Text-DiFuse: An Interactive Multi-Modal Image Fusion Framework based on Text-modulated Diffusion Model

Hao Zhang[*], **Lei Cao**[*], **Jaiyi Ma**[†]
Electronic Information School
Wuhan University
Wuhan, China
{zhpersonalbox, jyma2010}@gmail.com, whu.caolei@whu.edu.cn

Abstract

Existing multi-modal image fusion methods fail to address the compound degradations presented in source images, resulting in fusion images plagued by noise, color bias, improper exposure, *etc*. Additionally, these methods often overlook the specificity of foreground objects, weakening the salience of the objects of interest within the fused images. To address these challenges, this study proposes a novel interactive multi-modal image fusion framework based on the text-modulated diffusion model, called Text-DiFuse. First, this framework integrates feature-level information integration into the diffusion process, allowing adaptive degradation removal and multi-modal information fusion. This is the first attempt to deeply and explicitly embed information fusion within the diffusion process, effectively addressing compound degradation in image fusion. Second, by embedding the combination of the text and zero-shot location model into the diffusion fusion process, a text-controlled fusion re-modulation strategy is developed. This enables user-customized text control to improve fusion performance and highlight foreground objects in the fused images. Extensive experiments on diverse public datasets show that our Text-DiFuse achieves state-of-the-art fusion performance across various scenarios with complex degradation. Moreover, the semantic segmentation experiment validates the significant enhancement in semantic performance achieved by our text-controlled fusion re-modulation strategy. The code is publicly available at https://github.com/Leiii-Cao/Text-DiFuse.

1 Introduction

Due to constraints in imaging principles and hardware technology, single-modal images fall short of accurately and comprehensively describing scenes, thereby limiting their utility in subsequent tasks. Hence, image fusion technology emerges as essential in this context [60, 28]. It aims to integrate useful information from multi-modal images, producing high-quality visual results that enhance both human and machine perception of scenes. Currently, image fusion technology has been integrated into various tasks, significantly advancing performance in related fields such as autonomous driving [47], intelligent security [57, 25], and disease diagnosis [14].

Over recent decades, rapid advancements in deep learning have propelled significant progress in image fusion. Deep learning-based methods have surpassed traditional approaches in fusion performance by a considerable margin. In the historical context, the evolution of image fusion closely aligns with the

[*]Equal Contribution
[†]Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

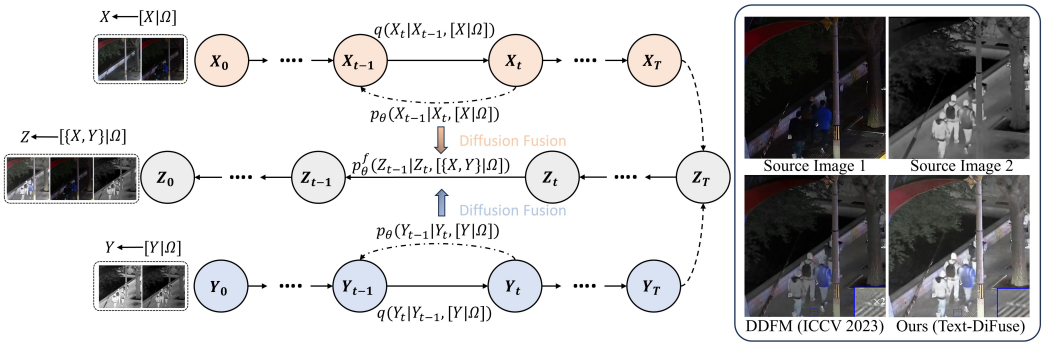

Figure 1: Our proposed explicit coupling paradigm of multi-modal information fusion and diffusion.

advancements in network paradigms. Autoencoders (AE) [19, 20], convolutional neural networks (CNN) [54, 49], generative adversarial networks (GAN) [29, 31], and Transformers [32, 44] represent a coherent axis of progress in image fusion. This phenomenon arises because there is no ground truth to supervise fusion learning. Therefore, fusion performance depends heavily on the continuous enhancement of the feature expression potential of neural network paradigms [55].

However, these methods falter in scenes with degradation, especially composite degradation, which we refer to as the **composite degradation challenge**. Essentially, current methods *prioritize multi-modal information integration without considering effective information restoration from degradation* [62, 26]. Last few years, the emergence of diffusion models has impressed many with their remarkable performance in visual restoration tasks [8, 16]. This prompts a natural question: Can diffusion models be utilized to tackle the challenge of multi-modal image fusion in scenes with complex degradation? According to the research status of diffusion model-based image fusion, two factors have hindered the implementation of this intuitive idea. First, visual restoration using diffusion models requires pairs of degraded and clean images, but in image fusion tasks, there is no clean fused image as ground truth due to the unsupervised nature of the task. Second, breaking through the paradigm that explicitly couples information fusion and diffusion is yet to be achieved.

Moreover, current fusion methods fail to account for the specificity of objects in the scene (*e.g.*, pedestrians, vehicles), applying the same fusion rules indiscriminately to both foreground and background. This lack of differentiation, termed the **under-customization objects limitation**, is unreasonable and may compromise the delineation of crucial objects [53, 64, 51]. Undoubtedly, maintaining the salience of foreground objects is crucial to satisfy both human and machine interest in them. This necessitates fusion models to possess the capability of interacting with users, achieving a "what you are interested in is what you get" approach.

To address the challenges of **composite degradation** and **under-customization objects** in multi-modal image fusion, we propose a novel interactive multi-modal image fusion framework based on the text-modulated diffusion model (Text-DiFuse). On the one hand, Text-DiFuse customizes a new explicit coupling paradigm of multi-modal information fusion and diffusion, eliminating complex degradation like color casts, noise, and improper lighting, as shown in Fig. 1. Specifically, it first applies independent conditional diffusion to data with compounded degradation, enabling degradation removal priors to be embedded into the encoder-decoder network. A fusion control module (FCM) is then embedded between the encoder and decoder to manage the integration of multi-modal features. This involves fusing multiple diffusion processes at the feature level, continuously aggregating multi-modal information while removing degradation during T-step sampling. To our knowledge, this is the first time information fusion is deeply and explicitly embedded in the diffusion process, effectively addressing compound degradation in image fusion tasks. On the other hand, to interactively enhance focus on objects of interest during diffusion fusion, we design a text-controlled fusion re-modulation strategy. This strategy incorporates text and a zero-shot location model to identify the objects of interest, thereby performing secondary modulation with the built-in prior to enhance their saliency. Thus, both the visual quality and semantic attributes of the fused image are significantly improved.

