We appreciate the thorough and constructive review provided by four expert reviewers **dxAk (R1)**, **SWSi (R2)**, **6XeZ (R3)**, and **BFry (R4)**. We will first address common questions (CQ), followed by responses(R) to questions (Q) from each reviewer. All these responses will be reflected in the final version.

CQ1: *Resolution and mask size and shape. (R1.Q2, R3.Q4)* **R:** Good point. Our model has been rigorously tested with mask ratios ranging from 0% to 80%. Although initially trained at 320x320 resolution due to computational limitations, GlobalPaint adapts well to higher resolutions like 320x640 and 640x320, maintaining high performance as shown in Supplementary Material Figure 3 and 4. With more computational resources, GlobalPaint could support larger resolutions. Regarding more mask shapes, we believe that rectangular is more applicable and leave other shapes as future work.

CQ2: *Analysis on global feature extraction. (R2.Q3, R4.Q1)* **R:** Good point. We apply all features, including the class token, from the penultimate layer of OpenCLIP instead of only the class token, which enhancing spatial detail information modeling throughout the framework. In the following Table, we compare our method with lightweight encoder (Lightweight-Enc) used in M3DDM, which uses Lightweight encoder features as global features. All models were trained for 40k iterations for evaluation due to limited rebuttal time. Results show that while Lightweight-Enc achieves comparable PSNR, SSIM, and LPIPS metrics, the FVD increased by 16.2% to 551.28, showing that our method significantly improves video outpainting quality, especially in motion naturalness.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
GlobalPaint	19.05	0.6777	0.1899	**474.59**
Lightweight-Enc	19.10	0.6762	0.1900	551.28

CQ3: *Efficacy Evaluation. (R3.Q2, R4.Q3)* **R:** Good suggestion. We report the trainable model parameters (Par.), computational complexity (FLOPs), inference GPU Memory (I-GPU), and inference time (I-Time) in table below. Our model has fewer parameters and higher inference efficiency. The unique classifier-free guidance design of M3DDM introduces additional computational complexity.

Method	Par.(M)	FLOPs(G)	I-GPU(GB)	I-Time(s)
M3DDM	1299	15667	30	44
GlobalPaint	1024	11177	31	17

R1.Q1 Q3: *Real-world applications and generalization to different types of videos.* **R:** We kindly remind the reviewer refer to the video results in the supplementary materials' show.html, including various real-world videos like sports and nature documentaries. Figure 1 shows our model successfully outpainting content in an animation video, demonstrating its versatility and practical applicability.

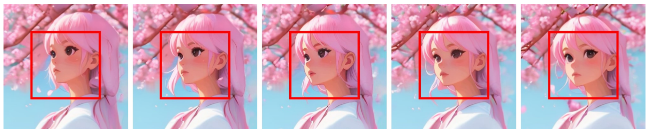

Figure 1: Results of animation video(red box region is input).

R2.Q2: *Two stage model* vs *one integrated model.* **R:** We streamlined the training process and reduced costs by using a two-stage model. GlobalPaint was trained on 4 A100 GPUs with a batch size of 32 for 720k iterations. In contrast, the resource-intensive M3DDM required 24 A100 GPUs and a batch size of 240 for 229k iterations, resulting in a computational cost about 2.4 times higher than ours.

R3.Q1 Q2: *Comparison with video inpainting model and Ewarp evaluation metric.* **R:** We compared GlobalPaint's effectiveness with the advanced video inpainting model, ProPainter. Inpainting models may produce blurry pixels when outpainting, potentially raising the PSNR. Unfortunately, due to the unavailable download link for the E_{warp} metric weight, we cannot access the pretrained weights and wait a response from the author. We plan to include the E_{warp} metric in the final version if the weights become available and more inpainting model comparisons.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
GlobalPaint	20.71	**0.7115**	**0.1685**	**251.6**
ProPainter(2023ICCV)	21.06	0.7065	0.2098	446.59

R3.Q3: *Denoising Steps vs. Video Quality.* **R:** The table below shows that as the number of steps increases, LPIPS and FVD decrease, indicating enhanced perceptual video quality and motion naturalness. Conversely, PSNR and SSIM increase at 20 steps due to more image blurriness and less detail, leading to greater pixel averaging. Overall, our default setting 50 steps provide a good balance between video quality and computational cost.

Steps	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
20	**20.91**	**0.7172**	0.1757	269.9
50	20.71	0.7115	0.1685	251.6
100	20.64	0.7089	**0.1667**	**247.3**

R4.Q1: *The effect of the number of learnable query feature.* **R:** The table below shows that reducing the global tokens to 64 nearly matches the performance of the default 256 tokens, while increasing to 1024 tokens significantly raises the FVD, likely from increased learning difficulty.

Method	G-Num.	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
GlobalPaint-64	64	19.10	0.6792	0.1906	477.12
GlobalPaint-256	256	19.05	0.6777	0.1899	**474.59**
GlobalPaint-1024	1024	19.01	0.6753	0.1883	513.14

R4.Q2: *Ablation study on window attention.* **R:** The table shows that the 5x5 window (EST-55T) outperformed the 3x3 (EST-33T) in the 40k model results. We haven't explored larger window sizes due to computational constraints but plan to analysis later.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
EST-55T	**18.99**	**0.6778**	**0.1902**	**528.48**
EST-33T	18.81	0.6731	0.1928	559.22

R4.Q4: *clarify on Confusing Results in the Table 1.* **R:** We learned from the M3DDM authors that their YouTube-VOS dataset is a non-standard version with extra videos, which could lower FVD due to its sensitivity to video count. Metrics like PSNR and SSIM are unaffected by the number of videos. In the final version, we will include complete metrics for Table 2 in the main text.

R4.Q5: *Frame rates.* **R:** The GlobalPaint model, trained at 24 fps, handles standard video speeds (24/30 fps) well and adapts effectively to higher fps. This strategy maximizes our limited resources while ensuring model flexibility and effectiveness.