AnyMEF: Unified Multi-Exposure Fusion with a Flexible Generative Model

Anonymous authors
Paper under double-blind review

Abstract

Multi-exposure fusion (MEF) combines images captured at different exposure levels for HDR imaging. However, existing MEF methods are typically tailored to specific exposure protocols, limiting their applicability in general scenarios. We formulate MEF as a unified observation-guided enhancement task and present AnyMEF, a generative framework that flexibly supports diverse MEF protocols. Built upon a pretrained generative model with multi-reference capabilities, AnyMEF introduces an extracted-attention-guided fusion mechanism (EAGF) to inject structural information from an arbitrary number of input images. Its synthetic training-data pipeline removes the assumption that the anchor must have a particular exposure, allowing any input frame to be selected as the anchor. In addition, AnyMEF allows adaptively routing image subareas to the transformer backbone, and implements a quality-based token selection strategy to improve efficiency. Experiments on multiple MEF benchmarks show that AnyMEF significantly outperforms existing methods while requiring only 20% as many TFLOPs and 6% as much inference time as UltraFusion.

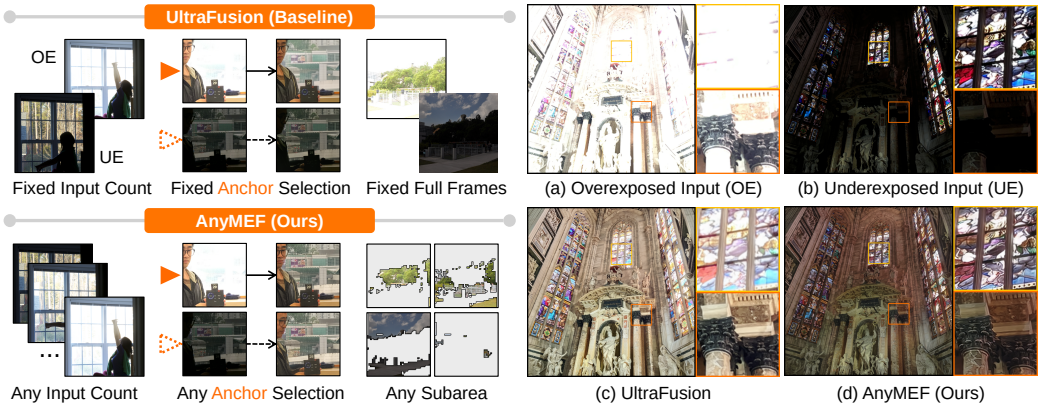

Figure 1: Left: Compared to existing methods that are trained for a specific MEF protocol, such as one that assumes the input must be 2 full images with one overexposed (OE) anchor and one underexposed (UE) reference, AnyMEF is a unified model that can take any number of frames as input, set any frame as the anchor without performance degradation, and adaptively route any subarea to the backbone. Right: Qualitative results on UltraFusionBench. AnyMEF restores more details compared to UltraFusion (Chen et al., 2025).

1 Introduction

Our world, with dazzling highlights and deep shadows, has a high dynamic range that often exceeds the sensing range of a single camera exposure, causing local regions of an image to be overexposed or underexposed. Facing this challenge, a common solution is Multi-Exposure Fusion (MEF): multiple images are sequentially captured with varying shutter

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.

Table 1: Various input and anchor protocols of existing methods.

Method	Input images	Anchor selection
RealHDRV (Shu et al., 2024)	3: Underexposed + Overexposed + Underexposed	OE frame in the middle
HSDS-MEF (Wu et al., 2025)	2: Overexposed + Underexposed	Misalignment not considered
UltraFusion (Chen et al., 2025)	2: Overexposed + Underexposed	Select overexposed
FreeMEF (Qu et al., 2026)	Any number of frames	Select middle-exposed
AnyMEF (Ours)	Any number of frames	Any frame

speeds, aperture sizes, or gain settings, and fused into an enhanced output image (Yue et al., 2026; Wu et al., 2025; Zhu et al., 2024; Qu et al., 2026; Jiang et al., 2023).

Though similar in spirit, many variants exist for the MEF task. We define the *protocol* of an MEF task as the combination of the following information: the number of input images, the exposure sequence used to capture them, and the choice of the anchor image, i.e., the image to which the output is required to be spatially and temporally aligned.

Existing methods summarized in Table 1 are typically trained for a specific protocol, and deviations from it lead to performance degradation. In practice, however, it is desirable for a single model to be able to incorporate additional complementary images when available and flexibly select any input frame as the anchor—for example, the frame in which the target is best focused. This raises a natural question:

Can we design a unified MEF model that is both powerful and compatible with any protocol?

In this work, we observe that strong aesthetic priors and multi-reference capabilities of generative models such as FLUX.2 (Black Forest Labs, 2025) hold promise for this purpose. We augment it with an extracted-attention-guided fusion mechanism (EAGF) to inject structural details from any number of input images, and design a synthetic training pipeline based on an observation-guided enhancement formulation, allowing any frame to be selected as the anchor. Moreover, unlike existing methods that uniformly process full-frame inputs, our method can adaptively route either full frames or selected informative token regions of flexible subareas to the Transformer backbone, allocating computation where it matters most. Experiments show that the resulting AnyMEF model significantly outperforms existing methods on MEF benchmarks while requiring only 20% as many TFLOPs and 6% as much GPU inference time as the existing method UltraFusion.

Our contributions can be summarized as follows:

- We formulate multi-exposure fusion as an observation-guided enhancement task, unifying it across MEF protocols with any input count, any anchor frame, and any input subarea;
- We design an extracted-attention-guided fusion mechanism (EAGF) and a quality-based token selection framework to adapt generative models to this unified task, making their strong generative priors applicable to multi-exposure fusion; and
- We propose AnyMEF, a flexible model for the unified task that achieves state-of-the-art performance on multi-exposure fusion benchmarks.

