

Blind Bitstream-corrupted Video Recovery via Metadata-guided Diffusion Model

Supplementary Material

This supplementary material is organized as follows:

- Correlation of Video Metadata and Corruption Patterns (Section 1).
- Frame Reconstruction Capabilities under Blind and Non-blind Settings (Section 2).
- Comparison of Mask Prediction (Section 3).
- More Visualization Results (Section 4).
- User Study (Section 5).
- Video Demo (Section 6).

1. Correlation of Video Metadata and Corruption Patterns

We showcase the correlation in Fig. 2. Bitstream corruption directly leads to the loss of motion vectors. Then H.264 decoder using incomplete motion vectors will decode unaligned images. It can be seen that the top and bottom halves of the truck in frame 009 are visibly unaligned.

2. Frame Reconstruction Capabilities under Blind and Non-blind Settings

To assess frame reconstruction capabilities, we test BSCVR and our approach using the same masks (Tab. 1). Compared to BSCVR, our method consistently performs better, with a 0.0043 improvement in LPIPS under GT masks (non-blind setting) and a 0.0059 improvement when using PMP masks. This indicates our model reconstructs more realistic frames by effectively exploiting generative priors of the diffusion model.

3. Comparison of Mask Prediction

We evaluate our Prior-driven Mask Predictor (PMP) against a fine-tuned SAM2 by measuring IoU with ground-truth (GT) masks. As shown in Tab. 2, PMP significantly outperforms SAM2 (65.4 vs 54.3 IoU). Since metadata implies the degradation of bitstream corruption and our PMP is metadata-aware, PMP can better handle complex corruption patterns. In addition, as shown in Tab. 1, both BSCVR and our method perform better when GT masks are used instead of PMP masks. Moreover, replacing SAM2 masks with our PMP masks leads to a substantial performance boost in the baseline (24.50 \rightarrow 25.64), indicating that enhancing mask quality will further improve video recovery.

In Fig. 6, we show a comparative analysis of mask prediction methods for two types of corruption patterns: spatial-only corruptions and temporally propagating corruptions. In the first example (a), involving blocking artifacts and misalignment, the corruption is limited to in-

Table 1. Frame reconstruction performance on DAVIS (PSNR \uparrow / LPIPS \downarrow).

Method	PMP masks	GT masks
BSCVR	25.64 / 0.0399	27.36 / 0.0316
Ours	26.05 / 0.0340	27.38 / 0.0273

Table 2. IoU comparison on DAVIS.

Method	IoU
SAM2-FT	54.3
PMP(Ours)	65.4

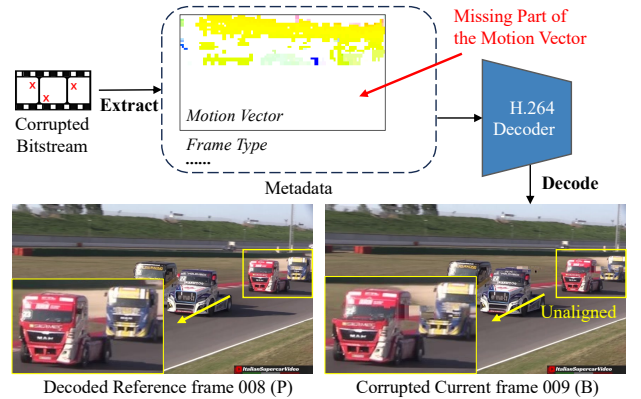


Figure 2. **Correlation of Video Metadata and Corruption Patterns.** Bitstream corruption causes missing motion vectors, leading to unaligned decoded frames. Zoom in for the best view.

dividual frames. Our method demonstrates a significantly better ability to detect these spatial distortions compared to the fine-tuned SAM2 [3], aligning closely with the ground truth (GT) masks. In the second example (b), which involves blocking artifacts combined with duplication artifacts, the corruption propagates temporally across consecutive frames. Our method outperforms SAM2 in capturing the temporal spread of distortions and producing masks that accurately represent the corruption patterns over time. These results emphasize the robustness of our approach in addressing both static and dynamic corruption scenarios, ensuring precise identification of corrupted regions and better alignment with ground truth.