In summary, we make the following contributions:

• We propose a novel explicit coupling paradigm of information fusion and diffusion, solving the compound degradation challenge in the task of multi-modal image fusion.

• A text-controlled fusion re-modulation strategy is designed, allowing users to customize fusion rules with language to enhance the salience of objects of interest. This interactively improves the visual quality and semantic attributes of fused images.
• We evaluate our Text-DiFuse on extensive datasets and verify its advantages over state-of-the-art methods in terms of degradation robustness, generalization ability, and semantic properties.

2 Related Work

Deep Multi-modal Image Fusion. As mentioned earlier, the progress in deep multi-modal image fusion is closely tied to updates in neural network paradigms. Initially, AE-based fusion methods [19, 20] utilize pre-trained encoders and decoders alongside hand-crafted fusion rules, resulting in performance bottlenecks. Subsequent methods introduce CNN [6, 5] and Transformer [37, 58] for end-to-end fusion guided by specific unsupervised loss, yielding improved performance. The introduction of GAN is groundbreaking due to their inherently unsupervised nature, enabling the preservation of important multi-modal features [29, 56]. However, the instability of the adversarial game often leads to non-equilibrium appearances in fused images [30]. Furthermore, the diffusion model is highly anticipated for solving image fusion and is used in two main ways: injecting features into CNNs for separate fusion and diffusion [52], or treating multi-modal images as conditions for implicit fusion [63]. However, both methods fail to utilize the diffusion model's degradation removal capabilities and struggle with complex degradation. In contrast, our Text-DiFuse embeds feature fusion into the diffusion process, ensuring robustness and aggregation of multi-modal information. Additionally, we use text combined with a zero-shot location model for user-customized fusion, enhancing object salience.

Diffusion Model. The impressive performance of the diffusion model [41, 13] makes it top-notch in visual generation. It constructs a Markov chain by progressively adding noise forward, and then estimates the underlying data distribution and uses inverse sampling to generate images. This natural property of degradation removal has made the diffusion model excel in visual restoration tasks [46, 61]. However, the practical application of the diffusion model is hindered by its slow T-step continuous sampling. Recent efforts have focused on enhancing sampling efficiency and sample quality [50]. For example, DDIM [42] extends the original denoising diffusion probability model to non-Markovian scenarios, requiring only discrete time steps during sampling to reduce costs. Furthermore, iDDPM [35] introduces an enhanced denoising diffusion probability model, parameterizing backward variance through linear interpolation and training with mixed objectives to acquire knowledge of backward variance. This approach increases log-likelihood and accelerates sampling rates without compromising sample integrity. Therefore, our method employs iDDPM to expedite sampling while upholding the quality of fused images.

Zero-shot Location. Establishing connections between unseen and seen categories using semantic information [18], zero-shot location models [27, 33] can understand unseen images to identify and locate designated objects. Representative zero-shot location models include GLIP [22], OWL-VIT [33], and Grounding DINO [27]. Additionally, methods like DiffSeg [40], PADing [12], and SAM [17] achieve finer pixel-level object localization. These powerful zero-shot location techniques provide a solid foundation for the implementation of our text-controlled fusion re-modulation.

3 Methodology

3.1 Problem Statement and Modeling

Let us formally define the research problem of this work: achieving multi-modal image fusion under degraded scenes while supporting text-controlled fusion re-modulation of objects of interest. The multi-modal image pair captured under degraded conditions is formulated $[\{X, Y\}|\Omega]$, in which $\{X, Y\}$ denotes clean multi-modal images (*e.g.*, infrared and visible images), and Ω indicates composite degradation (*e.g.*, color casts, noise, and improper lighting). We aim to process degraded multi-modal images to obtain a clean fused image: $Z = \Gamma([\{X, Y\}|\Omega])$. The function Γ must handle two tasks: degradation removal R and information fusion F. There are two routes: concatenation ($\Gamma = R + F$) and coupling ($\Gamma = R \biguplus F$). Concatenation overlooks the intrinsic connection between degradation removal and fusion, leading to limited performance (see comparative experiments).

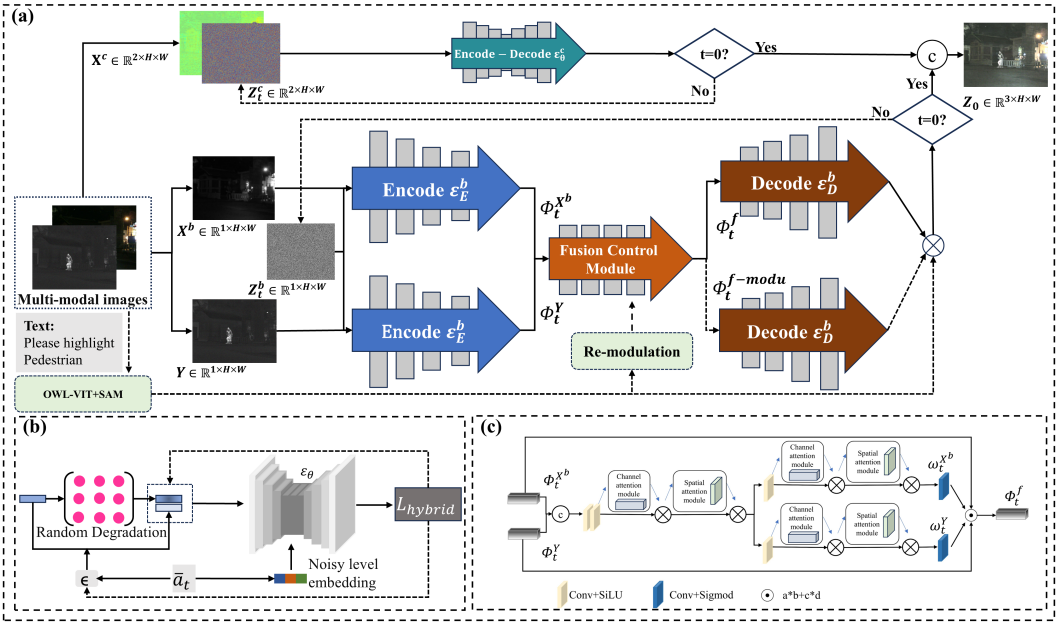

Figure 2: The pipeline of our Text-DiFuse. (a) The text-controlled diffusion fusion process; (b) training diffusion model for degradation removal; (c) the detailed structure of fusion control module.

