# Fusion in Your Way: Aligning Image Fusion with Heterogeneous Demands via Direct Preference Optimization

Supplementary Material

## A. Methodology details

This section details the network architecture, hyperparameters, and the preference data annotation process.

### A.1. Network architecture

We use a Transformer-based U-Net architecture as the backbone of our latent diffusion model. To incorporate semantic guidance, the input text prompts $c_t$ are encoded into semantic embeddings with a pre-trained CLIP ViT-L/14 [1, 21] text encoder $\tau_\theta$, denoted as

$$c_{\text{emb}} = \tau_\theta(c_t). \tag{10}$$

These embeddings are subsequently injected into the U-Net via cross-attention. Specifically, the intermediate spatial features $\phi(z_t)$ from the U-Net and the text embeddings $c_{\text{emb}}$ are projected into query $Q$, key $K$, and value $V$ matrices, which is denoted as

$$Q = W_Q \cdot \phi(z_t), \quad K = W_K \cdot c_{\text{emb}}, \quad V = W_V \cdot c_{\text{emb}}, \tag{11}$$

where $W_Q$, $W_K$, and $W_V$ are learnable projection matrices. The cross-attention output then guides the denoising process, *i.e.*,

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{12}$$

where $d$ represents the channel dimension of the keys and queries. This mechanism ensures the fusion output aligns with the specific property. The detailed parameters of PALDM and PCLDM are presented in Table 4.

### A.2. Preference data collection and annotation

**Human feedback.** To capture subjective visual quality, we design a custom user interface (Figure 9) that facilitates the collection of human preferences. Human annotators are presented with candidate fused images generated by PALDM and are asked to select regions according to their perceptual preferences. The segment anything model (SAM) [2, 13] is then applied to accurately generate the corresponding preference-region mask $I_m$ for the selected areas. Specifically, three annotators annotate the feedback datasets for LLVIP, MSRS, and RoadScene, respectively, and we train distinct RLHF models for each dataset.

Table 4. Hyperparameters for the LDMs.

| Configuration | PALDM | PCLDM |
|---|---|---|
| Downsampling | 4 | 4 |
| Latent shape ($z$) | $64 \times 64 \times 9$ | $64 \times 64 \times 9$ |
| Diffusion steps | 1000 | 1000 |
| Noise schedule | linear | linear |
| $N_{\text{params}}$ | 420.26M | 179.51M |
| Channels | 224 | 224 |
| Depth | 2 | 2 |
| Channel multiplier | 1,2,3,4 | 1,2,3,4 |
| Attention resolutions | 8,4,2 | 8,4,2 |
| Head channels | 32 | 32 |
| Batch size | 8 | 8 |
| Epochs | 56 | 20 |
| Learning rate | 5e-6 | 1e-5 |

**VLM-based feedback.** We employ QWEN3-Omni-Think [3, 35] as a VLM-based feedback mechanism to rank all candidates based on overall image quality and fidelity. We construct three VLM-based feedback datasets for LLVIP, MSRS, and RoadScene, respectively, and train separate RLVF models for each. To ensure the VLM focuses strictly on technical fusion quality rather than high-level semantic content or artistic aesthetics, the model is instructed to act as an "Image Quality Rater". The step-by-step instructions are as follows.

**Instruction**

- You will be given five images labeled 1–5. Assess only perceptual image quality.
- Focus solely on *signal-level quality*: sharpness, detail retention, structure fidelity, naturalness, and visible artifacts (noise, halos, ringing, blocking, exposure, color cast).
- Ignore content and aesthetics.
- Provide ranking from best to worst using ">" as separators (e.g., `3 > 1 > 5 > 2 > 4`).

Based on the generated ranking, we designate the highest-ranked image as the preferred sample $I_0^w$ and the lowest-ranked image as the rejected sample $I_0^l$. In this sce-

