Dirichlet-Constrained Variational Codebook Learning for Temporally Coherent Video Face Restoration — Supplementary Materials

Anonymous ICCV submission

Paper ID 8594

1. Experiment Details

Dataset. For blind face restoration (BFR) task, we adopt the degradation pipeline from [7] to generate low-quality (LQ) videos through a multi-stage process:

$$x = \{[(y \otimes k_\sigma)_{\downarrow_r} + n_\delta]_{\text{FFMPEG}_q}\}_{\uparrow_r}, \quad (1)$$

where the HQ video y is first convolved with a Gaussian kernel k_σ, and then downsampled to scale r. After that, additive Gaussian noise n_δ is added to the video, and then the video coding degradation is implemented through FFPMEG [4] with a quality factor q. Finally, the LQ video is resized back to 512×512. During training, we randomly sample σ, r, δ and q for [1, 21], [1, 4], [0, 20], [10, 30] respectively.

Settings. For all facial restoration training tasks, we employ the Adam [3] optimizer with a batch size of 4 and train for 100k iterations. The initial learning rate is set to 5×10^{-5}, and is gradually reduced to 2×10^{-5} using a cosine annealing strategy, after which it remains constant. To ensure training stability, we apply gradient clipping with a maximum norm of 1.0 to the model gradients during training. Our method is implemented based on Pytorch Framework and trained use two NVIDIA A100 GPUs.

Metrics The evaluation framework for our facial restoration model incorporates multiple complementary measurement dimensions. For image restoration quality assessment, we employ three conventional metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [6] quantify pixel-wise reconstruction fidelity at the individual frame level, while Learned Perceptual Image Patch Similarity (LPIPS) [9] evaluates human-aligned visual similarity through deep feature analysis. To specifically monitor facial identity preservation during restoration, we implement Identity Similarity (IDS) measurements by cosine similarity of the off-the-shelf identity detection network ArcFace [1]. To comprehensively evaluate pose alignment in restored video sequences, we integrate the Average Keypoint Distance (AKD) [8], which calculates mean landmark deviations between detected landmarks and the generated and ground-truth video frames.

At the temporal analysis level, we employ the Fréchet Video Distance (FVD) [5] to conduct a comprehensive quality assessment by comparing the distributions of generated videos with reference samples across both spatial and temporal domains. Given that our task involves video face restoration, it is crucial to evaluate facial continuity. To this end, we utilize the method proposed by [2] to extract five facial landmarks from both restored and real faces. We then compute the L2 distance between the landmark displacements of real faces and those of restored faces, which we term the Temporal Landmark Motion Error (TLME). This metric effectively assesses the temporal consistency of facial components throughout the video sequence.

2. More Experiments

Generalization on Out-of-Domain Data. We compare the methods on out-of-domain videos for the blind video face restoration task, and the results are shown in Figure 3. On the videos that are very different from the training data, our approach produces more reasonable results and shows superior generalization ability.

Temporal Stability. For both blind video face restoration and inpainting, we visualize the prediction results on VFHQ-Test dataset to evaluate the temporal stability of the methods. As shown in Figure 1, our method produces more consistent faces and preserves temporal continuity over time.

Temporal Transformer. We compare different Transformer architectures in Table 1. By aggregating features at multiple time steps using Transformer, we achieve lower FVD for blind video face restoration on VFHQ-Test dataset, meaning that our design is helpful to produce face videos closer the ground truth. Moreover, the FVD is further reduced by feeding the concatenation of multi-scale (16, 32,

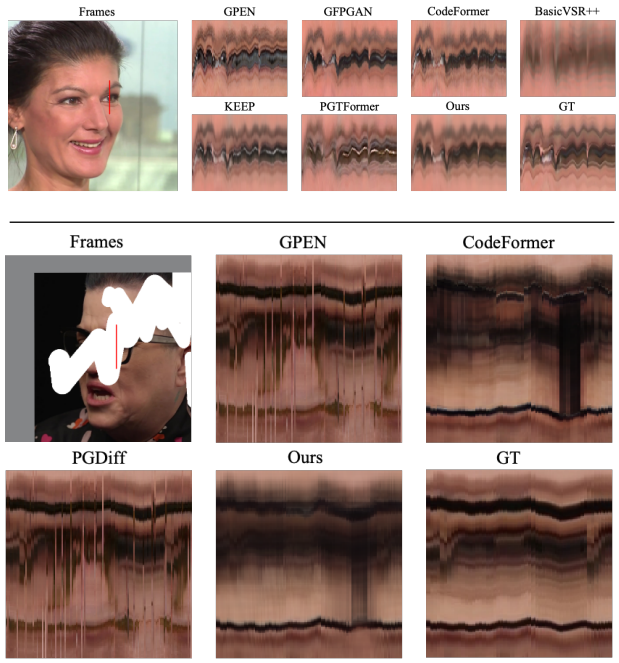

Figure 1. Comparison of temporal stability with other state-of-the-art methods for face restoration and inpainting on the VFHQ-Test dataset. We selected a column (red line) along the subject's eye and plotted its temporal variations over time. Our method exhibits significantly mitigated temporal jitter, enhancing spatial consistency across restored frames and preserving temporal continuity over time.

64) features of the encoder to the Transformer.