4. More Visualization Results

To comprehensively evaluate the effectiveness of our recovery method, we divide our visual demonstration into two parts: subtle corruption patterns, where the visual disruptions are minor but still noticeable, and severely degrading corruption patterns, where the distortions significantly impair perceptual quality. This structured presentation showcases the versatility and robustness of our method in addressing a wide range of challenges, highlighting its ability to restore spatial and temporal coherence in both mild and extreme scenarios.

In the first part, we focus on subtle corruption patterns

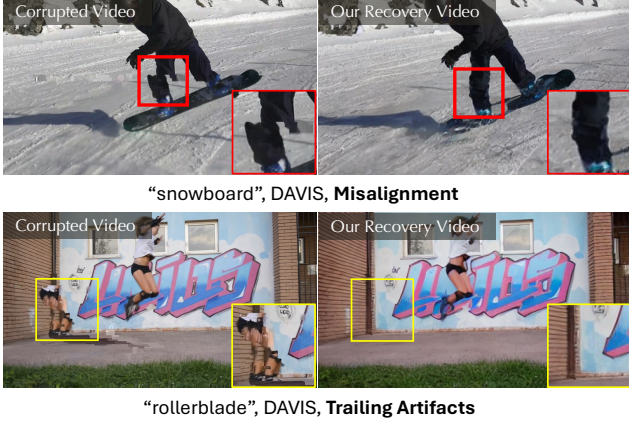


Figure 3. **Recovery Visualization for Subtle Corruption Patterns.** This figure highlights the recovery results for corruption patterns that only slightly affect visual quality, such as (1) misalignment and (2) trailing artifacts. Despite their subtle nature and limited visual disruption, these patterns can still impact perceptual consistency. The results demonstrate the sensitivity of our model in detecting and addressing even minor corruptions, ensuring high-quality restoration and maintaining coherence in the recovered videos.

that cause only minor visual disturbances but still impact the overall perceptual experience, as shown in Fig. 3. Specifically, we address (1) misalignment, which introduces slight spatial inconsistencies within one frame, and (2) trailing artifacts, which create faint ghosting effects or residual traces of previous frames. While these distortions are less severe, they can disrupt the smoothness and coherence of a video. Our recovery method demonstrates exceptional sensitivity to these minor corruptions, effectively realigning spatial components and eliminating residual artifacts. The results ensure seamless transitions between frames and preserve visual coherence, emphasizing the precision of our approach in handling subtle imperfections.

In the second part, we address severely degrading corruption patterns, as illustrated in Fig. 5. These include (1) color artifacts, which cause unnatural hues and distortions; (2) blocking artifacts, characterized by visible block boundaries and sharp discontinuities; (3) duplication artifacts, which disrupt spatial coherence by creating repeated content; and (4) texture loss, which removes essential fine details required for scene comprehension. Our recovery method effectively tackles these severe distortions by restoring natural colors, reconstructing textures, aligning repeated regions, and smoothing blocky artifacts. The results demonstrate the robustness of our method in restoring perceptual consistency and visual fidelity even in highly degraded scenarios.

Additionally, the results presented also highlight our method’s ability to restore temporal consistency, a crucial aspect for maintaining natural perception in video content. Severe distortions often disrupt the continuity of motion and

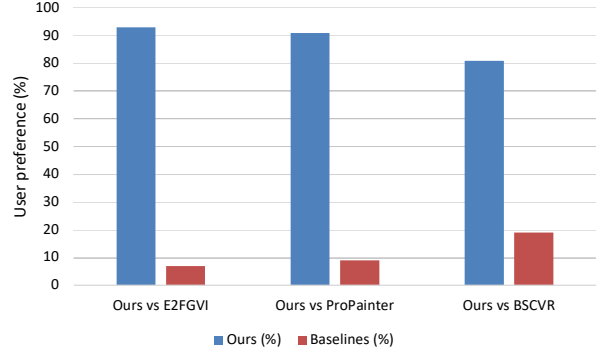


Figure 4. **User Study Results.** Our results are preferred by human users over the previous methods.

structure across frames, leading to abrupt changes and incoherent sequences. Our method successfully eliminates such disruptions, delivering smooth transitions and consistent motion across frames, thereby ensuring a natural and immersive viewing experience even in dynamic video scenes.