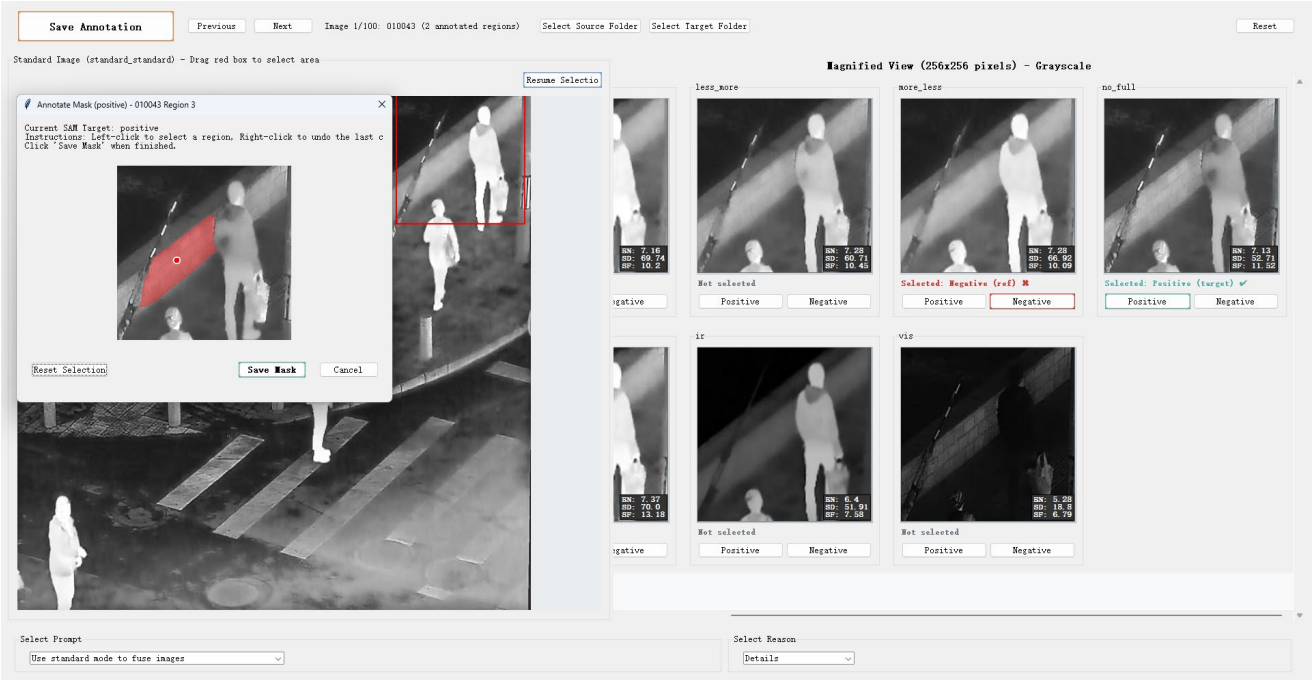


Figure 9. User interface for human feedback annotation. The left side of the interface displays the entire fused image for the user to select a region of interest. The right side shows the fusion results of other candidates within that region, allowing for the annotation of preferred and rejected samples. The pop-up window functions as a SAM-based interface where the user clicks to select the preference region.

nario, the preference is considered global, and the preference mask $I_m$ is set to encompass the entire image.

**Task-driven feedback.** Downstream evaluations are conducted to examine whether the preference outputs produced by PCLDM align with objective performance metrics.

For the semantic segmentation task, we first process the fused candidate images with SegFormer [4, 32] and compare the resulting predictions with the ground-truth labels. The preferred sample $I_0^w$ and the rejected sample $I_0^l$ are determined by evaluating a weighted combination of the mean Intersection over Union (mIoU) and per-class accuracy. The final preference-region mask $I_m$ is defined as the intersection of the predicted masks from the preferred sample, the rejected sample, a baseline reference sample, and the ground-truth mask.

For objective detection, the fused images are sent into a YOLOv11 [5, 12] to obtain detection results. We generate preferences based on detection performance by computing a weighted sum of mean Average Precision (mAP) and accuracy, which determine the preferred $I_0^w$ and rejected $I_0^l$ samples. We select image patches containing detection targets to serve as the preference regions for IDPO, and the preference mask $I_m$ is set to cover these selected patches.

## B. Additional experimental results

This section presents additional qualitative results and introduces five other evaluation metrics to comprehensively assess the performance of DPOFusion.

### B.1. Qualitative comparisons

The extended qualitative comparison results are presented in Figure 10. On the RoadScene dataset, the fusion results of RLVF exhibit richer texture details, with the digital logo on the wall appearing significantly clearer. For the LLVIP dataset, the RLHF fusion results preserve complete traffic markings, aligning closely with human visual perception. Compared to the baseline methods, the fusion results of both RLHF and RLVF meet human perception requirements, offering higher contrast and richer texture details.