Therefore, we choose the coupling route and introduce the diffusion model for degradation removal. Then, the key question becomes how to integrate the diffusion model with information fusion. Notably, the diffusion operates as a continuous process with multi-step sampling, making it difficult to incorporate fusion. To address this, we propose a novel **explicit coupling paradigm of information fusion and diffusion**, as illustrated in Fig. 2 (a). Specifically, the diffusion model is initially trained on data with compound degradation, incorporating the degradation removal prior into the encoder ε_E^b and decoder ε_D^b. During T-step reverse sampling, the multi-modal encoded features are continuously passed to the FCM for fusion, aiding in the reconstruction of the final fused image by the decoder. This essentially consolidates multiple diffusion processes into a single one, effectively integrating degradation removal and information fusion. However, the above methodology has not yet addressed the re-modulation of objects of interest in image fusion. To this end, we further develop a **text-controlled fusion re-modulation strategy** to highlight the objects of interest, thereby enhancing subsequent semantic decision performance. This strategy can be formulated as $Z = \Gamma([\{X, Y\}|\Omega], L)$, where L represents the user-defined language command. Specifically, we utilize text combined with the zero-shot location module to identify and locate the objects of interest. This knowledge triggers the re-modulation of diffusion fusion to enhance the saliency of the objects with in-built contrast-enhancement prior, thereby significantly improving perceptual quality for both humans and machines.

3.2 Explicit Coupling Paradigm of Information Fusion and Diffusion

Diffusion for Degradation Removal. Embedding the degradation removal prior into the encoder-decoder network of the diffusion model forms the foundation for diffusion fusion. We consider three primary types of degradation: color casts, noise, and improper lighting. They cover both nighttime and daytime negative imaging conditions and can be considered comprehensive to a certain extent. In our model, we separate brightness and chrominance components, and perform independent diffusion for them. Here, we describe and represent these two diffusion processes consistently. For clean components s (brightness or chrominance), the corresponding degraded versions $\Omega(s)$ involve composite degradation like color casts, noise, and improper lighting. These degraded components are fed into the encoder-decoder network as conditions. As shown in Fig. 2 (b), in the forward diffusion process, the original clean component s, denoted as s_0 at step 0, is progressively added with Gaussian noise over T steps to obtain s_T. In the reverse sampling process, the encoder-decoder network is guided to estimate the mean $\mu_\theta(s_t, \Omega(s), t)$ and variance $\sum_\theta(s_t, \Omega(s), t)$ of s_{t-1},

progressively approaching the clean component with the conditions $\Omega(s)$. Drawing upon iDDPM [35], the optimization can be defined as:

$$\nabla_\theta \parallel \epsilon_t - \epsilon_\theta(s_t, \Omega(s), t) \parallel^2 + \lambda D_{\mathrm{KL}}(q(s_{t-1}|s_t, s_0, \Omega(s))||p_\theta(s_{t-1}|s_t, \Omega(s))), t > 1 \tag{1a}$$

$$\nabla_\theta \parallel \epsilon_t - \epsilon_\theta(s_t, \Omega(s), t) \parallel^2 - \lambda \log p_\theta(s_0|s_1, \Omega(s)), t = 1 \tag{1b}$$

where ∇_θ means optimization by gradient descent, ϵ_t denotes the added noise in the forward diffusion process, D_{KL} is the regularization term based on KL divergence, and q and p_θ represent the prior and posterior probability distributions, respectively. ϵ_θ is the noise predictor.

Fusion Control Module for Information Fusion. For information fusion, we design an FCM to aggregate encoded features during the diffusion process, as shown in Fig. 2 (c). It primarily consists of convolution layers with CBAM [45]. In them, integrating spatial and channel attention mechanisms helps perceive the importance of multi-modal features on a wider scale, promoting rational feature fusion. To reduce the solution space, FCM generates weight coefficients for fusion rather than directly predicting fused features, allowing faster convergence in multi-step sampling of the diffusion process.

Diffusion Fusion. Everything is ready, and now we can seamlessly integrate information fusion with diffusion, termed diffusion fusion. Given the degraded multi-modal image pairs $[\{X, Y\}|\Omega]$, we assume X is a color image and Y is a grayscale image. This assumption aligns with most multi-modal image fusion scenarios, such as visible and infrared image fusion, and MRI and PET image fusion. By separating components, we obtain the brightness component $[X^b|\Omega]$ and chrominance component $[X^c|\Omega]$. First, using the trained diffusion model ε_θ^c, we remove the degradation in the chrominance component $[X^c|\Omega]$, obtaining a clean and reasonable chrominance component X_0^c.

Then, the processing of paired $[\{X^b, Y\}|\Omega]$ involves diffusion fusion, which achieves simultaneous degradation removal and information fusion. Specifically, given randomly sampled Gaussian noise $Z_T^b \sim N(0, I)$, $[\{X^b, Y\}|\Omega]$ are regarded as the condition input to the shared encoder ε_E^b in another diffusion model, obtaining features $[\{\Phi_t^{X^b}, \Phi_t^Y\}|\Omega]$ at step t:

$$[\{\Phi_t^{X^b}, \Phi_t^Y\}|\Omega] = \varepsilon_E^b(Z_t^b, [\{X^b, Y\}|\Omega], t), t \in \{T, \cdots 0\}. \tag{2}$$

The FCM generates weight coefficients $\{\omega_t^{X^b}, \omega_t^Y\}$ for fusing these multi-modal features:

$$[\Phi_t^f|\Omega] = [\{\Phi_t^{X^b}, \Phi_t^Y\}|\Omega] \odot \{\omega_t^{X^b}, \omega_t^Y\}, \tag{3}$$

where $[\Phi_t^f|\Omega]$ is the fused feature with residual degradation at step t, and $\odot$ denotes the Hadamard product. Subsequently, the fused feature is fed into the decoder ε_D^b to predict the contained noise $\epsilon_\theta(t)$ at step t, and the relevant variable $\upsilon_\theta(t)$ for learning the variance:

$$\epsilon_\theta(t), \upsilon_\theta(t) = \varepsilon_D^b([\Phi_t^f|\Omega], t). \tag{4}$$