5. User Study

In this section, we conduct a user study to further compare previous video recovery methods, *i.e.*, E²FGVI [1], ProPainter [5], BSCVR [2] and our M-GDM. We invite a total of 30 participants for this user study. Each participant is presented with 20 recovery video sets: the input video, the recovered video by one of the previous methods, our recovered video. We ask them to select the visually better video from each set. The final results are summarized in Fig. 4. Through comparison, we conclude that our method clearly outperforms previous baselines.

6. Video Demo

We also provide demo videos to showcase the results evaluated in additional scenarios from the test set [4]. Please refer to the supplementary materials for the *Demo-012257ffcf.mp4* and *Demo-5fc34880f7.mp4* videos. These videos visualize the predictions of E²FGVI [1], ProPainter [5], BSCVR [2] and our M-GDM, along with the corrupted video and the ground truth. The results demonstrate that our approach significantly outperforms other state-of-the-art (SOTA) methods.

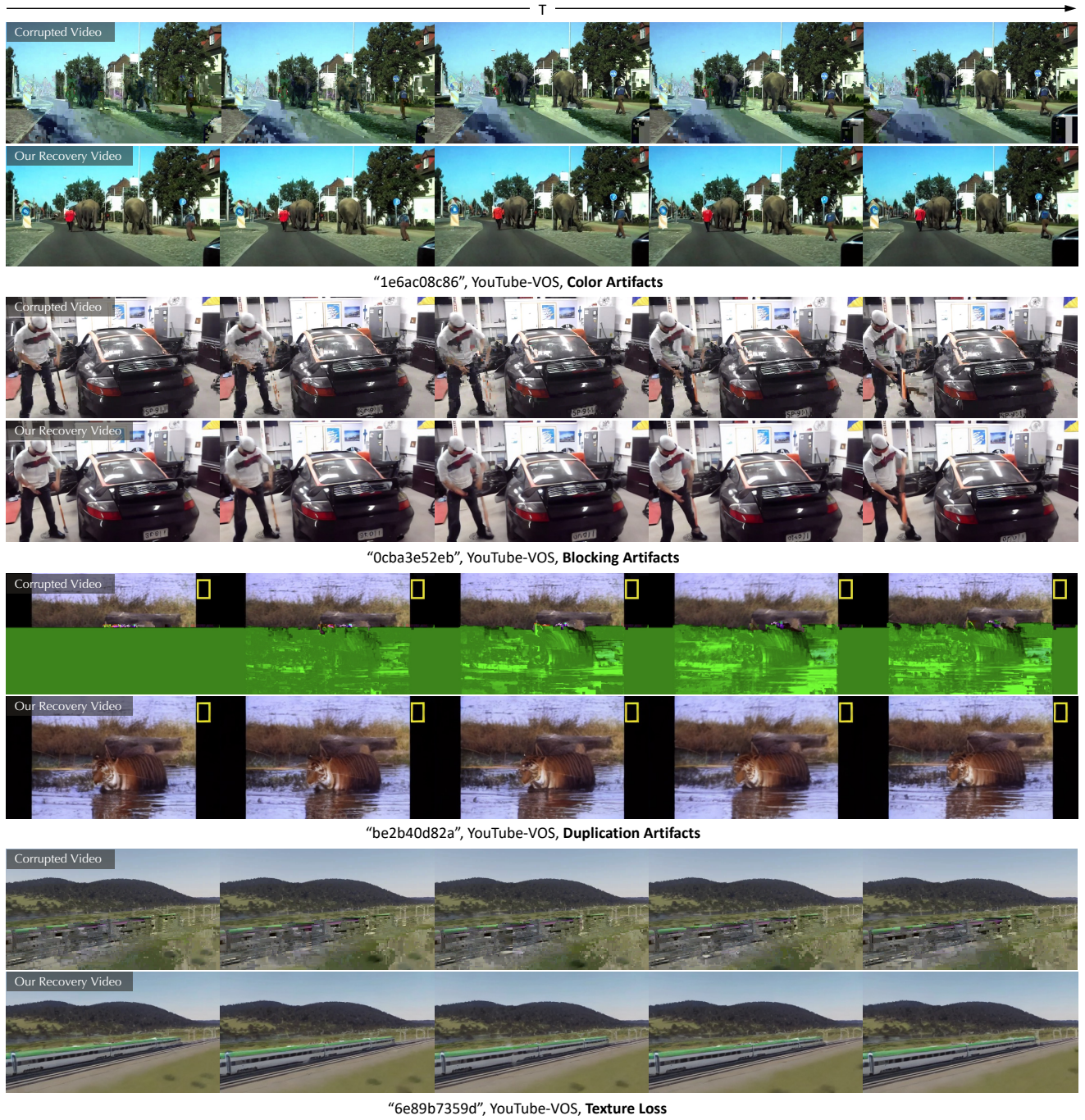
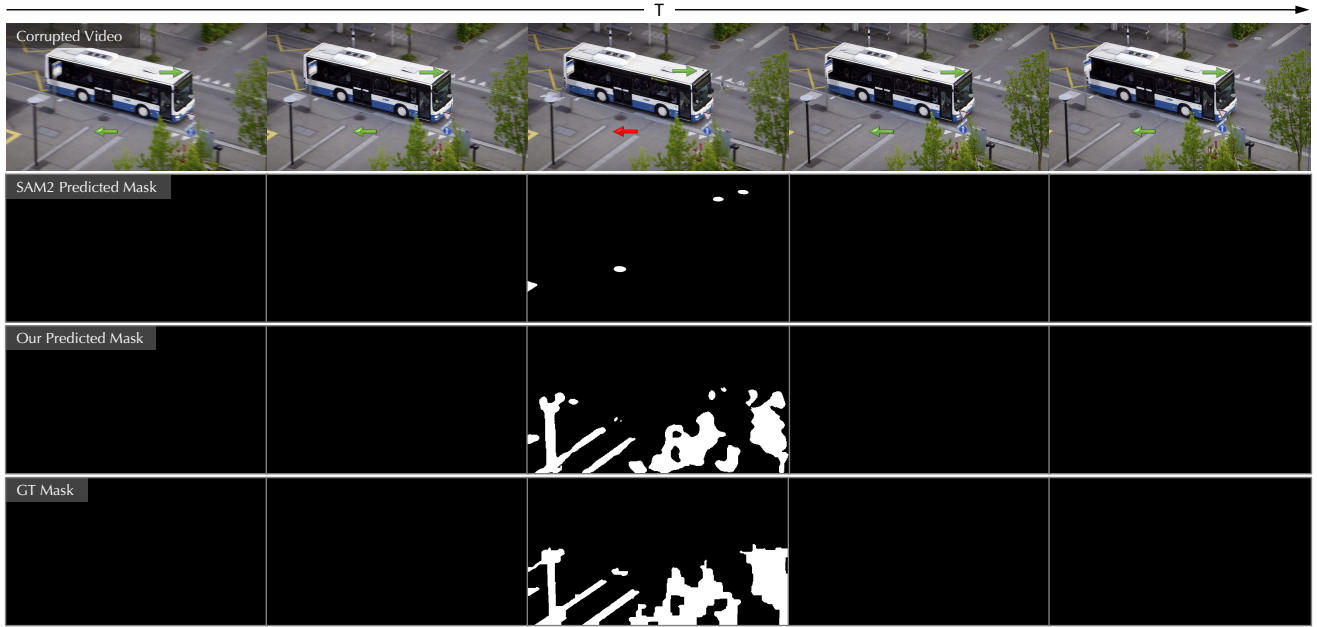
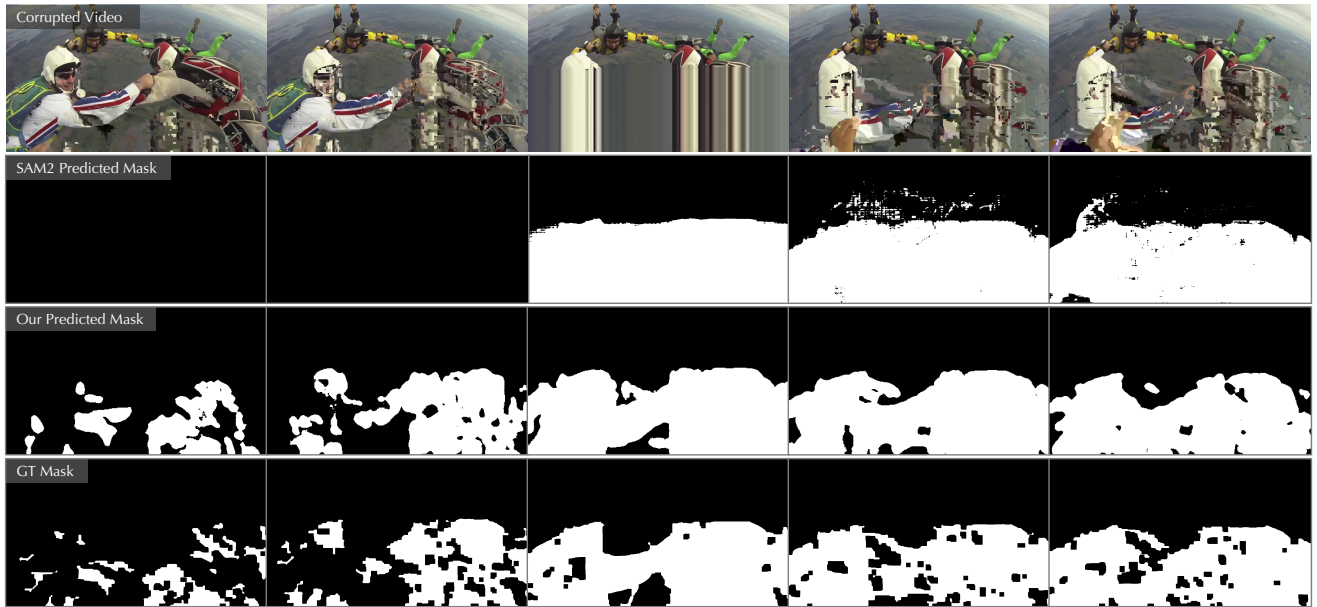


