	0.2260
DGAF-VSR	CVPR'26	VSR	30.2435	0.8596	0.3308	0.2454	27.0057	66.6288	0.0448	0.1961
HATIR	AAAI'26	IVSR	32.3733	0.8940	0.2621	0.1953	23.7566	54.5624	0.0418	0.1466
Ours	-	IVSR	**34.3555**	**0.9182**	**0.2344**	0.2059	28.9877	43.2735	**0.1098**	**0.2453**

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Implementation Details. CAMO is built upon CogVideoX1.5 Yang et al. (2025) and trained on four RTX PRO 6000 GPUs with a total batch size of 8. We use 25-frame clips at 320×640 resolution and randomly drop up to 5 frames for augmentation. The model is trained for 10,000 iterations using AdamW with a learning rate of 2×10^{-5}. The loss weight λ_{motion} is set to 0.5.

Datasets and Metrics. Following the experimental protocol of HATIR Zou et al. (2026b), we conduct experiments on FLIR-IVSR Zou et al. (2026b) and $\mathrm{M}^3\mathrm{FD}$ Liu et al. (2022). For FLIR-IVSR, the first 135 videos are used for testing and the remaining ones for training. For $\mathrm{M}^3\mathrm{FD}$, we randomly select 100 images, repeat each for 30 frames, and apply turbulence distortion, downsampling, and noise. We evaluate PSNR, SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), DISTS Ding et al. (2020), MUSIQ Ke et al. (2021), BRISQUE Mittal et al. (2012), DOVER Wu et al. (2023), and FasterVQA Wu et al. (2022).

Comparative Methods. We compare with general VSR methods FMA-Net Youk et al. (2024), STAR Xie et al. (2025), DLoRAL Sun et al. (2025), and DGAF-VSR Xu et al. (2026), the IVSR method HATIR Zou et al. (2026b), and turbulence mitigation methods Turb-Seg-Res Saha et al. (2024), DATUM Zhang et al. (2024a), and MambaTM Zhang et al. (2025). The latter three are cascaded with IA-RT Xu et al. (2024) for SR. All methods are trained on the same datasets.

4.2 COMPARISON WITH EXISTING METHODS

Quantitative Comparisons. As shown in Table 1, CAMO achieves the best results across all eight metrics on FLIR-IVSR, demonstrating strong reconstruction fidelity, perceptual quality, and temporal consistency. On $\mathrm{M}^3\mathrm{FD}$, it also achieves the best PSNR, SSIM, LPIPS, DOVER, and FasterVQA, while remaining competitive on the other metrics. Although DLoRAL obtains better MUSIQ and BRISQUE scores on $\mathrm{M}^3\mathrm{FD}$, its substantially lower PSNR and SSIM indicate degraded structural fidelity, consistent with the visual distortions observed in qualitative results. Overall, CAMO provides a favorable balance between faithful reconstruction and perceptual quality under complex degradations. As shown in Figure 1 (bottom-right), CAMO maintains consistently strong performance across different turbulence intensities. Failure cases are analyzed in Appendix A.8.

Qualitative Comparisons. As shown in Figure 4, CAMO restores sharper, more faithful structural details across various scenes. Turb-Seg-Res, MambaTM, and FMA-Net tend to over-smooth fine textures, while DLoRAL and HATIR often introduce unrealistic high-frequency artifacts. In con-

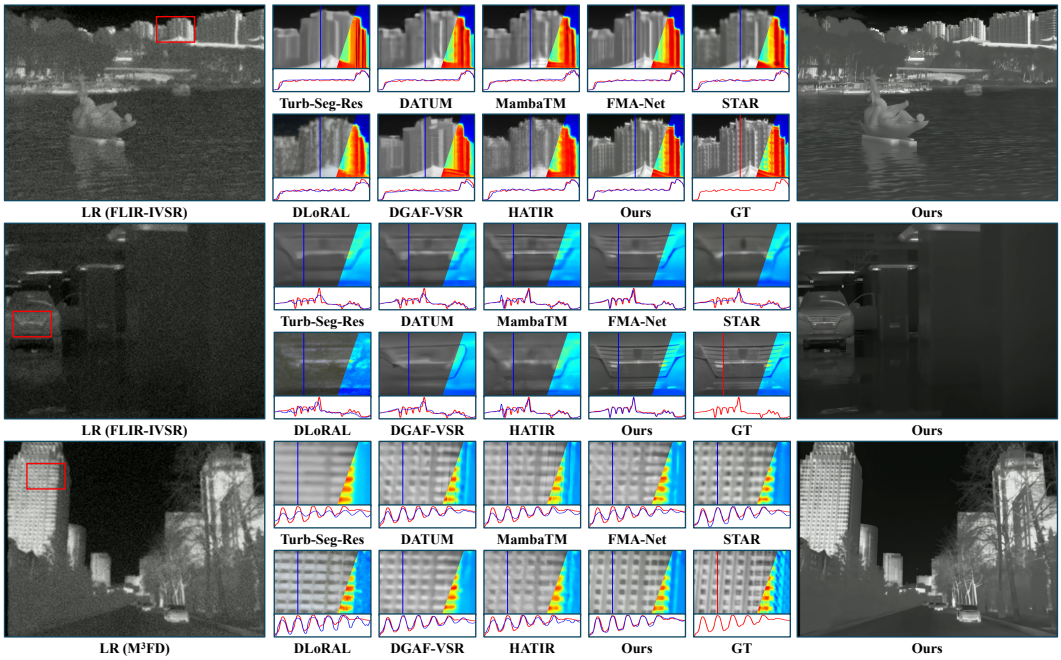

Figure 4: Qualitative comparison with SOTA methods on FLIR-IVSR and M^{3}FD datasets. The graph illustrates grayscale fluctuations along the blue-marked sampling line, and red-marked sampling line denotes the HR.

