# BiNar: Bitstream Native Restorer for Bitstream-corrupted Image Restoration

**Anonymous authors**
Paper under double-blind review

## Abstract

Bitstream-corrupted image restoration (BCIR) aims to reconstruct faithful images from corrupted compressed bitstreams. However, corruption occurs in bitstream-domain may lead to unpredictable error propagation and distortion amplification in pixel-domain, and even result in complete undecodability by standard decoders, posing significant challenge to existing image restoration paradigms. To address this challenge, we propose **Bi**tstream **Na**tive **R**estorer (**BiNar**), a generative restoration framework that leverages fine-grained bitstream-native semantics to guide bitstream-corrupted image restoration. In BiNar, Bitstream Semantic Modeling (**BSeM**) extracts robust informative semantics and models sequential bytes directly from corrupted bitstream, enabling to generate fine-grained bitstream descriptions without relying on pixel decoding. Based on the descriptions, the Corruption-adaptive Multimodal Semantic Aggregation (**CMSA**) further extracts adaptive visual modal cues in decoded corrupted images and aggregates to bitstream descriptions. The multimodal aggregated fine-grained descriptions can serve as coherent restoration guidance. Finally, Blind Corruption Localization (**BCoL**) exploits the segmentation prior of visual foundation models to identify and locate error regions that requires recovery in corrupted images. The located error regions and the fine-grained descriptions jointly serve as visual prompt and language prompt of diffusion models image restoration, respectively. Extensive experiments under diverse bitstream corruption conditions show that BiNar outperforms all State-of-the-arts image restoration models regardless of perceptual quality or objective quality. Ablation studies demonstrate the effectiveness of fine-grained bitstream-native semantics in BCIR task.

## 1 Introduction

Image restoration aims to recover clean images from degraded inputs by exploiting the texture, structure, and semantic cues that remain visible. Faithful restoration therefore depends on whether these cues remain informative about the original image. However, bitstream corruption can severely distort or even eliminate these cues. Compressed image bitstreams are highly vulnerable to corruption during storage and transmission. Since bitstream corruption occurs before image decoding, localized errors may propagate through the decoding process and disrupt subsequent coded content Lam & Reibman (1995); Bourdon et al. (2005). As a result, even limited corruption can produce widespread and unpredictable pixel-domain distortions, including color shifts, structural displacement, missing content, and severe visual artifacts Liu et al. (2023b). Under such degradation, semantic modeling from the decoded image becomes increasingly challenging, as the observed pixels may no longer faithfully represent the original content. In more severe cases, bitstream corruption can even cause decoding failure, leaving no usable visual information from which semantic cues can be extracted. Such degraded or missing semantics make faithful image restoration particularly challenging.

As illustrated in Fig. 1, existing methods for bitstream-corrupted image restoration (BCIR) mainly attempt to recover visual information after decoding or exploit thumbnail guidance for restoration Liu et al. (2023b); Hu et al. (2025). When decoding produces severe color distortion, structural displacement, or missing regions, the remaining pixels provide limited information for recovering fine structures and semantic details. Moreover, thumbnail guidance is not universally available. Some compressed image files do not contain an embedded thumbnail, since common image formats such as Exif and JFIF do not require one Hamilton (2004). Even when a thumbnail is originally present,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

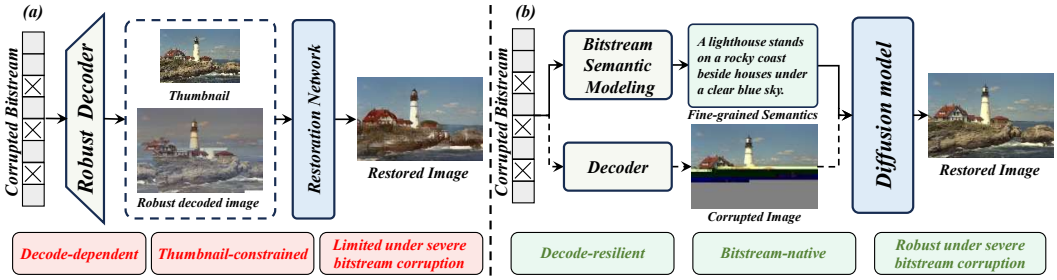


Figure 1: Paradigm comparison of bitstream-corrupted image restoration. (a) Conventional bitstream-corrupted image restoration relies on decoded visual observations and auxiliary thumbnail guidance Liu et al. (2023b). (b) Our method removes this dependence by extracting semantic guidance directly from corrupted bitstream for robust image restoration.

corruption of the file structure or associated metadata may prevent it from being correctly parsed and extracted Wiegand et al. (2003). As a result, this additional visual cue may also be unavailable in challenging corruption cases, further limiting restoration based on decoded visual information. A bitstream error can disrupt the decoding process and propagate to subsequent content, causing a much larger region to become distorted in the pixel Kuo et al. (2019). However, this propagation does not imply that the corresponding information has been equally lost from the compressed representation. Large portions of the coded data may remain unchanged even when the decoded image exhibits severe color distortion, structural displacement, or missing regions. These surviving bytes can therefore preserve information about the original image that is no longer faithfully expressed in the decoded observation. This makes the corrupted bitstream itself a promising source of semantic information for restoration.