Figure 5. **Recovery Visualization for Severely Degrading Corruption Patterns.** The figure illustrates the recovery results for corruption patterns that severely impact visual quality, including (1) color artifacts, (2) blocking artifacts, (3) duplication artifacts, and (4) texture loss. These patterns represent highly disruptive scenarios where visual degradation significantly affects the viewer's experience. The recovery results demonstrate the method's ability to restore perceptual consistency and effectively mitigate severe visual disruptions.



(a) "bus", DAVIS, Blocking Artifacts + Misalignment (See caption for details)



(b) "4a90394892", YouTube-VOS, Blocking Artifacts + Duplication Artifacts

Figure 6. Comparison of Different Mask Prediction Methods. The figure illustrates the predicted masks for two types of corruption patterns: (a) spatial-only corruption patterns, and (b) corruption patterns emerging across frames. For each case, we compare the performance of our method with fine-tuned SAM2 [3], alongside the ground truth (GT) masks. Our method demonstrates a clear advantage in accurately identifying both spatial and temporal corruptions. Note that in the middle frame of (a), there is a misalignment with the top part as the bottom part is not moving (as it should be).

References

- [1] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. [2](#)
- [2] Tianyi Liu, Kejun Wu, Yi Wang, Wenyang Liu, Kim-Hui Yap, and Lap-Pui Chau. Bitstream-corrupted video recovery: a novel benchmark dataset and method. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [1](#), [4](#)
- [4] Linjie Yang, Yuchen Fan, and Ning Xu. The 2nd large-scale video object segmentation challenge-video object segmentation track, 2019. [2](#)
- [5] Shangchen Zhou, Chongyi Li, Kelvin C.K Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2023. [2](#)