2 Related Works

HDR Imaging. HDR information exceeding a single image's dynamic range can be restored with many methods including fusing multiple frames (Niu et al., 2021; Kong et al., 2024; Li et al., 2025; Yue et al., 2026), using information from additional sensors such as event cameras (Yang et al., 2023; Chang et al., 2023), or even hallucinating with generative priors (Liu et al., 2026). HDR imaging algorithms focus on restoring real illuminance relationships, usually with HDR formats as output.

Multi-Exposure Fusion. MEF algorithms, which output tonemapped images of HDR scenes, have long been used for image quality enhancement (Mertens et al., 2007). Facing the need for robustness in the wild, where ghosting artefacts can be caused by motion and misalignment (Fang et al., 2020; Liu et al., 2022), a large variety of learning-based methods from ultrafast lookup tables (Jiang et al., 2023), task-customized mixture-of-adapter mod-

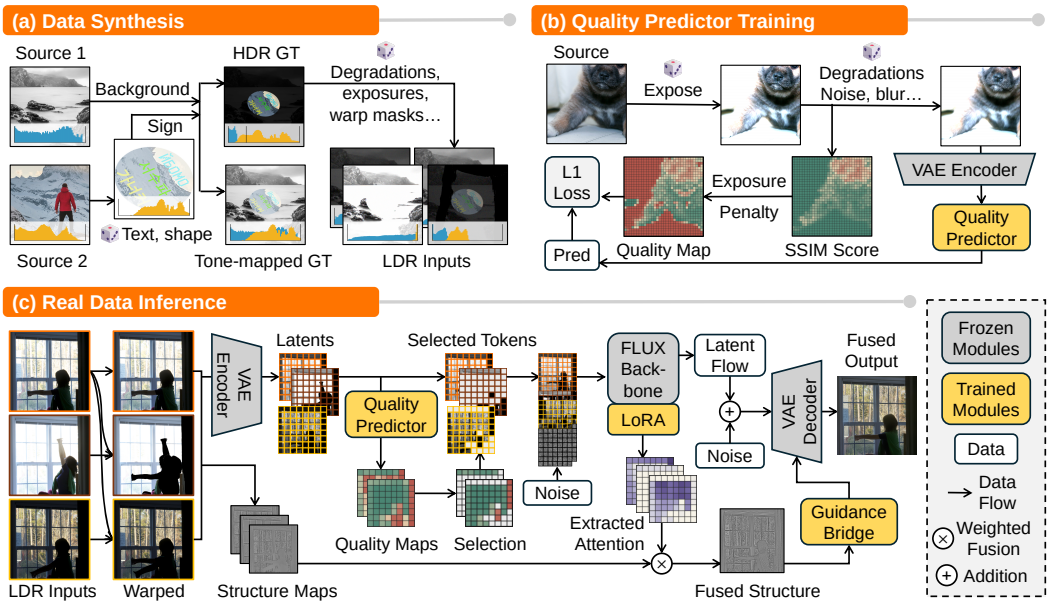

Figure 2: (a) Our data synthesis pipeline models non-monotonic tone mapping with synthetic illuminated signs. (b) We train a lightweight quality predictor for token selection, also trained with synthetic degraded images. (c) The LoRA-finetuned backbone, guidance bridge and quality predictor of AnyMEF are trained in three stages sequentially. At inference time, they work together: the quality predictor selects the tokens that are routed into the backbone; attention maps are extracted from the backbone to guide the fusion of structure maps; and the fused structure map is injected into the VAE decoder through the guidance bridge to help decode the output latent of the backbone.

els (Zhu et al., 2024) to heavy generative models (Chen et al., 2025) have been designed. A limitation of these methods is that they are often specialized to a specific MEF protocol. FreeMEF (Qu et al., 2026) allows using any number of images for input, but it still limits the anchor frame to be a middle-exposed frame.

Generative Priors for Image Enhancement. Generative models have primarily been developed for high-level tasks such as text-to-image generation, but their strong natural-image priors have increasingly been found useful for low-level image restoration and enhancement. This line of research has progressed from diffusion-based approaches (Lin et al., 2024; Xu et al., 2024; Kim et al., 2025; Zhang et al., 2026) to flow-matching-based generative models (Deng et al., 2025; Martin et al., 2025; Fei et al., 2026). The main challenges in this direction include the fundamental trade-off between perceptual enhancement and fidelity to the original image, as well as the high computational cost of using large base models.

3 Method

3.1 Problem formulation and data synthesis

Early multi-exposure fusion (MEF) methods formulate the problem as a weighted fusion task, in which spatially varying weight maps are estimated to combine differently exposed images (Mertens et al., 2007). In practice, however, these images are captured sequentially, and the resulting time gap can cause object and camera motion, leading to artifacts such as ghosting. UltraFusion (Chen et al., 2025) instead models MEF as a guided inpainting problem, where overexposed regions in an OE image are inpainted with soft guidance from a UE reference. However, this formulation is limited to an OE–UE input pair with the OE image fixed as the anchor.

In this work, we formulate MEF as a unified *observation-guided enhancement task*. Given N multi-exposure images captured at times $t_1, \ldots, t_N$, we assume corresponding ideal HDR images $H_1, \ldots, H_N$. The captured images $O_1, \ldots, O_N$ are degraded observations with unknown degradations d_i:

$$O_i = \mathrm{Deg}(H_i, d_i). \tag{1}$$

These degradations include exposure errors, motion blur, sensor noise, demosaicing artifacts, defocus blur, and imperfect white balance, with motion blur and sensor noise more common in OE and UE images, respectively.

When O_k is selected as the anchor, the goal is not simply to reconstruct H_k. Instead, the model should generate an enhanced image E_k corresponding to the content and viewpoint of H_k, with sufficient detail and perceptual quality for LDR display. This may require spatially varying, non-monotonic tone mapping. For example, an illuminated sign in a night scene may be much brighter than its surroundings. Directly preserving this brightness ratio can make the sign unreadable or render the rest of the scene too dark. A better rendering applies separate tone mapping to the sign and its surroundings, allowing both regions to retain sufficient detail, much like the human visual system adapts to different regions when viewing a high-dynamic-range scene. We hence define the MEF task as:

$$E_k = \mathrm{MEF}\left(\{O_i\}_{i=1}^{N}, k\right). \tag{2}$$

The observations need not include both overexposed and underexposed images: they may be all overexposed, all underexposed, or affected by different degradations. An observation O_i need not be a complete image; a spatially local observation, such as a subarea of image patches, can still provide useful guidance. Because brightness and tone mapping are ambiguous, multiple LDR images may be consistent with the content of H_k; MEF should produce the most aesthetically pleasing one.