### B.2. Quantitative comparisons

Table 5 presents the quantitative comparison results of DPOFusion with other methods on five additional metrics. RLVF and RLHF achieve optimal or comparable values across all three datasets. For LLVIP, RLHF obtains four best values and one second-best value, indicating that the fusion results of RLHF exhibit higher contrast and richer texture details. Meanwhile, RLVF consistently achieves the best or second-best values on the VIFF metric across the three

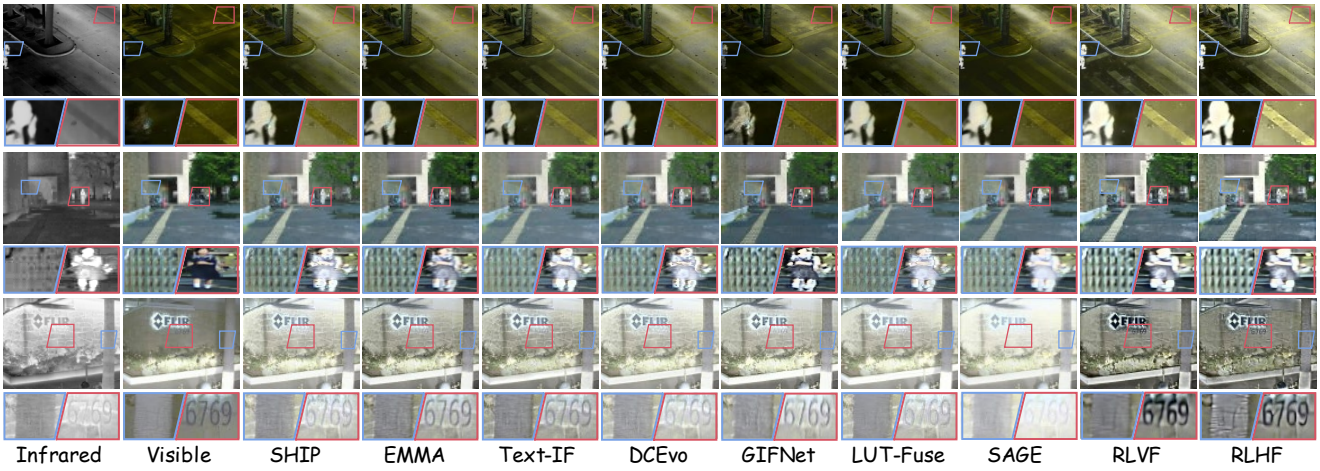


Figure 10. Qualitative comparison of DPOFusion against state-of-the-art fusion methods on the LLVIP, MSRS, and RoadScene datasets.

Table 5. Quantitative comparison of our methods against state-of-the-art fusion methods on the LLVIP, MSRS, and RoadScene datasets. The best values are in **bold**, and the second-best values are underlined.