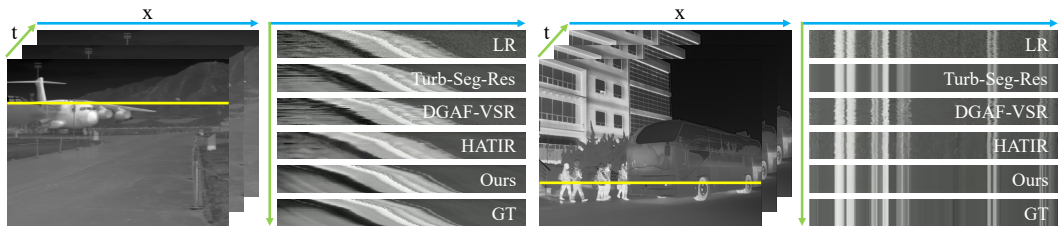

Figure 5: Temporal consistency comparison with SOTA methods on FLIR-IVSR and M^{3}FD datasets. The temporal profile is obtained by stacking the yellow line across frames.

trast, CAMO better preserves local structures and produces intensity profiles closer to the ground truth. Appendix A.5 provides more qualitative comparisons.

Figure 5 further demonstrates improved temporal consistency. Competing methods exhibit noticeable structural drift and discontinuous trajectories in the spatio-temporal slices, whereas CAMO produces smoother and continuous temporal patterns that better match the ground truth. This indicates more stable reconstruction with reduced flickering and turbulence-induced temporal artifacts. Appendix A.6 provides more temporal consistency comparisons.

Efficiency Analysis. As shown in Figure 6, CAMO achieves the highest FasterVQA score while running at 2.32 FPS, outperforming all diffusion-based competitors in inference efficiency. This favorable trade-off mainly benefits from the temporal compression strategy and one-step restoration formulation, which avoids costly iterative sampling. Compared with the one-step DLoRAL, CAMO is nearly $1.9\times$ faster while delivering substantially higher video quality. Detailed comparisons of sampling steps, model complexity, and runtime are provided in Appendix A.7.

4.3 Ablation Study

Effect of the Proposed Components. As shown in Figure 7, removing any proposed component degrades the overall performance, while the complete model achieves the best results across all

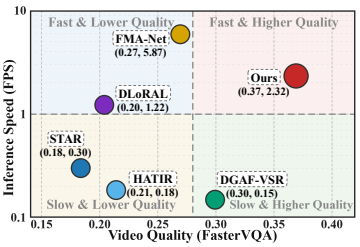

Figure 6: Efficiency comparison in terms of perceptual quality and runtime.

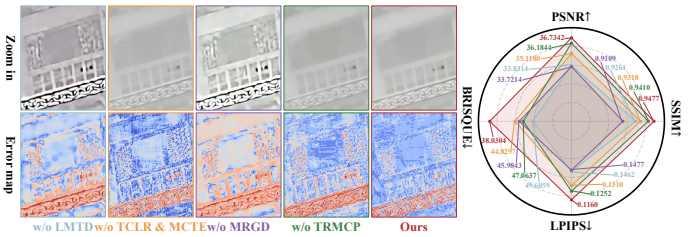

Figure 7: Visual (left) and quantitative (right) comparison of different ablation variants, including zoomed-in regions and corresponding error maps.

metrics. In particular, removing LMTD notably affects both reconstruction fidelity and perceptual quality, confirming the importance of disentangling real motion from turbulence-induced variations. TCLR/MCTE, MRGD, and TRMCP provide complementary improvements by respectively enhancing motion-aware restoration, regulating unreliable temporal cues, and providing turbulence-aware reliability guidance. The visual results are consistent with the quantitative findings. The ablated variants exhibit blurred or distorted local structures and larger reconstruction errors, whereas the full model preserves finer structural details with fewer high-error regions.

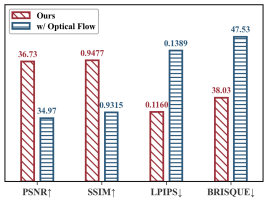

Figure 8: LMTD versus optical-flow replacement.

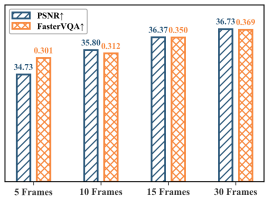

Figure 9: Performance versus input length.

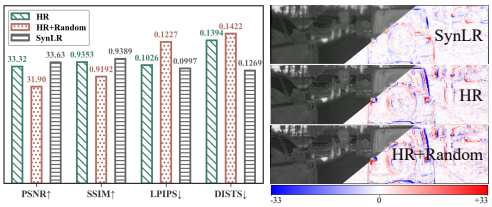

Figure 10: Quantitative comparison and error maps against **LR inputs**.

Comparison with Optical Flow. As shown in Figure 8, LMTD outperforms its optical-flow-based replacement (using SpyNet Ranjan & Black (2017)) across all metrics. This demonstrates that conventional optical flow is insufficient to characterize turbulence-corrupted temporal variations, whereas explicitly disentangling real motion from turbulence provides more reliable temporal cues.

Temporal Length Analysis. As shown in Figure 9, increasing the input length improves video quality with diminishing gains beyond 15 frames. This indicates that CAMO remains effective with short sequences and that its gains are not primarily attributable to longer temporal context.

Turbulence Disentanglement Validation. To validate whether LMTD captures turbulence-specific perturbations, we inject the disentangled turbulence features into clean HR videos to synthesize SynLR, and compare it with HR and HR degraded by random blur and distortion. As shown in Figure 10, SynLR achieves the best PSNR, SSIM, LPIPS, and DISTS against the real LR observations. The error maps also show that SynLR better reproduces real turbulence variations, while random perturbations lead to mismatched structures. Additional qualitative comparisons and error-map analysis are provided in Appendix A.9.