Based on this formulation, we construct a data synthesis pipeline from high-quality image datasets, as illustrated in Fig. 2(a). We treat each source image, assumed to be properly exposed and color-corrected, as the ground-truth enhanced image E. We apply inverse gamma correction and related operations to transform it into a pseudo-HDR representation. We then apply random exposure changes, noise, and blur to this representation to generate input observations $\{O_i\}$.

Since the above degradations preserve a monotonic mapping from H to E, they cannot model locally non-monotonic tone mapping. We therefore introduce a *synthetic illuminated sign* mechanism. Specifically, we randomly select two source images, use one as the scene background, and turn a randomly cropped region from the other into a text-bearing sign. We compose the pseudo-HDR input and enhanced target differently: the sign is made brighter than the background in the pseudo-HDR input, while the target preserves their original brightness ratio. Ablation results in Fig. 5 show that this augmentation improves enhancement in real night scenes with illuminated signs.

Following UltraFusion, we also add random black regions corresponding to areas masked by the warp-consistency check, improving robustness to such masks during inference.

3.2 One-step generative backbone

To support an arbitrary number of input images, we adopt a transformer-based model that naturally handles variable-length inputs. Specifically, we use FLUX.2-klein-4B (Black Forest Labs, 2025) as our base model. It provides strong image priors and natively supports image generation conditioned on multiple references, making it well suited to our task formulation.

The base model consists of a VAE encoder Enc, a transformer backbone $\mathcal{B}_\theta$, and a VAE decoder Dec. As shown in Fig. 2(c), given the observations of a MEF problem, reference images are first warped to the anchor frame using a pretrained optical flow model RAFT (Teed & Deng, 2020), and possibly occluded pixels failing the forward-backward warp consistency

check are masked black. The images are then encoded into latents $\{\mathrm{Enc}(O_i)\}_{i=1}^{N}$ and split into tokens, each corresponding to a 16×16 patch in the original image. Rotary Position Encoding (Su et al., 2024) encodes both the spatial position of each token and the image to which it belongs. We encode the anchor frame as the first frame, with other observations following. For non-anchor frames, we only keep a selected subset of the latent tokens at inference time; to adapt to any selection algorithm, at training time we perform random token dropout.

We denote z_{init} as a randomly initialized latent, $z_{\mathrm{gt}} = \mathrm{Enc}(E_k)$ as the latent of the ground truth enhanced image that we would like to approach, and $z_t = tz_{\mathrm{init}} + (1-t)z_{\mathrm{gt}}$ as an intermediate point along the flow path. As a flow-predicting model, the backbone is trained to predict the velocity from any intermediate point towards the ground truth:

$$\mathcal{B}_\theta(t, \mathrm{Enc}(O_k), \{\mathrm{Enc}(O_i)\}_{i\neq k}, z_t) = v_t = z_{\mathrm{init}} - z_{\mathrm{gt}}. \tag{3}$$

At inference time, we use a single Euler step for efficient generation:

$$\hat{z}_{\mathrm{gt}} = z_{\mathrm{init}} - \mathcal{B}_{\theta^*}(1, \mathrm{Enc}(O_k), \{\mathrm{Enc}(O_i)\}_{i\neq k}, z_{\mathrm{init}}), \qquad \hat{E}_k = \mathrm{Dec}(\hat{z}_{\mathrm{gt}}), \tag{4}$$

Here, $\hat{E}_k$ and $\hat{z}_{\mathrm{gt}}$ denote the predicted enhanced image and its latent representation, respectively. Through this end-to-end training process, the model learns to align the generated content with the anchor-defined viewpoint and produce an image whose exposure and color distribution resemble those of the high-quality enhanced image E_k.

We finetune the model backbone using LoRA (Low-Rank Adaptation, Hu et al. (2022)) with a latent-space MSE loss, while keeping the VAE encoder and decoder frozen. This preserves the pretrained generative prior and general functionality while adapting the model to exposure fusion.

3.3 Extracted-attention-guided refiner

A key challenge of latent-flow-based generative models for image enhancement is the information bottleneck of the VAE. An RGB patch with $3 \times 16 \times 16$ values is encoded into a 128-dimensional visual token, reducing the representation dimensionality by 83.3%. This bottleneck can discard fine details, motivating us to inject structural information back into the decoding process. However, the N input images contain both useful details and degraded content: selection and fusion require image-quality awareness comparable to that needed for MEF itself.

We observe that the finetuned transformer backbone implicitly provides such quality awareness through its internal attention patterns. Specifically, in each transformer block ℓ, the anchor tokens, reference tokens, and current latent tokens iterate after attending to each other. After the final block, the tokens corresponding to the current latent are projected to predict the output flow. As visualized in Fig. 5(k), we find that the attention between output positions and other tokens correlates with the quality of the corresponding content, especially in later blocks.

Let $a_{p,i}^{(\ell)}$ denote the attention from output position p to the corresponding token in reference image i at block ℓ, averaged over attention heads. We aggregate the attention from the set of selected late blocks $\mathcal{L}$, normalize it across references, and use them to fuse each area of the structure maps $\{S_i\}_{i=1}^{N}$:

$$a_{p,i} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} a_{p,i}^{(\ell)}, \qquad w_{p,i} = \frac{a_{p,i}}{\sum_{j=1}^{N} a_{p,j}}, \qquad S_{\mathrm{fused}}(p) = \sum_{i=1}^{N} w_{p,i} S_i(p). \tag{5}$$

The fused structure map is fed into a lightweight guidance bridge, which extracts structure features at three spatial scales and injects them into the corresponding decoder levels through zero-initialized residual adapters.