Then, the mean and variance of Z_{t-1}^b can be obtained according to:

$$\mu_\theta(Z_t^b, [\{X^b, Y\}|\Omega], t) = \frac{1}{\sqrt{\alpha_t}}(Z_t^b - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(t)), \tag{5}$$

$$\textstyle\sum_\theta(Z_t^b, [\{X^b, Y\}|\Omega], t) = exp(\upsilon_\theta(t)\log\beta_t + (1-\upsilon_\theta(t))\log\tilde{\beta}_t), \tag{6}$$

where β_t represents the variance associated with the forward diffusion process, using the notation $\alpha_t = 1 - \beta_t$ and $\overline{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Additionally, we parameterize the variance between β_t and $\tilde{\beta}$ in the logarithmic domain using the technique of iDDPM [35], where $\tilde{\beta} = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$. Then, Z_{t-1}^b can be computed according to:

$$Z_{t-1}^b = \mu_\theta(Z_t^b, [\{X^b, Y\}|\Omega], t) + \sqrt{\textstyle\sum_\theta(Z_t^b, [\{X^b, Y\}|\Omega], t)} \cdot z, \tag{7}$$

where z denotes the randomly sampled Gaussian noise $z \sim N(0, I)$ when $t > 1$, otherwise $z = 0$. According to Eqs. (4)-(7), each sample will derive an $\hat{Z}_0^b$:

$$\hat{Z}_0^b(Z_t^b, \epsilon_\theta(t)) = \frac{Z_t^b - \sqrt{1-\overline{\alpha}}\epsilon_\theta(t)}{\sqrt{\overline{\alpha}}}. \tag{8}$$

Notably, $\hat{Z_0^b}(Z_t^b, \epsilon_\theta(t))$ indicates the corresponding fake final fused image that is derived from the results of any step of sampling. Therefore, we construct constraints to guide the FCM in retaining beneficial information during the diffusion fusion process. Considering pixel intensity and gradient as two basic elements that describe images, we specify intensity loss $\mathcal{L}_{int}$ and gradient loss $\mathcal{L}_{grad}$ to emphasize the preservation of significant contrast and rich texture:

$$\mathcal{L}_{int} = \| \ |\hat{Z_0^b}(Z_t^b, \epsilon_\theta(t))| - \max\{|X^b|, |Y|\} \ \|, \tag{9}$$

$$\mathcal{L}_{grad} = \| \ \nabla\hat{Z_0^b}(Z_t^b, \epsilon_\theta(t)) - \max\{\nabla X^b, \nabla Y\} \ \|, \tag{10}$$

where $\max$ is the maximum function, ∇ is the Sobel gradient operator, X^b and Y are clean source components after diffusion. The total loss is summarized as:

$$\mathcal{L}_{diff-fusion} = \gamma_{int}\mathcal{L}_{int} + \gamma_{grad}\mathcal{L}_{grad}, \tag{11}$$

where γ_{int} and γ_{grad} control the balance of these terms, set to 1 and 0.2, respectively. After optimization, we obtain the clean fused brightness component $Z^b = Z_0^b$. The purified chrominance component X^c is used as the fused image's chrominance component: $Z^c = X_0^c$. Finally, stitching Z^b and Z^c yields the final fused image Z with accurate colors, minimal noise, and proper lighting. Through these designs, information fusion and diffusion have been fully and explicitly coupled, achieving multi-modal image fusion while removing compound degradation.

3.3 Text-controlled Fusion Re-modulation Strategy

The above diffusion fusion constitutes the **basic version** of our method. Now, we aim to expand it into a **modulatable version**, allowing users to re-modulate the fusion process based on personalized needs, enhancing the perception of objects of interest. Firstly, we use state-of-the-art zero-shot localization models to identify and locate objects of interest based on text commands. Specifically, we introduce OWL-VIT [33] for detecting objects of interest with open-word input. Then, SAM [17] provides pixel-level positioning of these objects, obtaining the mask M. Subsequently, M is fed into the re-modulation block to generate fusion modulation coefficients $\{\kappa^{X^b}, \kappa^Y\}$. This block incorporates a built-in contrast-enhancement prior, aiming to maximize the contrast between the object area and the background in the fused image, thus improving the salience of the objects. Consequently, the multi-modal feature fusion in the diffusion process changes from Eq. (3) to:

$$[\Phi_t^{f-modu}|\Omega] = [\{\Phi_t^{X^b}, \Phi_t^Y\}|\Omega] \odot \{\omega_t^{X^b}, \omega_t^Y\} \odot \{\kappa^{X^b}, \kappa^Y\}. \tag{12}$$

In non-object areas, the original distribution of diffusion fusion should be maintained:

$$\{Z_t^b\}^{re-modu} = (1 - M) \cdot Z_t^b + M \cdot \{Z_t^b\}^{mod}. \tag{13}$$

The modulated fused image enhances the saliency of objects compared to the before, making it more suitable for subsequent advanced tasks. We prove this in the re-modulation verification section.

4 Experiments

Configuration. We evaluate our method on two typical multi-modal image fusion scenarios: infrared and visible image fusion (IVIF) and medical image fusion (MIF). For IVIF, we use the MSRS dataset [43], with 485 training and 100 testing image pairs. For MIF, we use the *Harvard medicine dataset*[3] with 160 training and 50 testing image pairs, covering CT-MRI, PET-MRI, and SPECT-MRI. Data augmentation like random flipping and cropping increases the training pairs to $12,888$ for IVIF and $6,408$ for MIF. Besides, generalization is evaluated on 60 pairs from LLVIP [15] and 25 pairs from RoadScene [49]. Competitors include 9 methods: RFN-Nest [20], GANMcC [31], SDNet [53], U2Fusion [49], TarDAL [25], DeFusion [23], LRRNet [21], DDFM [63], and MRFS [58]. Five metrics are used: EN [39], AG [3], SD [38], SCD [2], and VIF [11]. The Adam optimizer with a learning rate of $2e^{-5}$ is used for parameter updates. Experiments are conducted on an NVIDIA RTX 3090 GPU and a 3.80 GHz Intel i7-10700K CPU.