| Methods | Reference | LLVIP Dataset | | | | | MSRS Dataset | | | | | RoadScene Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | VIF | SCD | SF | DF | VIFF | VIF | SCD | SF | DF | VIFF | VIF | SCD | SF | DF | VIFF |
| U2Fusion | TPAMI[22] | 0.340 | 1.315 | 11.574 | 4.287 | 0.435 | 0.248 | 1.198 | 7.235 | 3.099 | 0.425 | 0.338 | 1.451 | 12.358 | 6.051 | 0.444 |
| DDFM | ICCV[23] | 0.376 | 1.399 | 13.398 | 6.785 | 0.428 | 0.369 | 1.387 | 8.856 | 4.738 | 0.482 | 0.360 | 1.696 | 13.115 | 7.075 | 0.511 |
| SHIP | CVPR[24] | 0.461 | 1.441 | 16.751 | 6.202 | 0.583 | 0.427 | 1.513 | 11.827 | 4.670 | 0.702 | 0.422 | 1.302 | 15.548 | 7.350 | 0.367 |
| EMMA | CVPR[24] | 0.467 | 1.584 | 14.825 | 5.444 | 0.591 | **0.523** | 1.629 | 11.559 | 4.387 | 0.773 | **0.430** | 1.629 | 15.744 | 7.161 | 0.568 |
| Text-IF | CVPR[24] | 0.491 | **1.641** | 15.154 | 5.579 | 0.628 | 0.506 | 1.395 | 10.831 | 4.469 | 0.757 | 0.390 | 1.594 | 16.401 | 7.692 | 0.538 |
| DCEvo | CVPR[25] | 0.484 | 1.546 | 15.882 | 5.341 | 0.599 | 0.514 | **1.665** | 11.460 | 4.326 | 0.756 | 0.410 | 1.476 | 13.860 | 6.246 | 0.442 |
| GIFNet | CVPR[25] | 0.356 | 1.505 | 20.188 | 6.378 | 0.486 | 0.324 | 1.411 | 12.748 | 3.808 | 0.531 | 0.346 | **1.728** | 17.882 | 7.452 | 0.580 |
| LUT-Fuse | ICCV[25] | 0.464 | 1.466 | 14.584 | 4.891 | 0.502 | 0.505 | 1.590 | 11.661 | 4.461 | 0.720 | 0.409 | 1.313 | 12.515 | 5.393 | 0.327 |
| SAGE | CVPR[25] | 0.365 | 1.488 | 13.857 | 4.821 | 0.482 | 0.346 | 1.417 | 10.420 | 3.682 | 0.590 | 0.352 | 1.528 | 10.206 | 4.527 | 0.387 |
| **RLVF** | Ours | 0.453 | 1.460 | 18.319 | 6.307 | 0.803 | 0.459 | 1.456 | **17.512** | **6.494** | **1.134** | 0.361 | 1.475 | **23.165** | **9.601** | **0.625** |
| **RLHF** | Ours | **0.495** | 1.603 | **20.733** | **6.959** | **0.914** | 0.448 | 1.414 | 17.070 | 6.269 | 1.045 | 0.351 | 1.098 | 15.818 | 6.668 | 0.415 |

datasets, demonstrating that RLVF effectively fine-tunes the model to generate higher-quality fused images.

## C. Experimental evaluation and analysis

This section discusses the effectiveness of IDPO on RLHF, RLVF, and RLDF.

### C.1. Fusion results with various text prompts

To verify the controllability of the property-aligned latent diffusion model (PALDM), we visualize the fusion results generated under different property-descriptive text prompts in Figure 11. By leveraging the proposed joint conditional loss and latent space interpolation strategy, PALDM can flexibly adjust the ratio of modal information in the fused image. In this work, we specifically set the interpolation level to $N = 5$. This design choice aims to improve selection efficiency while constraining the solution space distribution of the LDMs. Consequently, it enables the model to more rapidly identify and select appropriate feature mappings within the existing solution space during the subsequent fine-tuning phase. As illustrated in Figure 11, the results exhibit a smooth semantic transition from full infrared information to full visible information. Specifically, images generated with mostly infrared prompts effectively highlight thermal targets, whereas those with mostly visible prompts retain richer background texture details. The diversity results ensure a comprehensive candidate set for the PCLDM fine-tuning.

### C.2. Analysis of IDPO efficacy

The qualitative results of DPOFusion are presented in Figure 12. With IDPO, the framework effectively adjusts the fusion results to align with human visual requirements, as evidenced by the richer texture details on the sidewalks in the LLVIP fusion results. IDPO also post-processes the fusion results to align with large model preferences, ensuring that results on the MSRS and RoadScene datasets retain