During training, the Flux backbone and VAE remain frozen. Since the generative backbone may produce an aesthetically preferred exposure and color distribution that differs from the ground truth, we first apply a robust per-channel mean–standard-deviation matching operation to the ground truth. Denoting the resulting target by $\tilde{E}_k$ and the refined output by $\hat{E}_k^{\text{ref}}$, we optimize the loss $\mathcal{L}_{\text{ref}} = \|\hat{E}_k^{\text{ref}} - \tilde{E}_k\|_1$. This encourages the refiner to restore structural details without altering the exposure and color distribution determined by the generative backbone.

3.4 Quality-based token selector

As the number of reference images increases, two issues emerge in the attention backbone. First, the computational cost grows with the total number of tokens, making inference increasingly expensive. Second, excessive reference tokens may dilute the attention assigned to informative observations and degrade performance. We therefore select reference tokens according to their predicted quality before feeding them into the backbone. Specifically, we retain all tokens from the anchor frame. At each spatial position, among the corresponding tokens from the remaining $N-1$ reference frames, we retain the k tokens with the highest predicted quality and discard the rest. We default to $k = 1$. Note that when $N = 2$ and $k = 1$, this strategy reduces to using both input images in full.

We first define "Quality" q. We start from a high-quality image I, apply a random exposure transformation to obtain an LDR image I_E, and then introduce local degradations, such as noise and blur, to obtain $I_{E,D}$. We use the 16×16-averaged SSIM error between $I_{E,D}$ and I_E, denoted e^{SSIM}, to measure local degradation severity. To account for extremely dark or bright regions, we further introduce an illuminance penalty e^{illum} based on the average luminance ℓ_p of each region. The target quality at position p, q_p, is thus defined as:

$$q_p = 1 - \max\left(e_p^{\text{SSIM}}, e_p^{\text{illum}}\right), \qquad e_p^{\text{illum}} = \max\left(\frac{0.1 - \ell_p}{0.1}, \frac{\ell_p - 0.9}{0.1}, 0\right). \tag{6}$$

We then train a model to predict the per-token quality map q given a degraded image. We encode the degraded image with the VAE encoder into a 128-channel latent, and use a lightweight fully convolutional residual network to predict one quality value for each latent token. In the AnyMEF inference pipeline, since these VAE latents are already computed for the generative backbone, the predictor reuses them without additional computation cost. The predicted quality maps are used for the token selection process.

4 Experiments

4.1 Training details

We train our models with data synthesized from a subset of the Unsplash-Lite (Unsplash, 2026) dataset, which consists of 10,662 high-quality images.

Stage 1: Backbone training. We initialize the backbone from the pretrained FLUX.2-klein-4B model and fine-tune it with low-rank adaptation (LoRA) of rank 16. The model is trained for 30,000 optimization steps on 512×512 crops. We use AdamW with a constant learning rate 1×10^{-5} and weight decay 1×10^{-4}.

Stage 2: Refiner training. We freeze the pretrained FLUX backbone and its autoencoder, and extract attention maps from the last 2 backbone transformer blocks for structure map fusion. The refiner, which extracts structure features and injects them into the decoder, is trained for 30,000 steps with batch size 1 using AdamW (learning rate 2×10^{-4}, weight decay 1×10^{-4}) and a cosine schedule with minimum learning rate 1×10^{-6}.

Stage 3: Quality predictor training. The predictor is trained for 100,000 steps with batch size 32 using AdamW (learning rate 2×10^{-4}, weight decay 1×10^{-4}) and cosine annealing to 1×10^{-6}, with an ℓ_1 loss.

Table 2: Quantitative results on DeepHDRVideo, RealHDRV and DeghostingIQA. Higher is better for all metrics and best results are highlighted in **green**.

Method/Protocol		DeepHDRVideo (5 images)			RealHDRV (8 images)			DeghostingIQA (3–10)	
		$TMQI_S$	MUSIQ	HyperIQA	$TMQI_S$	MUSIQ	HyperIQA	MUSIQ	HyperIQA
MEFLUT (Jiang et al., 2023)	2-OE	0.722	57.110	0.453	0.803	63.860	0.534	68.340	0.580
HSDS-MEF (Wu et al., 2025)	2-OE	0.696	58.370	0.468	0.805	63.930	0.525	68.950	0.575
TC-MoA (Zhu et al., 2024)	2-OE	0.714	56.430	0.473	0.823	63.680	0.537	69.000	0.575
FreeMEF (Qu et al., 2026)	2-OE	0.711	54.540	0.414	0.830	64.670	0.526	66.450	0.489
UltraFusion (Chen et al., 2025)	2-OE	0.719	57.920	0.467	0.838	66.650	0.573	69.980	0.590
AnyMEF (Ours)	2-OE	**0.747**	73.370	0.661	**0.842**	74.670	0.697	75.120	0.685
FreeMEF (Qu et al., 2026)	N-OE	0.706	56.910	0.440	0.832	66.240	0.538	65.820	0.502
AnyMEF (Ours)	N-OE	0.745	**73.860**	**0.668**	0.835	**75.080**	**0.705**	**75.350**	**0.689**

4.2 Evaluation protocols

Datasets. To comprehensively evaluate the performance of methods, we conduct experiments on 5 MEF datasets: MEFB (Zhang, 2021) and UltraFusionBench (Chen et al., 2025), with 2 images per sample; DeepHDRVideo (Chen et al., 2021), with 5; RealHDRV (Shu et al., 2024), with 8; and DeghostingIQA (Fang et al., 2020), with 3–10 images per sample. DeepHDRVideo and RealHDRV provide ground truth HDR images, while the others do not.

Protocols. To ensure fair comparison between methods, we design 3 evaluation protocols:

- 2-OE: Models are provided with 2 images: one overexposed and one underexposed. The OE image is set to be the anchor frame. It is used to fairly compare against existing models designed for OE+UE protocols.
- 2-UE: Each sample provides the same (OE, UE) pair as in the 2-OE protocol, but the anchor is set to be the UE frame.
- N-OE: Models are provided with as many images as available — 5, 8 and 3–10 for DeepHDRVideo, RealHDRV and DeghostingIQA respectively. The anchor frame is set to be the same OE image as in the 2-OE protocol.