[3] `https://www.med.harvard.edu/AANLIB/home.html`

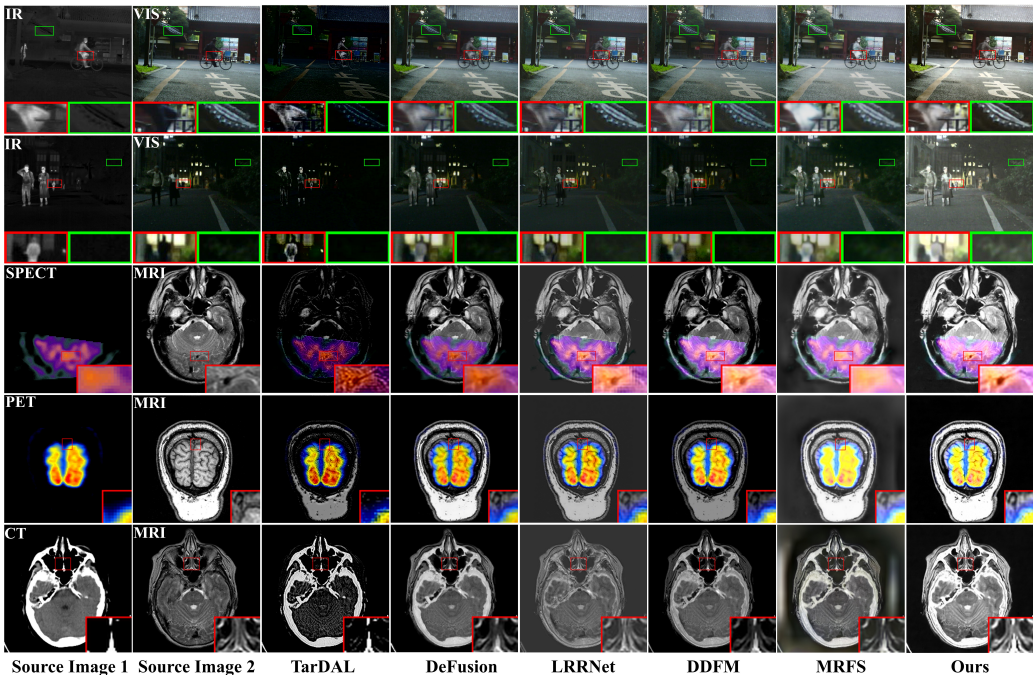

Figure 3: Visual comparison of image fusion methods.

Table 1: Quantitative comparison of image fusion methods. **Blod**: the best; underline: second best.

Methods	MSRS DataSet					Havard Medicine Dataset				
	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
RFN-Nest (InF'21)	5.89	1.84	26.03	1.41	0.63	5.34	4.07	63.65	1.58	0.43
GANMcC (TIM'21)	6.03	1.91	25.58	1.37	0.65	5.38	5.13	55.00	1.11	0.43
SDNet (IJCV'21)	4.90	2.33	16.35	0.91	0.49	5.56	6.22	46.64	0.48	0.38
U2Fusion (TPAMI'22)	5.19	2.46	24.82	1.21	0.52	5.22	6.08	53.20	1.03	0.41
TarDAL (CVPR'22)	3.30	2.04	18.52	0.63	0.15	5.66	5.13	41.94	1.12	0.18
DeFusion (ECCV'22)	6.22	2.31	32.34	1.36	0.75	4.96	4.47	55.45	0.93	0.48
LRRNet (TPAMI'23)	5.89	2.19	26.64	0.75	0.52	5.34	5.39	45.89	0.59	0.40
DDFM (ICCV'23)	5.81	2.65	24.98	1.37	0.62	5.00	5.04	63.53	1.59	0.48
MRFS (CVPR'24)	6.91	2.67	40.95	1.23	0.75	**7.24**	4.41	70.75	1.53	0.41
Ours (Text-DiFuse)	**7.08**	**3.31**	**47.44**	**1.44**	**0.76**	6.44	**7.31**	**80.19**	**1.69**	**0.49**

Comparative Experiments. We first compare the basic version of our Text-DiFuse with current state-of-the-art fusion methods, and the qualitative results are shown in Fig. 3. The first two rows depict IVIF results, demonstrating our method's ability to correct color casts, restore scene information under low-light conditions, and suppress noise. The last three rows display MIF results, which show that our method can highlight physiological structure information while maintaining functional distribution. In contrast, competitors are unable to achieve such information recovery and still suffer significantly weakened appearance. The quantitative results in Table 1 demonstrate our method's advantages over other fusion techniques. For further fairness, we introduce state-of-the-art low-light enhancement (CLIP-LIT [24]), denoising (SDAP [36]), and white balance (AWB [1]) algorithms as the pre-processing steps for these competitors, with results presented in Fig. 4 and Table 2. Clearly, our method still outperforms these comparative methods. This is because these added pre-processing steps for information recovery are entirely independent of information fusion, so they cannot mine habits that are more conducive to modal complementarity, leading to limited performance.

Generalization Evaluation. Next, we directly test the model trained on the MSRS dataset on the LLVIP and RoadScene datasets to evaluate the generalization ability of the proposed method. We select a daytime scene with overexposure and a low-light nighttime scene, and the qualitative results are shown in Fig. 5. It can be observed that our Text-DiFuse still maintains high-quality degradation removal and information fusion capabilities. In particular, it has two-way information recovery functions such as overexposure correction and low-light enhancement, producing visually satisfying

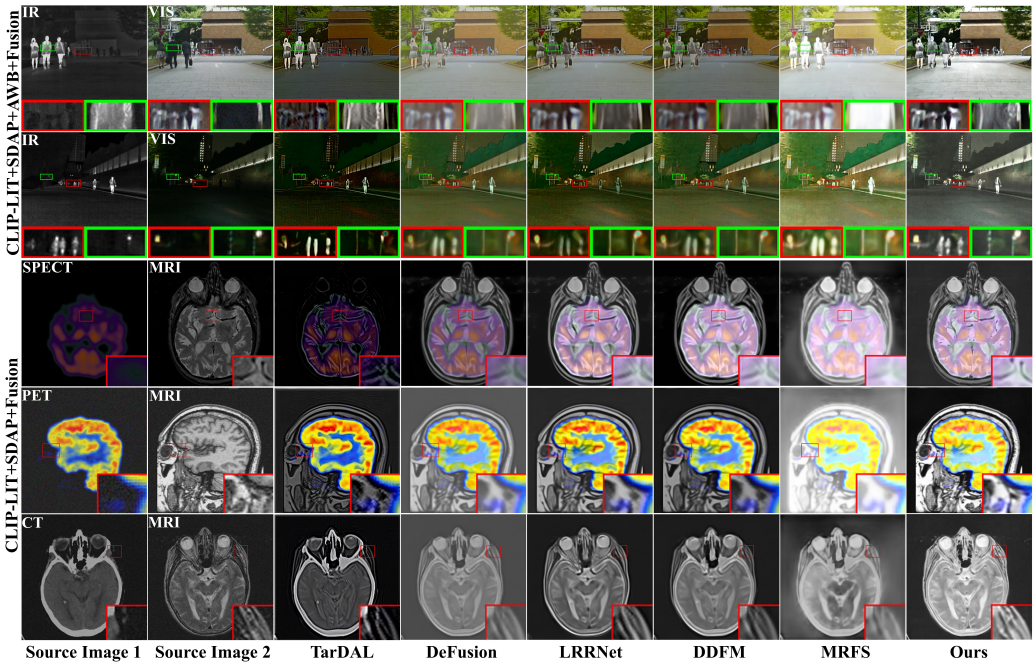

Figure 4: Visual comparison of enhancement plus image fusion methods.

