1. TRAINING DETAILS

We used E^2FGVI HQ [2] and ProPainter [3] for a baseline pre-trained generator. The generator and discriminator are trained simultaneously using Adam optimizer for $5 \cdot 10^4$ iterations. Learning rate is set to $4 \cdot 10^{-5}$ for both models. For E^2FGVI, we set $\lambda_{\text{rec}} = \lambda_{\text{valid}} = 1, \lambda_{\text{flow}} = 0.01, \lambda_{\text{adv}} = 0.04$, and $\alpha_{\text{local}} = \alpha_{\text{global}} = 0.5$. For ProPainter, we set the values same as E^2FGVI except $\lambda_{\text{flow}} = 1$ and $\lambda_{\text{adv}} = 0.01$. During training, all frames are resized into 432×240 and the number of local frames and non-local frames (See E^2FGVI [2]) are set to 5 and 3, respectively. During evaluation and test, following the previous practices, we use sliding window with the size of 10.

Masks. While our primary target is outpainting 4:3 videos to 16:9 videos ($m = 1/4$), we fine-tuned the generator to mask ratio of minimum $1/12$ to maximum $1/3$ to increase robustness of the model.

Model architecture. For FEM, we stack three 3D convolutional layers with a spatial stride size of 2. The receptive field is $\approx 2^3 \cdot 7 = 56$ which is similar to the width of the outpainted region when mask ratio $m = 1/4$, 54. For FCM, we also stack three 3D convolutional layers with a spatial stride size of 2. The receptive field is $\approx 2^6 \cdot 7 = 448$ which is larger than the width of the training data, 432.

2. EXTENDED RESULTS

Here we present the VFID results of Tab. 3.

Method	$1/3$	$1/6$
Dehan *et al.* [4]	0.130	0.071
M3DDM [9]	0.277	0.120
E^2FGVI[2]	0.217	0.095
ProPainter[3]	0.193	0.105
Ours (E^2FGVI)	0.204	0.092
Ours (ProPainter)	0.156	0.075

Table 4. VFID by the outpainting ratios on the DAVIS dataset.

3. EXTENDED ABLATION STUDIES

3.1. Ablation on Additional Generator

Discriminator	PSNR	SSIM	VFID
w/o Fine-tuning	25.55	0.7861	0.193
T-PatchGAN [1]	26.06	0.7907	**0.167**
Ours	**26.24**	**0.7916**	0.177

Table 5. Quantitative comparison of discriminator design on DAVIS dataset and FuseFormer [15] generator.

As shown in Tab. 5, our fine-tuning framework increases the performance of FuseFormer [15] in both PSNR and SSIM metrics, compared to the T-PatchGAN discriminator. Thus, effectiveness of our method is not restricted to E^2FGVI[2] and ProPainter[3], and can be used with any video inpainting model.

3.2. Flow loss weight

λ_{gen}	λ_{flow}	PSNR ↑	SSIM ↑	VFID ↓
1	0.01	26.61	0.9385	0.139
1	0.1	26.43	0.9375	0.146
1	1.0	26.26	0.9363	0.147

Table 6. Ablation study on the flow loss weight on the DAVIS dataset. Note that E^2FGVI baseline is trained to $\lambda_{\text{flow}} = 1$.

As shown in Tab. 6, lower flow weight in generator loss led to a slight increase in all metrics. This is expected since the inpainting task that incorporates object mask during training is better for learning the flow estimation.

3.3. Generative loss weight

α_{inter}	α_{global}	PSNR ↑	SSIM ↑	VFID ↓
0.9	0.1	26.50	0.9383	0.149
0.1	0.9	26.31	0.9365	0.137
0.5	0.5	26.61	0.9385	0.139

Table 7. Ablation study on the local and global loss weight on the DAVIS dataset.

As shown in Tab.7, different configurations of hyperparameters do not markedly affect the performance in all metrics, highlighting the robustness of our method to hyperparameters.