Metrics. We use the no-reference metrics MUSIQ (Ke et al., 2021) and HyperIQA (Su et al., 2020) to assess the aesthetic quality of results. For datasets with ground truth images, we additionally report $TMQI_S$, the structural fidelity component of TMQI (Yeganeh & Wang, 2013). We use $TMQI_S$ rather than the full TMQI, which also includes a no-reference statistical naturalness term, to reduce redundancy with the no-reference metrics.

More details on the evaluation protocols can be found in Appendix A.2.

4.3 Results

2-OE Comparison. We first perform a fair comparison among all methods on the 2-OE protocol, which is the protocol most existing methods were trained to be specialized for. Quantitative results can be seen in the 2-OE rows in Table 2 and Table 3, showing that the performance of AnyMEF significantly exceeds other methods by a large margin. Qualitative results in Fig. 4 show that AnyMEF restores more vivid colors with less artefacts.

Multi-input comparison. We further test AnyMEF and FreeMEF on N-OE, the protocol which gives models the chance to utilize additional input images, such as in the example of Fig. 3 where details are missing in both 2-OE images. Results in Table 2 show that extending the input of AnyMEF from 2-OE to N-OE slightly improves performance. Note that since AnyMEF only routes the top $k = 1$ reference token at each position, the computation cost of AnyMEF on the two protocols is approximately the same.

Robustness to anchor selection. FreeMEF and UltraFusion use anchor-specific processing, so we swap their OE and UE inputs to evaluate the 2-UE protocol, where the UE image serves as the anchor. As shown in Table 3, both methods suffer a substantial

Table 3: Quantitative results on MEFB and UltraFusionBench. Higher is better for all metrics and best results are highlighted in **green**.

Method/Protocol		MEFB (2 images)		UltraFusionBench (2 images)	
		MUSIQ	HyperIQA	MUSIQ	HyperIQA
MEFLUT (Jiang et al., 2023)	2-OE	66.790	0.622	67.770	0.609
HSDS-MEF (Wu et al., 2025)	2-OE	66.850	0.603	67.500	0.588
TC-MoA (Zhu et al., 2024)	2-OE	63.870	0.548	68.770	0.604
FreeMEF (Qu et al., 2026)	2-OE	63.280	0.521	66.860	0.537
UltraFusion (Chen et al., 2025)	2-OE	68.890	0.649	70.560	0.637
Ours	2-OE	72.550	**0.709**	**73.770**	**0.683**
FreeMEF (Qu et al., 2026)	2-UE	60.190 (−3.100)	0.514 (−0.007)	53.840 (−13.020)	0.479 (−0.059)
UltraFusion (Chen et al., 2025)	2-UE	61.120 (−7.770)	0.575 (−0.073)	53.020 (−17.540)	0.503 (−0.134)
Ours	2-UE	**72.660** (+0.100)	**0.709** (+0.000)	73.530 (−0.240)	0.671 (−0.012)

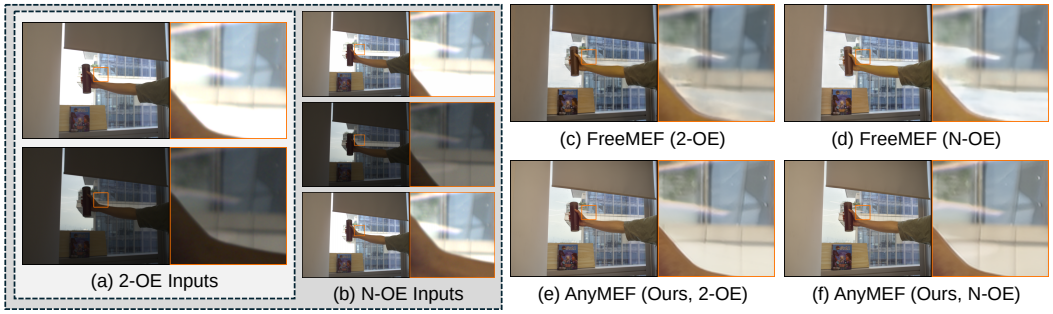

Figure 3: Qualitative results on DeepHDRVideo. In the 2-OE inputs, the vertical wall seam in the background is overexposed in the OE image and occluded in the UE image. The N-OE protocol provides images with more information and allows AnyMEF to restore the seam accordingly.

performance drop due to this protocol mismatch, whereas AnyMEF remains stable when switching the anchor from OE to UE.

Computation cost. Table 4 lists inference computation costs on the RealHDRV 2-OE protocol, profiled on a single NVIDIA H20-3e GPU. As a generative model, AnyMEF costs much more computation than FreeMEF, HSDS-MEF and MEFLUT. However, AnyMEF requires less than 20% as many TFLOPs and 6% as much inference time as UltraFusion, which is also based on generative priors. A breakdown of the cost of AnyMEF components can be seen in Appendix A.3.

4.4 Ablation studies

Refiner. Without the refiner, the information bottleneck caused by the VAE encoder and decoder leads to distorted details in outputs, which are especially perceptible in areas with small text. An example is shown in Fig. 5(d): the text "LOS MEUR" is distorted in the directly decoded output of the backbone, and fixed by the refiner in Fig. 5(c).

Synthetic signs. We retrained AnyMEF with a data pipeline without the synthetic illuminated sign augmentation. In its outputs, night signs become blurry, such as in Fig. 5(e).

Extracted attention. We substitute the extracted-attention guided fusion (EAGF) mechanism with directly using average weights to fuse the structure maps. As compared in Fig. 5(i) and (j), EAGF provides stronger structure signals for guidance.

Top-k $k = 1$. In our current implementation, only $k = 1$ reference tokens are selected for each position. Experiments show that using a larger k value costs more computation, while not improving performance. See Appendix A.4 for results.