5 Conclusion

We presented CAMO, a causal-role-aware framework for turbulent infrared video super-resolution. CAMO explicitly disentangles constructive real motion from destructive turbulence-induced pseudo-motion. Guided by the thermo-radiative motion coherence prior, the disentangled components are used to suppress turbulence perturbations and enhance temporally consistent details, while reliability-guided diffusion modulation prevents uncertain temporal cues from being amplified during generation. Extensive experiments demonstrate strong reconstruction fidelity and robustness under varying turbulence conditions.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, figure refinement, and feedback on the presentation of the methodology and experiments. All AI-assisted content was independently reviewed, verified, and revised by the authors. The authors take full responsibility for the final content of this paper.

ETHICS STATEMENT

Our experiments use the existing publicly available FLIR-IVSR and M^3FD datasets and involve no new collection of human-subject data. In the infrared imagery used in this work, facial features and vehicle license plates are not identifiable, thereby limiting potential privacy risks. We use these datasets solely for academic research and follow their respective terms of use.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the model architecture, training objectives, implementation details, dataset protocols, evaluation settings, and additional experimental results needed to reproduce our study. We also provide an anonymous project page with qualitative comparisons and release the training and inference code, configurations, and pretrained models to facilitate independent verification.

REFERENCES

Nantheera Anantrasirichai, Alin Achim, Nick G Kingsbury, and David R Bull. Atmospheric turbulence mitigation using complex wavelet-based fusion. *IEEE Transactions on Image Processing*, 22(6):2398–2408, 2013.

Haoran Bai, Xiaoxu Chen, Canqian Yang, Zongyao He, Sibin Deng, and Ying Chen. Vivid-vr: Distilling concepts from text-to-video diffusion transformer for photorealistic video restoration. *arXiv preprint arXiv:2508.14483*, 2025.

Haoming Cai, Jingxi Chen, Brandon Y Feng, Weiyun Jiang, Mingyang Xie, Kevin Zhang, Ashok Veeraraghavan, and Christopher Metzler. Convrt: Consistent video restoration through turbulence with test-time optimization of neural video representations. *arXiv preprint arXiv:2312.04679*, 2023.

Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4947–4956, 2021.

Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5972–5981, 2022.

Stanley H Chan. Tilt-then-blur or blur-then-tilt? clarifying the atmospheric turbulence model. *IEEE Signal Processing Letters*, 29:1833–1837, 2022.

Junyang Chen, Jiangxin Dong, Long Sun, Yixin Yang, and Jinshan Pan. Stcdit: Spatio-temporally consistent diffusion transformer for high-quality video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38281–38290, 2026a.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. Dove: Efficient one-step diffusion model for real-world video super-resolution. *Advances in Neural Information Processing Systems*, 38:85218–85237, 2026b.

Nicholas Chimitt and Stanley H Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated zernike coefficients. *Optical Engineering*, 59(8):083101–083101, 2020.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

Ajay Jaiswal, Xingguang Zhang, Stanley H Chan, and Zhangyang Wang. Physics-driven turbulence image restoration with stochastic refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12170–12181, 2023.

Darui Jin, Ying Chen, Yi Lu, Junzhang Chen, Peng Wang, Zichao Liu, Sheng Guo, and Xiangzhi Bai. Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning. *Nature Machine Intelligence*, 3(10):876–884, 2021.

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.

Charles D Kuglin. The phase correlation image alignment method. In *IEEE Int. Conf. on Cybernetics and Society, 1975*, pp. 163–165, 1975.

Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022.

Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5811, 2022.

Zhengyao Lv, Menghan Xia, Xintao Wang, and Kwan-Yee K Wong. Duo-vsr: Dual-stream distillation for one-step video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16333–16344, 2026.

Zhiyuan Mao, Ajay Jaiswal, Zhangyang Wang, and Stanley H Chan. Single frame atmospheric turbulence mitigation: A benchmark study and a new physics-inspired transformer model. In *European conference on computer vision*, pp. 430–446. Springer, 2022.

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

Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2720–2729. IEEE, 2017.

B Srinivasa Reddy and Biswanath N Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE transactions on image processing*, 5(8):1266–1271, 1996.

Claudio Rota, Marco Buzzelli, and Joost van de Weijer. Enhancing perceptual quality in video super-resolution through temporally-consistent detail synthesis using diffusion models. In *European Conference on Computer Vision*, pp. 36–53. Springer, 2024.

Ripon Kumar Saha, Dehao Qin, Nianyi Li, Jinwei Ye, and Suren Jayasuriya. Turb-seg-res: A segment-then-restore pipeline for dynamic videos with atmospheric turbulence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25286–25296, 2024.

Yujing Sun, Lingchen Sun, Shuaizheng Liu, Rongyuan Wu, Zhengqiang Zhang, and Lei Zhang. One-step diffusion for detail-rich and temporally consistent video super-resolution. *Advances in Neural Information Processing Systems*, 38:172821–172841, 2025.

Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3360–3369, 2020.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, et al. Seedvr2: One-step video restoration via diffusion adversarial post-training. In *International Conference on Learning Representations*, volume 2026, pp. 40987–41009, 2026.

Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

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

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pp. 538–554. Springer, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20144–20154, 2023.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17108–17118, 2025.

Jingyi Xu, Meisong Zheng, Ying Chen, Minglang Qiao, Xin Deng, and Mai Xu. Rethinking diffusion model-based video super-resolution: Leveraging dense guidance from aligned features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38248–38257, 2026.

Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2546–2555, 2024.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European conference on computer vision*, pp. 224–242. Springer, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Rajeev Yasarla and Vishal M Patel. Learning to restore a single face image degraded by atmospheric turbulence using cnns. *arXiv preprint arXiv:2007.08404*, 2020.

Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fma-net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 44–55, 2024.

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

Xingguang Zhang, Nicholas Chimitt, Yiheng Chi, Zhiyuan Mao, and Stanley H Chan. Spatio-temporal turbulence mitigation: A translational perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2899, 2024a.

Xingguang Zhang, Zhiyuan Mao, Nicholas Chimitt, and Stanley H Chan. Imaging through the atmosphere using turbulence mitigation transformer. *IEEE Transactions on Computational Imaging*, 10:115–128, 2024b.

Xingguang Zhang, Nicholas Chimitt, Xijun Wang, Yu Yuan, and Stanley H Chan. Learning phase distortion with selective state space models for video turbulence mitigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2127–2138, 2025.

Mingxuan Zhou, Shuang Li, Yutang Zhang, Jing Geng, Yirui Shen, Jingxuan Kang, Fuzhen Zhuang, and Shuigen Wang. Thermal diffusion matters: Infrared spatial-temporal video super-resolution through heat conduction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2146–2155, 2026.

Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2535–2545, 2024a.

Xingyu Zhou, Leheng Zhang, Xiaorui Zhao, Keze Wang, Leida Li, and Shuhang Gu. Video super-resolution transformer with masked inter&intra-frame attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25399–25408, 2024b.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43482–43493, 2026.

Yang Zou, Jun Ma, Zhidong Jiao, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Toward real-world infrared image super-resolution: A unified autoregressive framework and benchmark dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16365–16375, 2026a.

Yang Zou, Xingyue Zhu, Kaiqi Han, Jun Ma, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Hatir: Heat-aware diffusion for turbulent infrared video super-resolution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):14095–14103, 2026b. doi: 10.1609/aaai.v40i16.38421. URL https://ojs.aaai.org/index.php/AAAI/article/view/38421.

A APPENDIX

Summary. We provide additional analyses and visualizations of CAMO, including motion–turbulence disentanglement (Sec. A.1), the detailed formulation of TRMCP (Sec. A.2), analysis of its response to turbulence strength (Sec. A.3), additional TRMCP maps (Sec. A.4), qualitative comparisons (Sec. A.5), temporal consistency comparisons (Sec. A.6), efficiency analysis (Sec. A.7), extreme-turbulence failure cases (Sec. A.8), and further validation of turbulence-specific feature disentanglement (Sec. A.9).

A.1 VISUALIZATION OF MOTION-TURBULENCE DISENTANGLEMENT

Figure 11 visualizes the latent temporal difference and the motion components disentangled by LMTD across different infrared video samples. The original temporal-difference features contain mixed responses from object motion and turbulence-induced variations, resulting in spatially entangled and irregular patterns. After disentanglement, the object-motion component mainly concentrates on coherent structures and moving boundaries, such as poles, vegetation, and object contours. In contrast, the turbulence component exhibits more scattered and spatially irregular responses without clear object structures. Similar patterns are observed across different scenes, showing that LMTD can separate structured motion cues from unstable turbulence-induced perturbations. This qualitative result supports the use of the two components for different restoration roles, with object motion providing temporal information and turbulence features indicating regions that require distortion suppression.

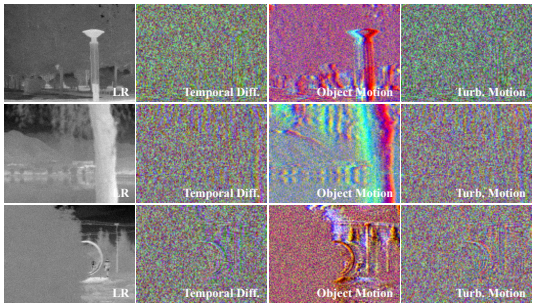

Figure 11: Visualization of motion disentanglement on different infrared video samples.

A.2 DETAILED FORMULATION OF TRMCP

Given a low-quality infrared video $\mathbf{X}^L$, we first convert each frame into a normalized pseudo-radiance representation. For a grayscale image $\mathbf{I}$, the pseudo-radiance response is defined as $\mathbf{\Theta} = \log(\mathbf{I}^\gamma + \epsilon)$, followed by frame-wise normalization to $[0, 1]$. This reduces the influence of different intensity ranges on coherence estimation.

For two adjacent frames $\mathbf{\Theta}_f$ and $\mathbf{\Theta}_{f+1}$, we extract overlapping local patches and apply a Hann window before Fourier transformation. Let $\mathcal{F}_f$ and $\mathcal{F}_{f+1}$ denote the spectra of two corresponding patches. We estimate their local translation $(\Delta y, \Delta x)$ using phase correlation with the normalized cross-power spectrum

$$\mathbf{C} = \frac{\mathcal{F}_{f+1}\mathcal{F}_f^*}{|\mathcal{F}_{f+1}\mathcal{F}_f^*| + \epsilon}. \tag{10}$$

The translation is obtained from the peak of the inverse Fourier response and is used to compensate for the dominant rigid motion between the two patches.

After motion compensation, TRMCP measures the remaining perturbations from phase, radiative, and edge coherence. According to the Fourier shift theorem, the phase change caused by translation $(\Delta y, \Delta x)$ is $2\pi(f_y \Delta y + f_x \Delta x)$. After removing this translation-induced phase change from the observed phase difference, we obtain the residual phase $\Delta\phi_{\text{res}}$. The phase residual is then computed as

$$E_\phi = \frac{\sum_{\boldsymbol{\omega}} A(\boldsymbol{\omega}) |\Delta\phi_{\text{res}}(\boldsymbol{\omega})|}{\sum_{\boldsymbol{\omega}} A(\boldsymbol{\omega}) + \epsilon}, \tag{11}$$

where $A(\boldsymbol{\omega}) = \sqrt{|\mathcal{F}_f(\boldsymbol{\omega})||\mathcal{F}_{f+1}(\boldsymbol{\omega})| + \epsilon}$ weights reliable frequency components. A larger E_ϕ indicates stronger phase variation that cannot be explained by coherent translation.