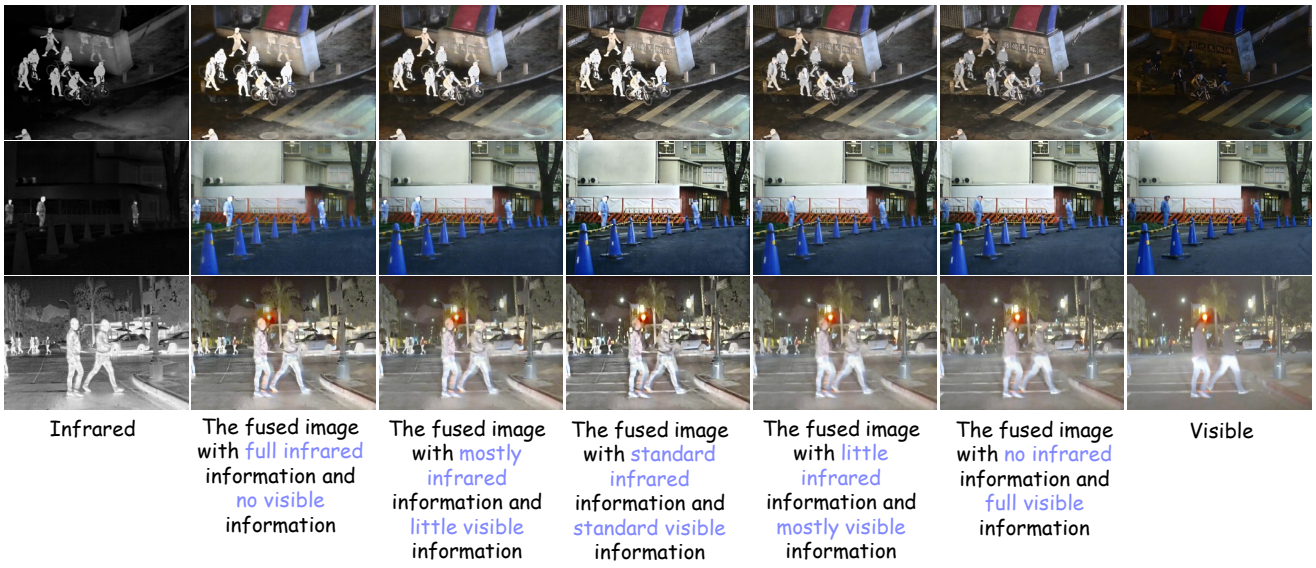


Figure 11. PALDM results conditioned on corresponding text prompts.


Figure 12. Qualitative results of the IDPO ablation study for DPO-Fusion.

complete task details. Finally, IDPO effectively fine-tunes the model to enhance performance in downstream tasks, increasing segmentation accuracy for small targets while improving dense object detection performance.

The quantitative analysis of the ablation study for IDPO on RLHF and RLVF is presented in Table 6. The model

Table 6. Ablation study of the IDPO effectiveness. The best values are shown in **bold**.

| LLVIP dataset | | | | | |
|---|---|---|---|---|---|
| Method | EN | SD | AG | MUS | CNN |
| w/o $\mathcal{L}_{\text{IDPO}}$ | 7.680 | 60.806 | 5.343 | 56.154 | 0.659 |
| RLVF | 7.554 | 53.299 | 5.439 | 56.164 | 0.660 |
| RLHF | **7.725** | **61.911** | **5.977** | **57.280** | **0.660** |
| MSRS dataset | | | | | |
| Method | EN | SD | AG | MUS | CNN |
| w/o $\mathcal{L}_{\text{IDPO}}$ | **7.244** | 55.158 | **5.833** | **39.707** | 0.532 |
| RLVF | 7.203 | **56.614** | 5.782 | 39.684 | 0.524 |
| RLHF | 7.138 | 53.385 | 5.601 | 39.140 | **0.545** |
| RoadScene dataset | | | | | |
| Method | EN | SD | AG | MUS | CNN |
| w/o $\mathcal{L}_{\text{IDPO}}$ | 7.348 | 45.137 | 6.587 | 43.564 | 0.492 |
| RLVF | **7.574** | **53.323** | **8.622** | **43.785** | **0.539** |
| RLHF | 7.210 | 40.535 | 5.957 | 41.431 | 0.472 |

fine-tuned with IDPO achieved improved or comparable results across all three datasets. The increases in EN, SD, and AG indicate that IDPO can effectively enhance the information content of fused images with human or VLM feedback, while MUSIQ and CNNIQA metrics demonstrate perceptual quality improvement.

Table 7 analyzes the effectiveness of IDPO on object detection with different training strategies. The first two lines of the results present the results of training YOLOv11 [5, 12] with raw infrared and visible images, as well as the model fine-tuned with PALDM. The remaining few lines

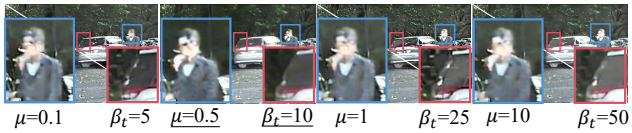


Figure 13. Sensitivity analysis of $\beta_t$ and $\mu$.