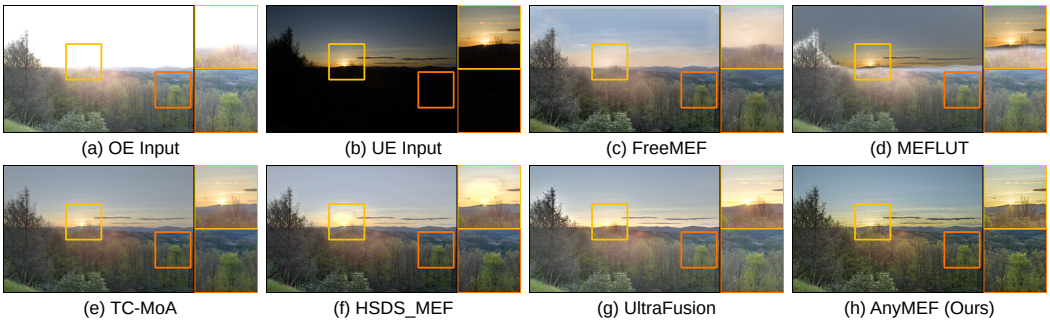

Figure 4: Qualitative results on the MEFB benchmark.

Table 4: Computation cost profiled on RealHDRV. Entries are reported as method-specific cost + RAFT alignment cost. FreeMEF internally deals with misalignment and does not need explicit alignment.

Method	Params (M)	TFLOPs	Latency (ms)
FreeMEF	8.92	1.27	252.22
MEFLUT	$0.0005 + 5.26$	$0.00003 + 5.13$	$1.52 + 344.34$
UltraFusion	$1855.15 + 5.26$	$562.01 + 5.13$	$36988.32 + 324.32$

Method	Params (M)	TFLOPs	Latency (ms)
HSDS-MEF	$1.17 + 5.26$	$1.79 + 5.13$	$108.27 + 352.67$
TC-MoA	$340.89 + 5.26$	$39.02 + 5.13$	$2110.06 + 338.11$
AnyMEF	$3977.73 + 5.26$	$107.42 + 5.13$	$1846.86 + 339.90$

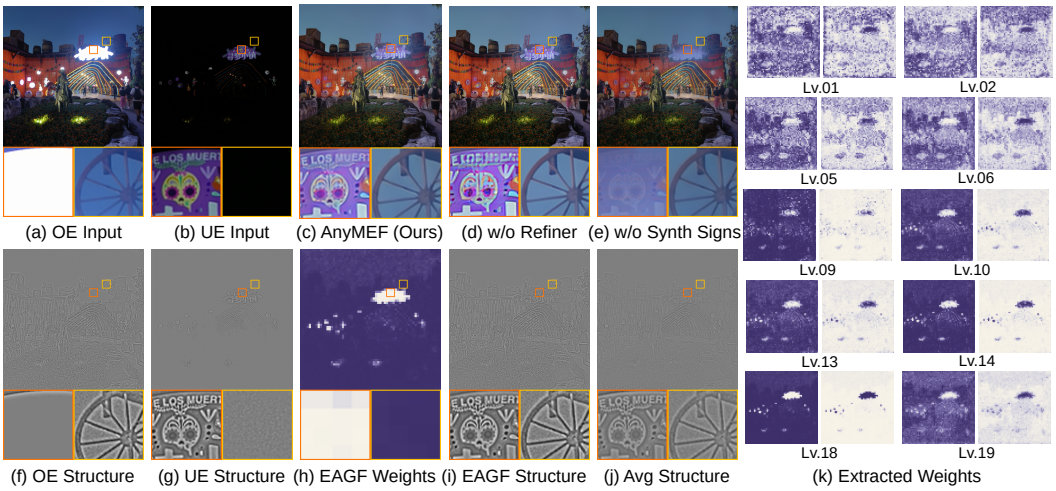

Figure 5: Qualitative ablation results and attention visualizations from UltraFusionBench.

Table 5: Ablation results. Higher is better for all metrics, with best highlighted in **green**.

Method/Protocol		DeepHDRVideo (5 images)			RealHDRV (8 images)		
		$TMQI_S$	MUSIQ	HyperIQA	$TMQI_S$	MUSIQ	HyperIQA
AnyMEF	N-OE	**0.745**	73.860	**0.668**	**0.835**	**75.080**	**0.705**
w/o Refiner	N-OE	0.723	**74.260**	0.657	0.817	74.830	0.692
w/o Synthetic Signs	N-OE	0.739	70.840	0.640	0.823	73.940	0.689
w/o Extracted Attention	N-OE	0.740	73.300	0.665	0.831	75.020	0.698

5 Conclusion

In this work, we formulate Multi-Exposure Fusion as an observation-guided enhancement task. We propose AnyMEF, a unified model that flexibly handles varying MEF protocols while exploiting generative priors. Experiments show that AnyMEF significantly outperforms existing methods. Limitations including high computation cost, lack of anchor-deblurring ability, potential color biases and experiments limited to FLUX are discussed in Appendix A.1.

AI use statement

In this work, we used generative AI tools to implement methods, provide feedback on research methodology and experiments, assist with translation and generate synthetic datasets (the pool of random text strings for the synthetic signs is AI-generated). We have not used generative AI tools to design research methodology or experiments, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to summarize and analyse existing literature, build research infrastructure such as experiment dashboards, format references, and help with LaTeX formatting. We have reviewed all AI-assisted work. We designed the code architecture and verified that each part of the LLM-generated code aligns with the purpose, and we checked LLM-summarized literature for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We trained our models on open source data. We will release our pretrained checkpoints as well as complete code including data processing scripts, training code and testing code, ensuring that all results can be reproduced.

References

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 FPS HDR video with a spike-RGB hybrid camera. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K. Wong, and Lei Zhang. HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. UltraFusion: Ultra high dynamic imaging using exposure fusion. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Junyuan Deng, Xinyi Wu, Yongxing Yang, Congchao Zhu, Song Wang, and Zhenyao Wu. Acquire and then adapt: Squeezing out text-to-image model for image restoration. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuming Fang, Hanwei Zhu, Kede Ma, Zhou Wang, and Shutao Li. Perceptual evaluation for multi-exposure image fusion of dynamic scenes. *IEEE Transactions on Image Processing (TIP)*, 29:1127–1138, 2020.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. LucidFlux: Caption-free photo-realistic image restoration via a large-scale diffusion transformer. In *Proc. of International Conference on Learning Representations (ICLR)*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. of International Conference on Learning Representations (ICLR)*, 2022.

Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, and Shuaicheng Liu. MEFLUT: Unsupervised 1D lookup tables for multi-exposure image fusion. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Exploiting diffusion prior for task-driven image restoration. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. SAFNet: Selective alignment fusion network for efficient HDR imaging. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024.

Xinyue Li, Zhangkai Ni, and Wenhan Yang. AFUNet: Cross-iterative alignment-fusion synergy for HDR reconstruction via deep unfolding paradigm. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024.

Aoyu Liu, Zhen Liu, Ziyi Wang, Dian Chen, Bing Zeng, and Shuaicheng Liu. ExpoCM: Exposure-aware one-step generative single-image HDR reconstruction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proc. of European Conference on Computer Vision (ECCV)*, 2022.

Ségolène Tiffany Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-play image restoration with flow matching. In *Proc. of International Conference on Learning Representations (ICLR)*, 2025.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Proceedings of the 15th Pacific Conference on Computer Graphics and Applications*, 2007.

Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson W. H. Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing (TIP)*, 30:3885–3896, 2021.

Lishen Qu, Yao Liu, Shihao Zhou, Jie Liang, Hui Zeng, Lei Zhang, and Jufeng Yang. There and back again: A flexible-frame transformer for multi-exposure fusion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2026.

Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world HDR video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jianlin Su, Ahmed Murtadha, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zachary Teed and Jia Deng. RAFT: Recurrent all pairs field transforms for optical flow. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020.

Unsplash. Unsplash Lite Dataset 1.4.0. Dataset, 2026. URL `https://unsplash.com/data`. Accessed September 24, 2026.

Guanyao Wu, Hongming Fu, Jinyuan Liu, Long Ma, Xin Fan, and Risheng Liu. Hybrid-supervised dual-search: Leveraging automatic learning for loss-free multi-exposure image fusion. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 2025.

Xiaogang Xu, Shu Kong, Tao Hu, Zhe Liu, and Hujun Bao. Boosting image restoration via priors from pre-trained models. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hojatollah Yeganeh and Zhou Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing (TIP)*, 22(2):657–667, 2013.

Huanjing Yue, Dawei Li, Shaoxiong Tu, and Jingyu Yang. F2HDR: Two-stage HDR video reconstruction via flow adapter and physical motion modeling. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xingchen Zhang. Benchmarking and comparing multi-exposure image fusion algorithms. *Information Fusion*, 74:111–131, 2021.

Yuhong Zhang, Hengsheng Zhang, Xinning Chai, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Diff-Restorer: Unleashing visual prompts for diffusion-based universal image restoration. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 36(4):4259–4274, 2026.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

A Appendix

A.1 Limitations

Computation efficiency. Compared with lightweight CNN-based methods, the heavy computation load of AnyMEF, despite being lower than that of UltraFusion, is a critical issue for practical applications. This is a common challenge faced by current generative-model-based image enhancement methods: in practical applications, they can typically only be used for cloud processing after image capture, rather than for real-time enhancement on mobile devices.

Restoring other degradations. Although we consider degradations such as motion blur in our problem formulation and introduce the corresponding degradation probabilities in the training data, our trained model currently cannot remove motion blur in the anchor frame during HDR fusion. However, if motion blur appears only in individual reference frames, AnyMEF can identify them through the quality predictor and token selection and avoid using them as references. In addition, AnyMEF can suppress image noise relatively effectively.

Color biases. Since we also include color grading within the scope of optimization, the color grading of the final output image may deviate visually to some extent. Future work will try to make it more controllable.

Beyond FLUX. In this work, our experiments are conducted only with FLUX.2-Klein-4B (Black Forest Labs, 2025) as the base model. We choose it because it has a relatively small number of parameters, strong performance, and is open-source. In fact, our adaptation method should also be applicable to various existing or future transformer-based generative models with multi-reference capabilities, which remains to be further explored.

A.2 Evaluation protocol details

We define deterministic rules for each dataset to determine which images are selected as inputs under the 2-OE, 2-UE, and N-OE protocols.

MEFB & UltraFusionBench. Each sample in these two datasets provides one overexposed image and one underexposed image. We use both images as inputs under all three protocols.

RealHDRV. Each RealHDRV sample provides HDR ground truth at eight time instants, together with one overexposed and one underexposed LDR image at each instant. We denote these images by $I_{t,\mathrm{HDR}}$, $I_{t,\mathrm{OE}}$, and $I_{t,\mathrm{UE}}$, where $t \in \{1, 2, \ldots, 8\}$. In the 2-OE protocol, the input set is $\{I_{4,\mathrm{OE}}, I_{5,\mathrm{UE}}\}$, with $I_{4,\mathrm{OE}}$ as the anchor and $I_{4,\mathrm{HDR}}$ as the ground truth. In the N-OE protocol, the input set is $\{I_{1,\mathrm{UE}}, I_{2,\mathrm{OE}}, I_{3,\mathrm{UE}}, I_{4,\mathrm{OE}}, I_{5,\mathrm{UE}}, I_{6,\mathrm{OE}}, I_{7,\mathrm{UE}}, I_{8,\mathrm{OE}}\}$, with $I_{4,\mathrm{OE}}$ as the anchor and $I_{4,\mathrm{HDR}}$ as the ground truth. In the 2-UE protocol, the input set is $\{I_{5,\mathrm{UE}}, I_{4,\mathrm{OE}}\}$, with $I_{5,\mathrm{UE}}$ as the anchor and $I_{5,\mathrm{HDR}}$ as the ground truth.