Table 2: Quantitative comparison of enhancement plus image fusion methods.

Methods		MSRS Dataset					Havard Medicine Dataset				
		EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
CLIP-LIT SDAP AWB	RFN-Nest	6.43	2.23	27.17	1.38	0.60	5.72	4.11	77.46	1.64	0.35
	GANMcC	6.25	2.06	24.55	1.31	0.57	5.80	5.28	66.37	1.19	0.31
	SDNet	5.84	2.99	20.26	1.08	0.52	5.91	6.00	60.83	1.15	0.30
	U2Fusion	6.55	3.55	29.08	1.32	0.58	5.68	6.09	71.59	1.56	0.32
	TarDAL	5.29	**4.42**	25.22	1.00	0.35	6.11	4.81	36.54	0.69	0.23
	DeFusion	6.31	2.07	25.52	1.16	0.59	6.08	4.27	67.77	1.38	0.35
	LRRNet	6.55	2.68	31.19	1.13	0.54	5.86	5.23	62.91	1.34	0.21
	DDFM	6.39	2.43	26.40	1.16	0.60	5.70	4.48	77.40	1.64	0.35
	MRFS	6.84	2.86	32.28	1.28	0.58	**7.18**	4.19	**87.53**	1.50	0.31
Ours (Text-DiFuse)		**7.08**	3.31	**47.44**	**1.44**	**0.76**	6.44	**7.13**	80.19	**1.69**	**0.49**

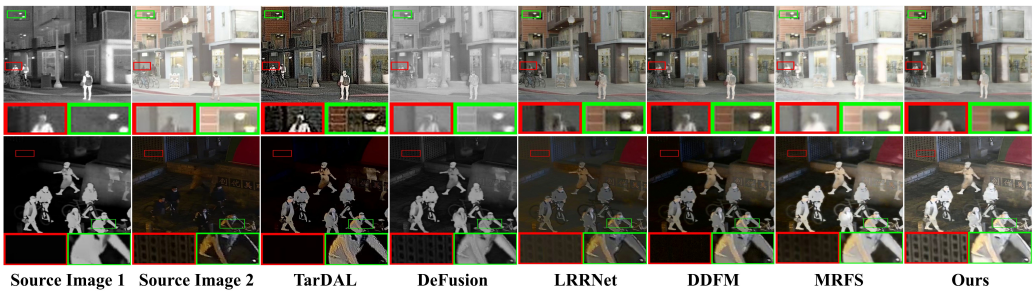

Figure 5: Visual results of generalization evaluation.

fused results. By comparison, other competitors lose useful information masked by overexposure or low light. We further prove the good generalization ability of our method in Table 3. In general, these results indicate that our Text-DiFuse can be applied more reliably in real scenarios.

Re-modulation Verification. We verify the performance gains brought by our text-controlled fusion re-modulation strategy on the MFNet dataset [10]. We select 6 state-of-the-art RGB-T segmentation methods for comparison, *i.e.*, MFNet [10], FEANet [4], EGFNet [7], CMX [59], GMNet [65], and MDRNet [48]. Besides, we train SegNext [9] on infrared images, visible images, the fused

Table 3: Quantitative comparison of generalization ability.

Methods	LLVIP Dataset					RoadScene Dataset				
	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑	EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
RFN-Nest	6.37	2.24	26.66	1.63	0.73	7.37	2.62	46.77	1.66	0.58
GANMcC	6.24	2.09	27.02	1.59	0.65	7.24	3.58	43.68	1.39	0.57
SDNet	6.00	2.74	23.05	1.24	0.62	7.18	4.86	40.63	1.16	0.66
U2Fusion	5.52	2.69	21.12	1.32	0.61	7.32	4.92	43.99	1.49	0.66
TarDAL	3.85	2.59	23.05	0.92	0.22	7.35	**11.84**	52.30	0.97	0.47
DeFusion	6.46	2.36	29.48	1.48	0.82	6.97	2.85	35.96	0.98	0.59
LRRNet	5.67	2.28	19.49	1.06	0.57	7.19	3.55	44.01	1.47	0.58
DDFM	6.46	3.51	30.64	1.72	0.70	7.30	3.63	44.19	1.57	0.65
MRFS	7.00	2.34	40.90	1.67	0.86	7.18	2.70	46.57	1.20	0.52
Ours (Text-DiFuse)	**7.08**	**3.99**	**41.78**	**1.73**	**0.87**	**7.46**	2.96	**52.84**	**1.67**	**0.66**

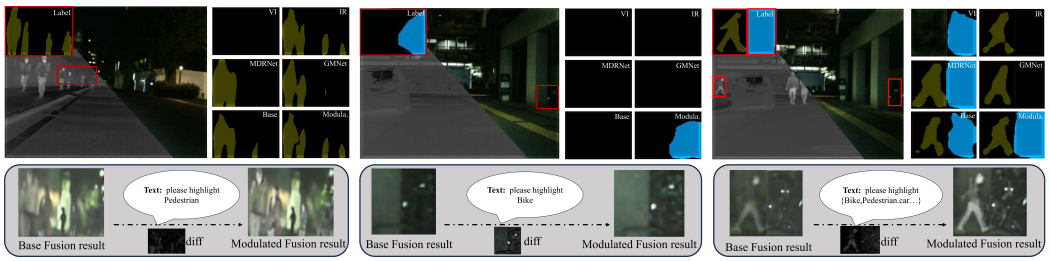

Figure 6: Visual results of re-modulation verification.

Table 4: Quantitative verification of re-modulation on semantic segmentation.