Method	SSIM↑	PSNR↑	LPIPS↓	FVD↓
(a) w/o	0.818	28.882	0.265	347.423
(b) single-scale	0.818	28.902	0.264	341.038
(c) multi-scale	**0.819**	**28.929**	**0.262**	**336.076**

Table 1. Ablation study of temporal Transformer for blind video face restoration on VFHQ-Test dataset. (a) means no temporal Transformer. (b) means feeding features with resolution 16×16 into Transformer. (c) means feeding concatenation of features with resolution 16×16, 32×32 and 64×64.

Number of Samples. During inference, we investigate the effect of sampling more $\hat{\mathbf{w}}$ from the posterior distribution and average the predictions $\mathbf{y}^{\texttt{pred}}$ as the final prediction. The results on VFHQ-Test dataset for blind video face restoration are shown in Table 2. We found the performance was stable with respect to the number of samples. Therefore, to save the computational cost we sample only one $\hat{\mathbf{w}}$ in other experiments.

Codebook Size. We conduct an ablation study to evaluate codebook capacity using sizes of 256, 512, and 1024 entries (Table 3). Increasing the codebook from 256 to 512 entries yields substantial improvements: AKD decreases by 20.5% (2.667→2.121) and FVD by 25.3% (449.907→336.216), indicating enhanced feature representation. Further expansion to 1024 entries provides marginal improvements—AKD decreases by 1.6% (2.121→2.088) and FVD by 0.04% (336.216→336.076)—suggesting performance saturation. Notably, all metrics achieve optimal values at 1024 entries, justifying this configuration for balancing representational capacity and computational efficiency.

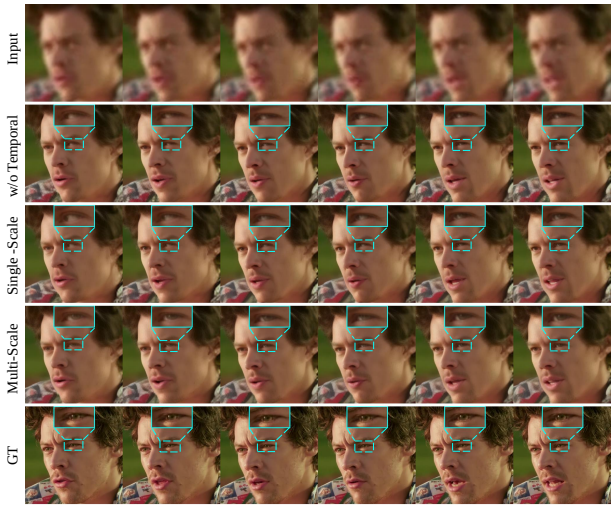

Figure 2. Ablation study on multi-scale design on the VFHQ-Test dataset. The visualized frames are sampled at 10-frame intervals. The multi-scale temporal design enhances temporal coherence, resulting in consistent gaze alignment for the central subject across sequential frames.

# $\hat{\mathbf{w}}$	PSNR↑	AKD↓	FVD↓	TLME↓
1	29.099	2.093	336.015	1.107
4	29.103	**2.091**	**332.290**	**1.099**
8	**29.104**	**2.091**	334.945	1.102
16	**29.104**	**2.091**	333.978	1.102

Table 2. Ablation for number of weight $\hat{\mathbf{w}}$ sampled from posterior distribution during inference for Blind Face Restoration on VFHQ-Test dataset. The final prediction is obtained by averaging predicted images decoded from all the sampled $\hat{\mathbf{w}}$.

3. Limitations and Future Work

Limitations. Despite demonstrating superior temporal consistency and restoration quality, two limitations remain. First, the sliding window inference strategy restricts temporal context to a local neighborhood (five frames), limiting long-range dependency modeling crucial for videos with abrupt scene changes. Second, the fixed codebook capacity may constrain representational diversity, occasionally yielding suboptimal reconstructions of rare facial attributes

Codebook sizes	PSNR↑	SSIM↑	LPIPS↓	IDS↑	AKD↓	FVD↓	TLME↓
256	28.033	0.801	0.289	0.889	2.667	449.907	1.203
512	28.916	**0.819**	0.264	**0.907**	2.121	336.216	1.110
1024	**28.929**	**0.819**	**0.263**	**0.907**	**2.088**	**336.076**	**1.091**

Table 3. Ablation for codebook sizes on the VFHQ-Test for blind face restoration.

Figure 3. Comparison of the BVFR on out-of-domain data. Our approach demonstrates more reasonable results that are more closely matched to the original images and appear more realistic. CodeFormer fails to generate reasonable outcomes, and the results of other methods also have some obvious unreasonable aspects.

or extreme expressions.

Future Work. Three promising directions emerge: (1) Developing adaptive codebook mechanisms that dynamically adjust code distributions based on video content could enhance feature diversity. (2) Extending the temporal context through hierarchical attention architectures may improve long-range dependency modeling. (3) Hybrid approaches integrating diffusion models with our variational framework could boost reconstruction fidelity for severely degraded inputs while maintaining temporal coherence.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[2] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10, 2006. 1

[5] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 1

[6] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[7] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 1

[8] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, et al. Human motion video generation: A survey. *Authorea Preprints*, 2024. 1

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1