Table 7. Quantitative comparison of the IDPO ablation study for object detection.

| Method | Bus | Car | Lamp | Moto. | People | Truck | @.5:.95 | @.5 | @.75 |
|---|---|---|---|---|---|---|---|---|---|
| w/o $\mathcal{L}_{\text{IDPO}}$ | 74.70 | 62.20 | 33.40 | 34.20 | 38.70 | 56.50 | 49.92 | 76.97 | 52.82 |
| RLDF-OD | **76.30** | **63.30** | **36.30** | **34.80** | **41.10** | **57.70** | **51.58** | **78.81** | **53.79** |
| w/o $\mathcal{L}_{\text{IDPO}}$ | 63.69 | 58.94 | 22.46 | 25.82 | **39.67** | **48.55** | 43.19 | 66.23 | 45.48 |
| $\beta_t = 10$ | 63.58 | **58.96** | 22.69 | 24.69 | 38.88 | 47.86 | 42.77 | 65.20 | 45.93 |
| $\beta_t = 50$ | 63.98 | 58.88 | **23.49** | 26.03 | 38.71 | 47.56 | 43.11 | 65.57 | 45.00 |
| $\beta_t = 500$ | **64.07** | 58.85 | 23.26 | **26.27** | 39.45 | 48.50 | **43.40** | **66.48** | **46.03** |

display the detection results training with IDPO fine-tuning results for different $\beta_t$. Table 7 indicates that detection accuracy improves as $\beta_t$ increases. This suggests that when utilizing the entire patch as a sample, a larger $\beta_t$ is necessary to effectively constrain the model. We supplement a sensitivity analysis for $\beta_t$ and $\mu$, as shown in Fig. 13. The settings $\beta_t = 10$ and $\mu = 0.5$ yield the clearest human details and the most complete vehicle contours.

To verify the fusion quality, as shown in Fig. 14, a blind study with 10 annotators on 20 images uses a 1–5 Likert scale to assess thermal target visibility and texture details (5 = excellent). RLHF exhibits the most prominent thermal targets and texture details.

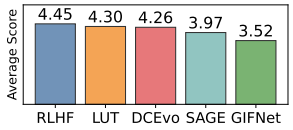


Figure 14. Preference evaluation result.

## C.3. Complexity and runtime comparison

We evaluate the model size, computational complexity, and inference time of various methods on input images of size $256 \times 256$. Specifically, for the core diffusion components of DPOFusion, the PALDM contains 420.26M parameters with 128.03 GFLOPs, while the PCLDM contains 179.51M parameters with 40.83 GFLOPs.

The DPOFusion performs the diffusion process in a compressed latent space ($64 \times 64$) rather than the original pixel space. Consequently, DPOFusion demonstrates significantly higher efficiency compared to pixel-space diffusion methods like DDFM. As shown in Table 8, DDFM requires approximately 7.8s per image, whereas DPOFusion reduces the inference time to 1.7s, achieving substantially fast inference among generative fusion methods.

While diffusion-based models generally incur higher computational costs than lightweight CNN-based approaches, DPOFusion is designed primarily as a preference-alignment framework. It serves as a powerful offline generator or teacher model to produce preference-aligned pseudo-labels. These high-quality samples can subsequently be used to supervise or distill lightweight models, allowing the transferred models to achieve real-time inference speeds while inheriting the preference-aligned capabilities of DPOFusion.

Table 8. Comparison of efficiency between DPOFusion and various methods.

| Method | Type | Size (M) | GFLOPs | Time (ms) |
|---|---|---|---|---|
| U2Fusion | General | 0.66 | 86.44 | 85.51 |
| SHIP | General | 0.55 | 35.16 | 33.23 |
| EMMA | General | 1.52 | 8.86 | 17.04 |
| Text-IF | Text-guided | 215.12 | 338.99 | 45.52 |
| DCEvo | General | 2.02 | 131.36 | 40.10 |
| GIFNet | General | 0.61 | 39.81 | 13.06 |
| LUT-Fuse | General | **0.0078** | - | 2.47 |
| SAGE | Semantic-aware | 0.14 | **4.34** | **1.60** |
| *Generative methods* | | | | |
| DDFM | General | 552.81 | 1840.49 | 7768.12 |
| DPOFusion | Preference-aligned | 778.4 | 2063.28 | 1709.86 |

## References


[1] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn*, pages 8748-8763, 2021. 1

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al., Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 1

[3] Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1

[4] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.*, 34:12077–12090, 2021. 2

[5] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 2, 4