Segmentation	Source	Background	Car	Person	Bike	Curve	Car Stop	Cuardrail	Color cone	Bump	mIoU
MFNet	RGB-T	96.26	60.95	53.44	43.14	22.94	9.44	0.00	18.80	23.47	36.49
FEANet	RGB-T	98.00	87.41	70.30	62.74	45.33	29.80	0.00	29.07	48.95	55.28
EGFNet	RGB-T	98.01	87.84	71.12	61.08	46.48	22.10	6.64	55.35	47.12	54.76
CMX-B2	RGB-T	97.39	84.23	67.12	56.93	41.11	39.56	18.94	48.84	54.42	58.31
GMNet	RGB-T	98.00	86.46	73.05	61.72	43.96	42.25	14.52	48.70	47.72	57.34
MDRNet	RGB-T	97.90	87.07	69.81	60.87	47.80	34.18	8.21	50.18	54.98	56.78
SegNext-Base	IR	97.79	84.89	70.73	56.29	41.94	24.15	7.60	35.91	48.64	51.99
	VI	97.93	88.29	62.42	63.67	35.34	36.95	5.77	51.20	47.74	54.37
	Our basis	98.11	88.66	70.00	64.30	43.07	30.25	11.95	55.14	56.27	57.53
	Our modulatable	98.18	88.32	72.23	65.02	44.79	33.11	13.76	56.32	55.97	**58.63**

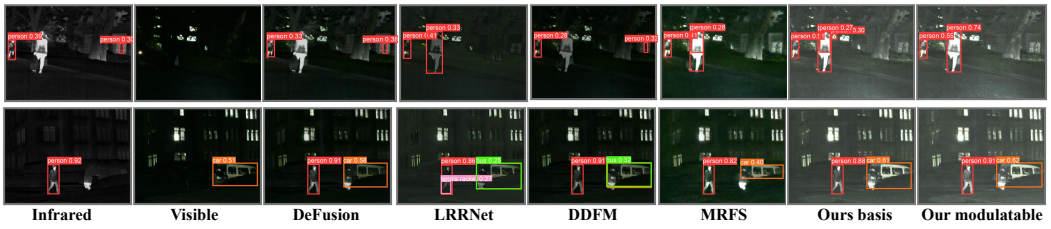

Figure 7: Visual verification in detection scenario.

images generated by the basic version of our method, and the modulatable version respectively to achieve segmentation. It can be seen from Fig. 6 that different language commands derive customized fused results, which promote the completeness and accuracy of semantic segmentation while visually highlighting the objects of interest. The quantitative results in Table 4 further prove that our re-modulation strategy can improve the semantic attributes, achieving the best segmentation scores.

Semantic Verification on Detection. We further verify the semantic gain brought by text modulation on the object detection task. Specifically, the MSRS dataset [43] is used, which includes pairs of infrared and visible images with two types of detection labels: person and car. Therefore, the text instruction is formulated as: "Please highlight the person and car", which guides our method to enhance the representation of these two types of objects in the fused image. Then, we adopt the YOLO-v5 detector to perform object detection on infrared images, visible images, and fused images generated by various image fusion methods. The visual results are presented in Fig. 7, in which more complete cars and people can be detected from our fused images while showing higher class

Table 5: Quantitative verification in detection scenario.

Detection	IR	VIS	DeFusion	LRRNet	DDFM	MRFS	Our basis	Our modulatable
mAP@0.5	71.9	74.8	86.6	86.3	88.6	82.0	87.3	**89.7**
mAP@[0.5:0.95]	48.4	47.3	60.1	58.9	59.4	53.2	56.3	**60.9**

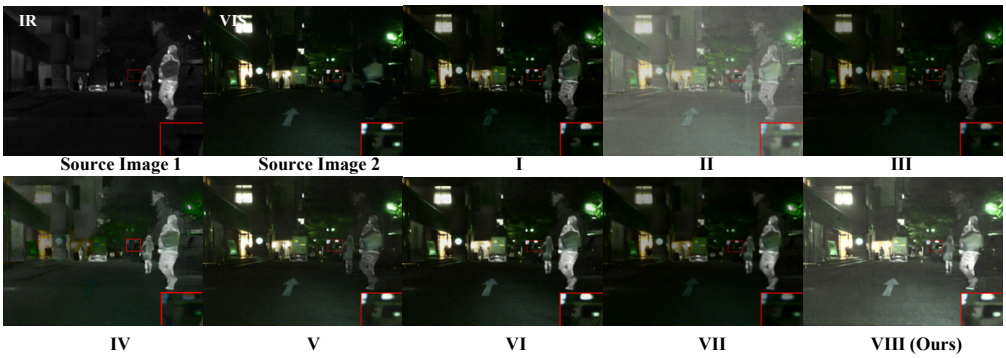

Figure 8: Visual results of ablation studies.

Table 6: Quantitative results of ablation studies.

Index	Diff.	$\mathcal{L}_{int}$	$\mathcal{L}_{grad}$	FCM	EN	AG↑	SD↑	SCD↑	VIF↑
I	✓	✓	✓	✗/max	5.71	1.90	25.66	1.30	0.49
II	✓	✓	✓	✗/add	6.08	1.99	20.14	1.11	0.58
III	✓	✓	✓	✗/mean	5.60	1.49	20.71	1.15	0.56
IV	✓	✓	✓	✗/variance	5.91	1.90	27.59	0.91	0.46
V	✓	✓	✗	✓	6.20	2.75	33.78	1.42	0.73
VI	✓	✗	✓	✓	6.67	3.25	45.60	1.43	0.76
VII	✗/AE	✓	✓	✓	6.37	2.26	37.48	1.42	0.69
VIII	✓	✓	✓	✓	**7.08**	**3.31**	**47.44**	**1.44**	**0.76**

confidence. Furthermore, we provide quantitative detection results in Table 5. It can be seen that the highest average accuracy is obtained from our fused images, demonstrating the benefits of text modulation. Overall, these results indicate that text control indeed provides significant semantic gains, benefiting downstream tasks.

Ablation Studies. We conduct ablation studies to verify the effectiveness of specific designs, involving eight variants: **I**: removing FCM with using maximum rule; **II**: removing FCM with using addition rule; **III**: removing FCM with using mean rule; **IV**: removing FCM with using variance-based rule [34]; **V**: removing $\mathcal{L}_{grad}$; **VI**: removing $\mathcal{L}_{int}$; **VII**: removing diffusion with using AE route; **VIII**: our full model. The visual results in Fig. 7 show that removing any of these designs results in a reduction of visual satisfaction. The quantitative scores in Table 5 also support this view. Overall, these designs in our Text-DiFuse collectively guarantee advanced fusion performance.

