

Supplementary Materials

Anonymous Authors

1 ABLATION STUDY

This section supplements the ablation study in the paper. In the ablation study, we conducted experiments to investigate the effectiveness of various components of the proposed dual-stage condition injection strategy. We are now designing experiments to validate another contribution of the paper, the heterogeneous feature alignment strategy.

1.1 Comparison with Different Schemes

As shown in Table 1, we test different schemes on the SpeakingFaces Dataset. Specifically, we have three comparison schemes:

(1) **Baseline.** We employed a pre-trained Arcface model, originally trained on visible images, alongside a skin color classification head also developed from visible images. We froze these pre-trained parameters for feature extraction purposes when working with thermal images. Table 1 shows that the domain gap between modalities significantly impacts performance. Despite applying our dual-stage condition injection strategy to refine the generation process, the extracted features from thermal images exhibit results that are comparable to Variant A in the ablation study of the paper, which solely employs the LDM framework for translation. Given that our strategy is designed to enhance performance, it's clear that directly leveraging feature extractors across modalities does not necessarily improve results and could potentially reduce model efficacy.

(2) **Fine-tune.** Fine-tuning the pre-trained Arcface model and skin-color classification head on thermal images aims to bolster the model's ability to extract identity features. This process can partially alleviate the degradation of feature extraction capability caused by domain discrepancy compared to the baseline. As shown in Table 1, all four metrics demonstrate improvement, confirming the paper's assertion that reducing the domain gap is both effective and necessary for feature extraction.

(3) **Alignment.** Our proposed heterogeneous feature alignment strategy surpasses the fine-tuning method, as evidenced by the results in Table 1. Across all metrics, our approach demonstrates superior performance, enhancing the model's ability to extract features from thermal images beyond both the baseline and fine-tuning techniques.

2 COMPARISON

This section supplements the Comparison subsection in the experimental section of the paper. Most of the visualizations of the model performance in the paper are based on the SpeakingFaces Dataset. To more comprehensively demonstrate the generalization ability of DiffTV, we provide additional visual comparison results on the ARL-VTF dataset. Similarly, our comparison methods mainly include HiFaceGAN [6], AxialGAN [1], DiFaReli [5], DiffuseIT [2], T2V-DDPM [4], and BBDM [3].

2.1 Qualitative Analysis

As depicted in Figure 1, we observe that thermal images in the ARL-VTF dataset contain fewer facial texture details compared to

Table 1: Ablation study on the SpeakingFaces Dataset. The results are obtained by the baseline and two designed schemes including our proposed one. The best result is highlighted by bold. Arrows indicate the desired direction for each metric, with \uparrow meaning higher is better, and \downarrow meaning lower is better.

Variants	FID \downarrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow
Baseline	34.68	0.1681	29.65	0.7143
+ Fine-tune	33.38	0.1649	29.79	0.7518
+ Alignment (Ours)	31.67	0.1553	30.42	0.7832

the SpeakingFaces Dataset, making the face translation task more challenging on this dataset. The facial details in the results of HiFaceGAN [6], DiFaReli [5], and DiffuseIT [2] in Figure 1 exhibit noticeable facial deformations. Additionally, other methods struggle with maintaining skin color consistency and preserving facial identity, with significant deviations from real faces. In contrast, our method delivers high-quality, accurate results that closely match the skin color and identity details of real faces. This not only proves the effectiveness of DiffTV but also underscores its robust generalization for thermal-to-visible face translation across varying scenarios.

REFERENCES

- [1] Rakhil Immidiseti, Shuowen Hu, and Vishal M Patel. 2021. Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan. In *IEEE International Joint Conference on Biometrics*. IEEE, 1–8.
- [2] Gihyun Kwon and Jong Chul Ye. 2023. Diffusion-based Image Translation using disentangled style and content representation. In *International conference on learning representations*.
- [3] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. 2023. Bbdtm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1952–1961.
- [4] Nithin Gopalakrishnan Nair and Vishal M Patel. 2023. T2V-DDPM: Thermal to visible face translation using denoising diffusion probabilistic models. In *International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–7.
- [5] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. 2023. DiFaReli: Diffusion face relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22646–22657.
- [6] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. 2020. Hifacegan: Face renovation via collaborative suppression and replenishment. In *Proceedings of ACM international conference on multimedia*. 1551–1560.



Figure 1: Comparative visualization of thermal-to-visible face translation results across various models, showcasing the advancements achieved with our DiffTV method against other approaches. The samples source from the ARL-VTF dataset.