We further align the second patch according to $(\Delta y, \Delta x)$ and denote it as $\widetilde{\mathbf{P}}_{f+1}$. The radiative residual is defined as $E_{\text{rad}} = \text{Mean}(|\mathbf{P}_f - \widetilde{\mathbf{P}}_{f+1}|)$, which measures the remaining local intensity

variation after motion compensation. We also compute Sobel gradient magnitudes $\mathbf{G}_f$ and $\widetilde{\mathbf{G}}_{f+1}$ and define the edge residual as $E_{\text{edge}} = \text{Mean}(|\mathbf{G}_f - \widetilde{\mathbf{G}}_{f+1}|)$ to measure local structural inconsistency.

After normalization, the three residuals are fused into a patch-level perturbation score:

$$Q_{\text{patch}} = \frac{w_\phi E_\phi + w_{\text{rad}} E_{\text{rad}} + w_{\text{edge}} E_{\text{edge}}}{w_\phi + w_{\text{rad}} + w_{\text{edge}} + \epsilon}, \tag{12}$$

where w_ϕ, w_{rad}, and w_{edge} control the contributions of the three coherence terms. The patch scores are mapped back to the image space and averaged in overlapping regions to obtain a perturbation map for each adjacent frame pair. Finally, these pair-wise maps are aggregated along the temporal dimension to obtain the frame-wise TRMCP perturbation map $\mathbf{Q}_{\text{ther}} \in \mathbb{R}^{B\times T\times 1\times H\times W}$. For intermediate frames, the responses from the previous and next frame pairs are averaged, while the first and last frames use their single adjacent pair. A larger $\mathbf{Q}_{\text{ther}}$ indicates stronger local thermo-radiative perturbation.

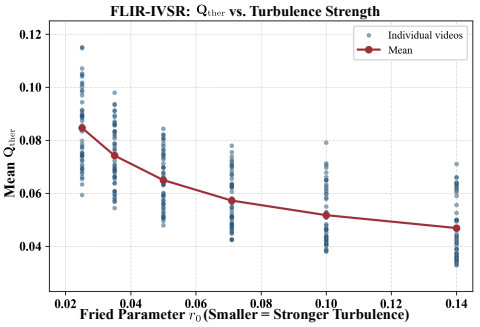

Figure 12: Response of $\mathbf{Q}_{\text{ther}}$ to different turbulence strengths on FLIR-IVSR.

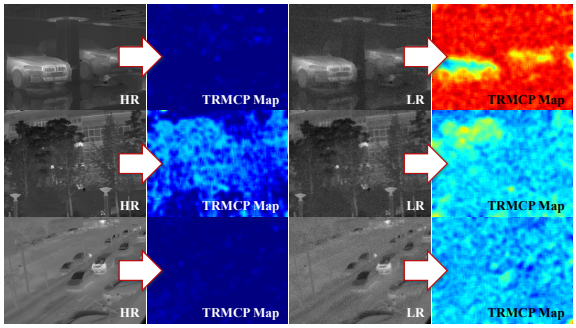

Figure 13: Visualization of TRMCP maps on different infrared video samples.

A.3 Analysis of TRMCP Response to Turbulence Strength

As shown in Figure 12, the blue points represent the $\mathbf{Q}_{\text{ther}}$ values of individual video samples at each turbulence level, while the red points denote their mean values. As the Fried parameter r_0 decreases, indicating stronger turbulence, the mean $\mathbf{Q}_{\text{ther}}$ increases consistently. Although individual videos show some variation at the same turbulence level, their overall responses shift toward larger $\mathbf{Q}_{\text{ther}}$ under stronger turbulence. This trend shows that TRMCP can effectively reflect changes in turbulence strength across different video contents, supporting $\mathbf{Q}_{\text{ther}}$ as a physical prior for motion reliability estimation.

A.4 Additional TRMCP Maps

As shown in Figure 13, the TRMCP maps exhibit clear differences between the HR and turbulence-degraded LR videos across different scenes. The HR samples generally produce weak responses dominated by low-value regions, whereas the LR samples show stronger and more spatially extensive responses. The spatial distributions also vary with scene content, indicating that TRMCP captures local coherence disruptions rather than simply responding to image intensity or texture. Although weak responses may remain in some HR regions due to scene motion or local radiative variations, the overall contrast between HR and LR samples demonstrates the sensitivity of TRMCP to turbulence-induced perturbations.

A.5 Additional Qualitative Comparisons

Figures 14 and 15 provide additional qualitative comparisons on FLIR-IVSR and M3FD. Existing methods often suffer from over-smoothed structures, unstable local textures, or noticeable intensity deviations. In contrast, our method recovers clearer object boundaries and finer structural details while maintaining more stable infrared intensity distributions. The corresponding intensity profiles