DeepHDRVideo. Each DeepHDRVideo sample provides five alternating-exposure LDR images and an HDR ground truth corresponding to the center time instant. It contains two types of samples: "high" samples, which provide $\{I_{1,\mathrm{OE}}, I_{2,\mathrm{UE}}, I_{3,\mathrm{OE}}, I_{4,\mathrm{UE}}, I_{5,\mathrm{OE}}, I_{3,\mathrm{HDR}}\}$, and "low" samples, which provide $\{I_{1,\mathrm{UE}}, I_{2,\mathrm{OE}}, I_{3,\mathrm{UE}}, I_{4,\mathrm{OE}}, I_{5,\mathrm{UE}}, I_{3,\mathrm{HDR}}\}$. In the N-OE and 2-OE protocols, we use only the high samples, with $I_{3,\mathrm{OE}}$ as the anchor and $I_{3,\mathrm{HDR}}$ as the ground truth. The input set for N-OE is $\{I_{1,\mathrm{OE}}, I_{2,\mathrm{UE}}, I_{3,\mathrm{OE}}, I_{4,\mathrm{UE}}, I_{5,\mathrm{OE}}\}$, while that for 2-OE is $\{I_{3,\mathrm{OE}}, I_{4,\mathrm{UE}}\}$. The 2-UE protocol uses only the low samples, with input set $\{I_{3,\mathrm{UE}}, I_{4,\mathrm{OE}}\}$, $I_{3,\mathrm{UE}}$ as the anchor, and $I_{3,\mathrm{HDR}}$ as the ground truth.

DeghostingIQA. We rank the valid images in each scene by the provided exposure values when available, and otherwise by their mean luminance. The brightest image is used as the anchor and the darkest image as the reference in 2-OE; the two are swapped in 2-UE. In

the N-OE protocol, the brightest image remains the anchor and all remaining images are used as references.

For testing convenience, we resize all test images and their corresponding ground truth images so that neither dimension exceeds 1024 pixels. Specifically, for images whose longer side is no greater than 1024, we test them at their original resolution; for images whose longer side exceeds 1024 while their shorter side is no greater than 1024, we crop the central 1024 pixels along the longer side; for images whose shorter side also exceeds 1024, we first resize them so that the shorter side has length 1024 and then crop the central 1024 pixels along the longer side.

A.3 Computational cost breakdown

In Table 6, we use the RealHDRV dataset, in which each sample provides eight reference images, as the profiling benchmark, and analyze the computational cost and runtime of each component in AnyMEF. It can be seen that the quality-map scorer and refiner that we additionally introduce both incur relatively low runtime costs. The FLUX backbone has the largest computational cost, reaching 151 TFLOPs when $k = 7$ (equivalent to using all eight images as input). Under our default $k = 1$ configuration, where the quality map is used to retain the top-1 reference token at each position, the backbone cost can be reduced to 74 TFLOPs.

Other major cost items include applying VAE Encoder embedding to each image and using the RAFT model to predict optical flow and warp each reference image to the anchor frame. In the current implementation, we process the images sequentially. Since processing the images in parallel can effectively reduce the total latency, we provide the estimated optimized latency in the "Optimal total" entries of the table. In addition, because the RAFT model is fully decoupled from the other parts of AnyMEF, using a more lightweight optical-flow estimation model can also reduce the cost.

A.4 Ablation on k.

In our quality-selection algorithm, we select the k reference tokens with the highest quality at each position, where k is set to 1 by default. In Table 7, we test different values of k from 0 to 7, where 0 means that the problem degenerates into blind restoration based solely on the anchor frame. It can be seen that $k = 1$ achieves the best results on both TMQI_S and HyperIQA; further increasing k significantly increases the computational cost and latency, but cannot further improve performance. This may be because $k = 1$ already uses the most informative reference tokens, and using more tokens with worse quality dilutes the model's attention.

Table 6: Module-level computation breakdown of AnyMEF on RealHDRV. Optimal total denotes the shorter runtime latency that can be achieved by parallelizing multi-image VAE encoding and RAFT alignment.

Stage	Count (unit)	TFLOPs		Latency (ms)	
		Per unit	Total	Per unit	Total
Protocol: N-OE, $k = 1$					
RAFT Warping	7 (pair)	5.13	35.94	359.1	2514.0
VAE Encoder	8 (image)	4.27	34.17	90.9	727.4
Quality Predictor	8 (image)	0.02	0.16	0.9	7.2
FLUX backbone	-	-	74.05	-	1638.4
VAE Decoder	-	-	19.39	-	183.7
Refiner	-	-	0.24	-	198.5
Total	-	-	**163.95**	-	**5269.2**
Optimal total: 359.1+90.9+0.9+1638.4+183.7+198.5=2471.5 *ms*					
Protocol: N-OE, $k = 7$					
RAFT Warping	7 (pair)	5.13	35.94	383.4	2683.8
VAE Encoder	8 (image)	4.27	34.17	90.6	724.7
Quality Predictor	8 (image)	0.02	0.16	0.9	7.4
FLUX Backbone	-	-	151.13	-	6930.7
VAE Decoder	-	-	19.39	-	183.8
Refiner	-	-	0.24	-	223.8
Total	-	-	**241.03**	-	**10754.1**
Optimal total: 383.4+90.6+0.9+6930.7+183.8+223.8=7813.2 *ms*					
Protocol: 2-OE, $k = 1$					
RAFT Warping	1 (pair)	5.13	5.13	349.6	349.6
VAE Encoder	2 (image)	4.27	8.54	90.7	181.5
FLUX backbone	-	-	79.25	-	1618.2
VAE Decoder	-	-	19.39	-	183.7
Refiner	-	-	0.24	-	55.3
Total	-	-	**112.55**	-	**2388.3**
Optimal total: 349.6+90.7+1618.2+183.7+55.3=2297.5 *ms*					

Table 7: Quality–cost trade-off of the top-k variants on the RealHDRV N-OE protocol. All metrics are higher-better, with best quality metrics highlighted in **green**.

k	TMQI_S	MUSIQ	HyperIQA	TFLOPs	Latency (ms)
0	0.8158	74.41	0.6981	143.28	4435.9
1	**0.8352**	75.08	**0.7047**	163.95	5259.2
2	0.8326	**75.11**	0.7036	182.25	5924.4
3	0.8309	75.06	0.7014	197.77	6630.2
4	0.8315	74.91	0.6978	209.54	7564.3
5	0.8327	74.61	0.6932	220.48	8514.3
6	0.8329	74.32	0.6890	230.90	9416.2
7	0.8329	74.11	0.6856	241.03	10420.2