5 Conclusion

This paper proposes a new interactive multi-modal image fusion framework based on the text-modulated diffusion model. On the one hand, it is the first to develop an explicit coupling paradigm for information fusion and diffusion models, achieving the integration of multi-modal beneficial information while removing composite degradation. On the other hand, a text-controlled fusion re-modulation strategy is designed. It incorporates text combined with the zero-shot location module into the diffusion fusion process, supporting users' language control to enhance the perception of objects of interest. Extensive experiments demonstrate that our method achieves better performance than current methods, effectively improving the visual quality and semantic attributes of fused results.

6 Acknowledgement

This work was supported by NSFC (62276192).

References

[1] Mahmoud Afifi, Marcus A Brubaker, and Michael S Brown. Auto white-balance correction for mixed-illuminant scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1210–1219, 2022.

[2] V Aslantas and E Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-International Journal of Electronics and Communications*, 69(12):1890–1896, 2015.

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

[4] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4467–4473, 2021.

[5] Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3333–3348, 2020.

[6] Wang Di, Liu Jinyuan, Fan Xin, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3508–3515, 2022.

[7] Shaohua Dong, Wujie Zhou, Caie Xu, and Weiqing Yan. Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing. *IEEE Transactions on Intelligent Transportation Systems*, 25(1): 657–669, 2023.

[8] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023.

[9] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022.

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

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

[12] Shuting He, Henghui Ding, and Wei Jiang. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11238–11247, 2023.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[14] Ling Huang, Thierry Denoeux, Pierre Vera, and Su Ruan. Evidence fusion with contextual discounting for multi-modality medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 401–411, 2022.

[15] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[16] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[18] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

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

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

[21] Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023.

[22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[23] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision*, pages 719–735, 2022.

[24] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023.

[25] 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 Ppattern Recognition*, pages 5802–5811, 2022.

[26] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132:1748–1775, 2023.

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

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

[29] 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.

[30] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[31] Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020.

[32] 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.

[33] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *Proceedings of the European Conference on Computer Vision*, pages 728–755, 2022.

[34] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. *arXiv preprint arXiv:2404.09977*, 2024.

[35] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of th International Conference on Machine Learning*, pages 8162–8171, 2021.

[36] Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, and Chao Ren. Random sub-samples generation for self-supervised real image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12150–12159, 2023.

[37] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023.

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

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

[40] Zhihao Shuai, Yinan Chen, Shunqiang Mao, Yihan Zho, and Xiaohong Zhang. Diffseg: A segmentation model for skin lesions based on diffusion difference. *arXiv preprint arXiv:2404.16474*, 2024.

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, pages 2256–2265, 2015.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, pages 1–20, 2021.

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

[44] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022.

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018.

[46] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023.

[47] Jun Xie, Jiazhen Dou, Liyun Zhong, Jianglei Di, and Yuwen Qin. A dual-mode intensity and polarized imaging system for assisting autonomous driving. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024.

[48] Lijie Xie, Fubao Zhu, and Ni Yao. Mdr-net: Multiscale dense residual networks for liver image segmentation. *IET Image Processing*, 17(8):2309–2320, 2023.

[49] 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.

[50] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[51] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[52] Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32: 5705–5720, 2023.

[53] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021.

[54] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12797–12804, 2020.

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

[56] Hao Zhang, Jiteng Yuan, Xin Tian, and Jiayi Ma. Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators. *IEEE Transactions on Computational Imaging*, 7:1134–1147, 2021.

[57] Hao Zhang, Tang Linfeng, Xinyu Xiang, Xuhui Zuo, and Jiayi Ma. Dispel darkness for better fusion: A controllable visual enhancer based on cross-modal conditional adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

[59] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023.

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

[61] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *Advances in Neural Information Processing Systems*, 36: 49703–49714, 2024.

[62] 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 Ppattern Recognition*, pages 5906–5916, 2023.

[63] 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*, pages 8082–8093, 2023.

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

[65] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: The main claims of this work are to solve the problem of multi-modal image fusion in degraded environments and to interactively achieve attention and enhancement of objects of interest. These claims correspond to our contributions and are verified in the methodological and experiment sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We discuss the limitations of this work in the supplementary material. Specifically, our proposed method has lower efficiency, which originates from the slow sampling process of the diffusion model. In the future, we will study the acceleration strategy of the diffusion model and further improve its integration in multi-modal image fusion to increase operating efficiency.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: This paper does not include theoretical results. Actually, it is a pioneering paradigm in which diffusion theory is explicitly embedded in multi-modal image fusion, achieving compound degradation removal while high-quality cross-modal information integration.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: This paper aims to propose a new multi-modal image fusion algorithm that can adapt to degraded scenes. It is completely reproducible. First, we describe in detail the experimental conditions of this work in the experimental configuration section, including datasets, training details, and computing hardware. Second, we provide a URL link to the code of our method in the abstract section, including all the necessary functions for training and inference. We also provide a README file within it to outline the necessary environment configuration for running, as well as detailed usage instructions. Together these ensure the experimental result reproducibility in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes] .

   Justification: In the abstract section, we provide a URL link to the code of our Text-DiFuse, in which all the necessary functional code for training and inference is included. Besides, we also provide a README file within it to outline the configuration of the environment required for running, as well as detailed usage instructions. In addition, all datasets used in this work are publicly available and we have provided accurate citations for them. We also describe the processing and partitioning of these datasets in the experimental configuration section.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: We describe in detail the experimental conditions of this work in the experimental configuration section, including datasets, training details, and computing hardware.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: The experimental results reported in this paper are the average of a large number of test results in the dataset. Therefore, they are statistically significant, being able to support and validate the contributions and claims of this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

   Justification: In the experimental configuration section, we provide the computing resources required to reproduce the experiments in this paper, including an NVIDIA RTX 3090 GPU and a 3.80 GHz Intel i7-10700K CPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

Justification: All data, codes, and methodologies involved in this paper comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes] .

Justification: We discuss the potential impacts of this work in the supplementary material. Specifically, this paper is devoted to solving the problem of multi-modal image fusion under degraded scenes to provide high-quality fused results suitable for human and machine perception. Therefore, it can be expected that this work will demonstrate positive social impacts in many fields. For example, it can help drivers better perceive the road conditions ahead in environments with poor visibility through information fusion, such as at night, to improve driving safety. For another example, it can help poor areas that only have low-quality medical imaging equipment to enhance the perception of the body's condition through information recovery and fusion, thereby assisting in disease diagnosis and treatment. As far as we know, this work does not appear to have any negative social impacts and the risks are extremely low.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes] .

    Justification: All data covered in this paper are publicly available, and we provide accurate citations for them.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes] .

    Justification: The code for our work is provided as a zip file, which already contains an MIT License.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.