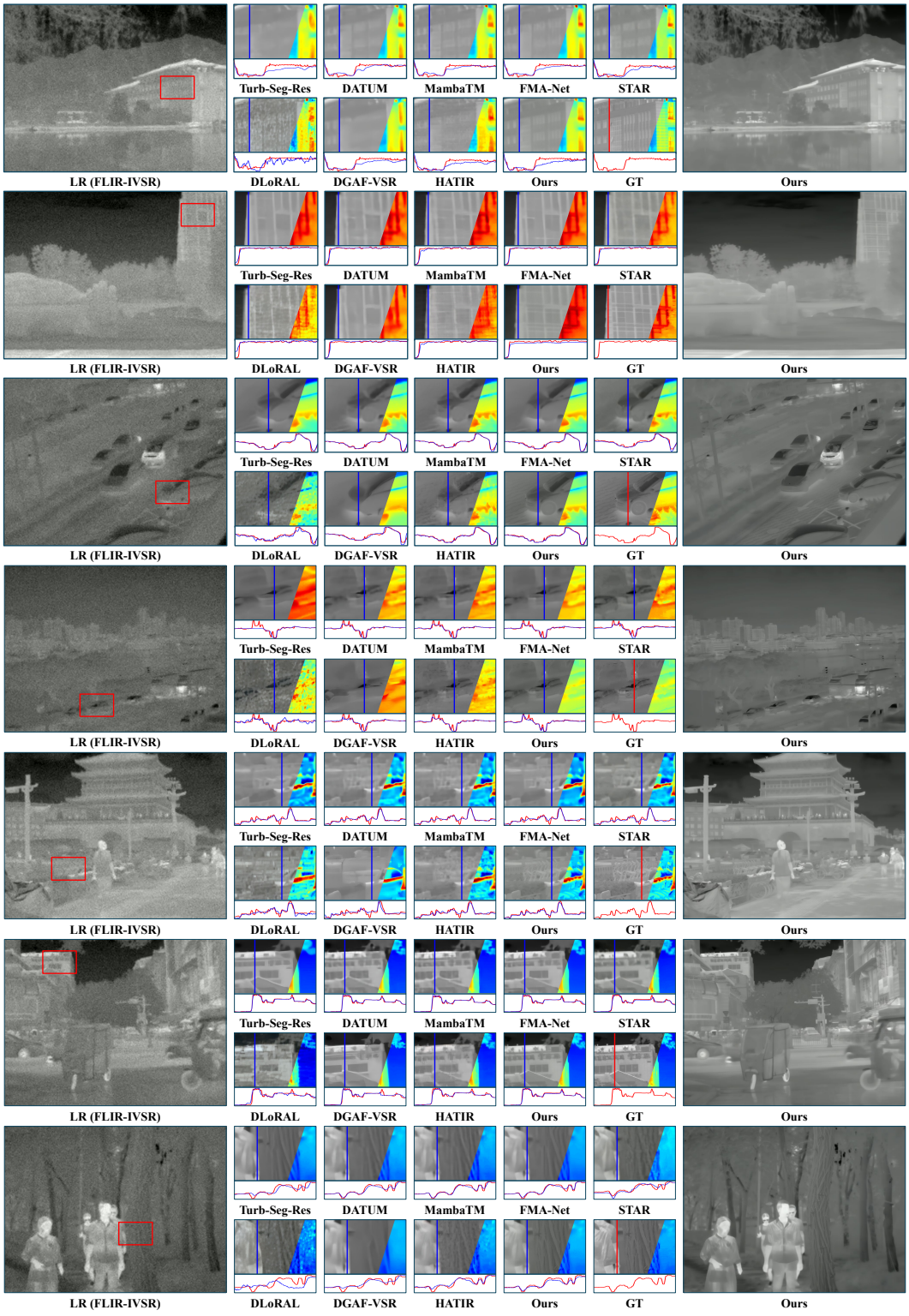

Figure 14: Qualitative comparison with SOTA methods on FLIR-IVSR datasets. The graph illustrates grayscale fluctuations along the blue-marked sampling line, and red-marked sampling line denotes the HR.

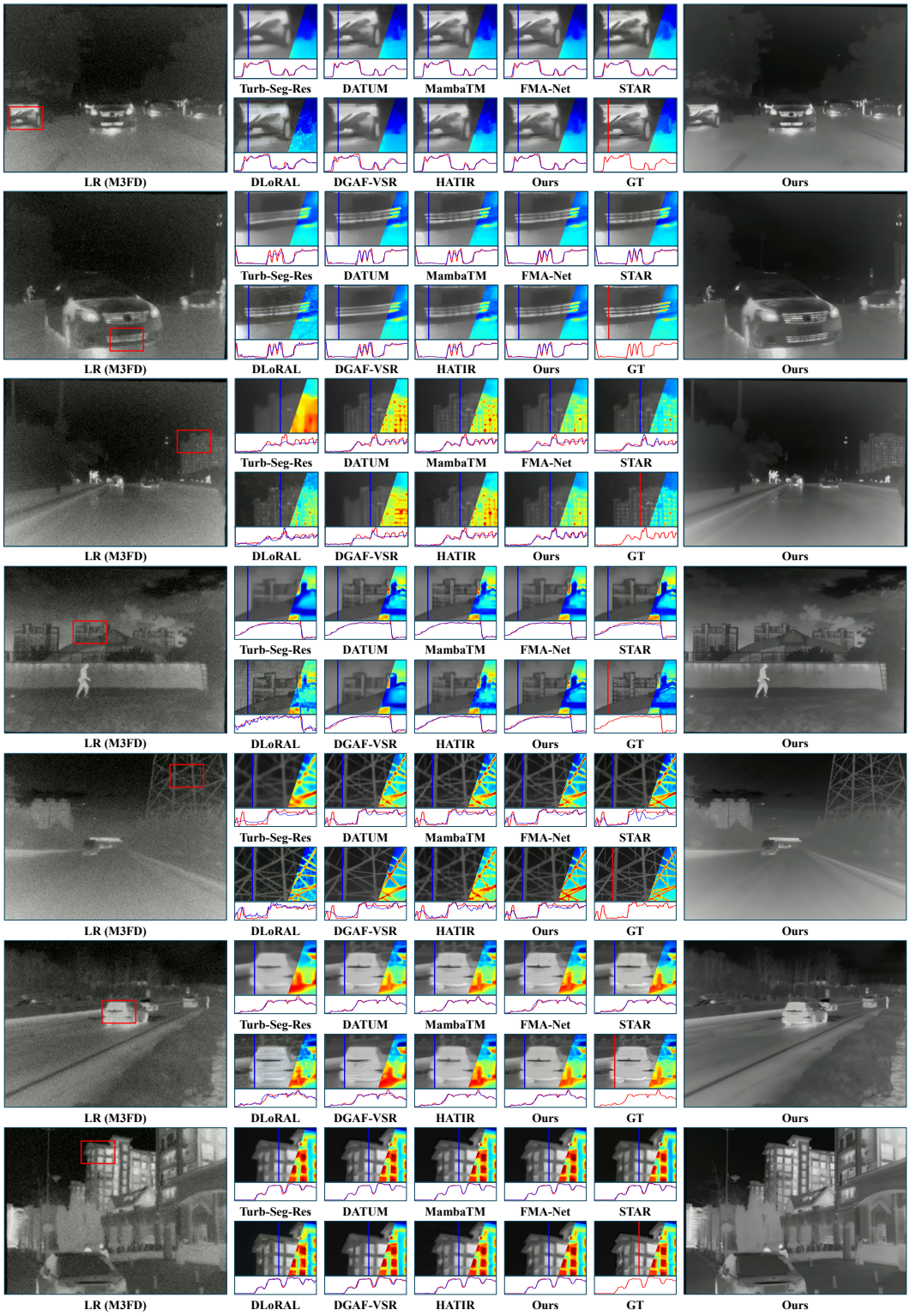

Figure 15: Qualitative comparison with SOTA methods on M^3FD datasets. The graph illustrates grayscale fluctuations along the blue-marked sampling line, and red-marked sampling line denotes the HR.

also follow the GT more closely, indicating improved structural fidelity and local radiometric consistency across different scenes.

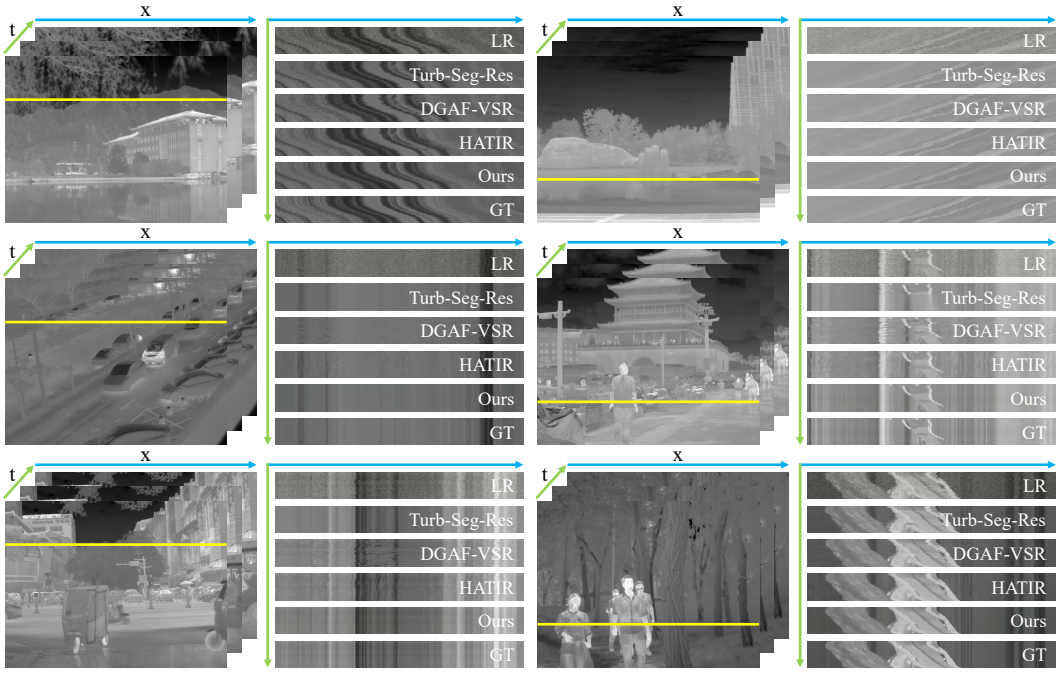

Figure 16: Temporal consistency comparison with SOTA methods on FLIR-IVSR datasets. The temporal profile is obtained by stacking the yellow line across frames.

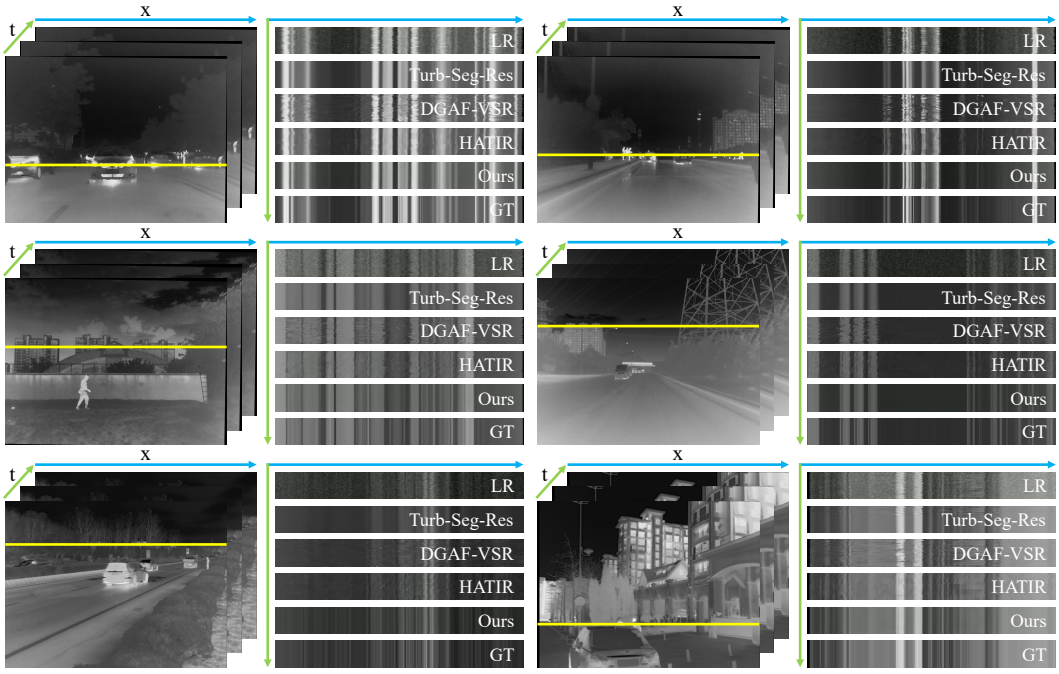

Figure 17: Temporal consistency comparison with SOTA methods on M^3FD datasets. The temporal profile is obtained by stacking the yellow line across frames.

Table 2: Efficiency comparison in terms of model parameters, diffusion steps, inference speed, and video quality. The inference speed is measured in frames per second (FPS) on 30-frame videos. "–" denotes a non-diffusion-based method. The best is in **bold**, while the second is underlined.

Metrics	FMA-Net	STAR	DLoRAL	DGAF-VSR	HATIR	Ours
Steps	–	15	1	50	50	1
Params (M)	**9.8**	2492.9	1771.6	1145.1	1579.9	5804.1
FPS	**5.871**	0.300	1.224	0.149	0.184	2.322
FasterVQA	0.2695	0.1838	0.2040	0.2994	0.2143	**0.3692**

A.6 Additional Temporal Consistency Comparisons

Figures 16 and 17 further compare temporal consistency on FLIR-IVSR and M3FD. The temporal profiles are obtained by stacking the pixels along the yellow sampling line across consecutive frames. Existing methods can partially suppress turbulence-induced variations, but noticeable temporal jitter, trajectory drift, or structural discontinuities remain. In comparison, our method produces smoother and more continuous temporal trajectories that are closer to the GT profiles, with fewer abrupt spatial shifts and intensity fluctuations. These results further demonstrate the effectiveness of motion–turbulence disentanglement for stable temporal reconstruction.



Figure 18: Failure cases under extremely strong turbulence. We regard $r_0 = 0.001$ as the 100% turbulence level and increase the degradation strength toward this extreme setting.

A.7 Additional Efficiency Analysis

As shown in Table 2, we compare different methods in terms of diffusion steps, model parameters, inference speed, and video quality. Although our method has the largest model size, it achieves the highest FasterVQA score of 0.3692 while maintaining an inference speed of 2.322 FPS, second only to the non-diffusion-based FMA-Net. This efficiency mainly benefits from the temporal compression mechanism of CogVideoX1.5 and the one-step diffusion inference used in our restoration framework. In contrast, STAR requires 15 sampling steps, while DGAF-VSR and HATIR require 50 steps, resulting in much lower inference speeds of 0.300, 0.149, and 0.184 FPS. Compared with DLoRAL, which also performs one-step diffusion inference, our method achieves a higher inference speed (2.322 vs. 1.224 FPS) and substantially better video quality (0.3692 vs. 0.2040 FasterVQA).

These results show that CAMO achieves a favorable balance between generative restoration quality and inference efficiency despite its relatively large parameter count.

A.8 Failure Cases under Extreme Turbulence

Figure 18 further investigates the performance boundary of CAMO under extremely strong turbulence. We regard $r_0 = 0.001$ as the 100% turbulence level, which represents an intentionally extreme condition far beyond the turbulence typically encountered in practical natural scenes. As the turbulence strength increases, the available structural information becomes progressively corrupted. Starting from approximately 82% turbulence, the restored results gradually become blurred and fine structures are increasingly difficult to recover. This degradation becomes more evident at 90%, where object boundaries and local details are noticeably weakened. At the 100% turbulence level, CAMO can still suppress part of the severe turbulence-induced distortion, but the underlying scene information has been heavily destroyed and a clear reconstruction can no longer be obtained. These results indicate that CAMO remains effective under severe turbulence, while its restoration capability reaches its limit when the input observations are excessively distorted and contain insufficient reliable structural information.

A.9 Qualitative Validation of Turbulence Disentanglement

Figure 19 qualitatively compares SynLR, the clean HR video, and HR degraded with random blur and distortion, where the corresponding error maps are computed against the real LR observations. Across different scenes, SynLR generally produces weaker and less structured residuals, indicating a closer match to the real turbulent observations. In contrast, the HR results exhibit clear residuals along object boundaries and local structures because they lack turbulence-induced distortions, while HR+Random still produces noticeable mismatched residuals despite introducing additional degradation. These results show that simply adding generic blur and distortion cannot reproduce the real turbulence pattern. By injecting the disentangled turbulence component into HR, SynLR better preserves the scene content while reproducing degradation characteristics consistent with the real LR observations, providing qualitative evidence that LMTD captures turbulence-specific perturbations.

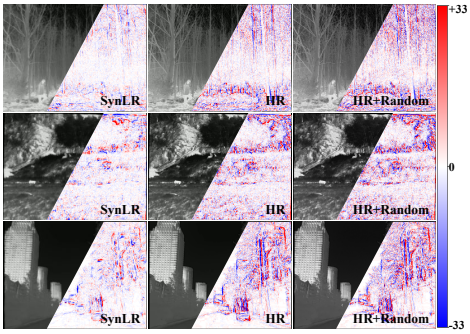

Figure 19: Visualization of synthetic low-quality infrared video samples.