To this end, we propose **Bitstream Native Restorer (BiNar)**, a generative restoration framework that exploits fine-grained bitstream-native semantics for bitstream-corrupted image restoration. The key contribution of BiNar is to establish a semantic pathway directly from corrupted bitstreams, providing informative restoration cues when pixel-derived semantics become unreliable or unavailable. Specifically, **Bitstream Semantic Modeling (BSeM)** directly models sequential bytes from corrupted bitstreams to extract fine-grained semantics without relying on pixel decoding. **Corruption-adaptive Multimodal Semantic Aggregation (CMSA)** further extracts useful visual cues from decoded corrupted images and aggregates them with the bitstream-derived semantics to obtain enhanced fine-grained descriptions. In parallel, **Blind Corruption Localization (BCoL)** exploits the segmentation prior of a visual foundation model to identify regions requiring recovery. The resulting fine-grained descriptions and localized corruption regions serve as language and visual prompts to jointly guide a pretrained diffusion model for image restoration. Extensive experiments across diverse bitstream corruption conditions show that BiNar achieves the best performance on seven of eight restoration metrics, while ablation studies further demonstrate the effectiveness of fine-grained bitstream-native semantics and the proposed components.

## 2 Related Work

### 2.1 Bitstream-Corrupted Visual Restoration

Early restoration of corrupted compressed images mainly relied on error concealment, using surviving spatial or transform-domain information to reconstruct damaged content (Gormish, 2003; Lee & Chen, 2003; Bourdon et al., 2005). More recent BCIR methods improve decoding robustness and exploit auxiliary information such as embedded thumbnails or neighboring coefficients for restoration (Liu et al., 2023b; Hu et al., 2025; Jiang et al., 2025). Generative approaches further introduce corruption localization, coding information, and visual foundation models to guide image or video recovery (Wang et al., 2025; Liu et al., 2023a; 2025; 2026). However, these methods still rely heavily on decoded visual information or auxiliary cues from the corrupted file. Under severe corruption, decoding artifacts or decoding failure can substantially reduce the information available for restoration, while auxiliary cues such as embedded thumbnails may also be unavailable or inaccessible.

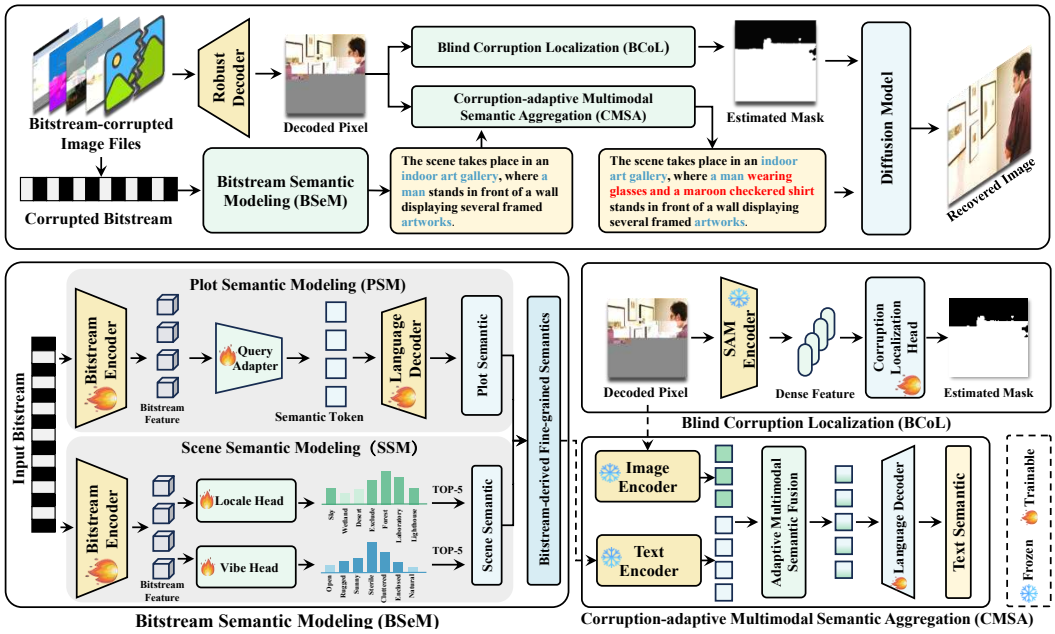


Figure 2: Overview of **BiNar**. BSeM extracts fine-grained bitstream-native semantics, CMSA aggregates complementary visual cues, and BCoL localizes corrupted regions. The resulting semantic and spatial prompts jointly guide diffusion-based restoration.

### 2.2 Semantic-Guided Image Restoration

Generative priors have substantially improved image restoration by enabling plausible reconstruction beyond the information directly available in degraded pixels. Early inpainting and diffusion-based methods recover missing content from known image regions (Yu et al., 2018; Suvorov et al., 2022; Lugmayr et al., 2022; Rombach et al., 2022). Recent restoration methods further incorporate large generative models and semantic conditions to improve perceptual quality and content consistency (Lin et al., 2024; Yu et al., 2024). More explicit semantic guidance has been introduced through vision-language models, textual prompts, and global or local visual representations (Luo et al., 2024; Wu et al., 2024a; Kong et al., 2025). However, such guidance is typically derived from degraded images and may become unreliable under severe bitstream corruption. This motivates extracting semantic information directly from the corrupted bitstream.

### 2.3 Bitstream-Domain Understanding

Bitstream-domain understanding aims to extract visual information directly from encoded data without relying on full pixel reconstruction. Early studies mainly exploited structured representations available in valid compressed images for recognition and segmentation (Duan et al., 2023; Ji & Karam, 2024). More recent work moves beyond codec-specific features and models raw byte sequences directly. ByteFormer (Horton et al., 2023), MEGABYTE (Yu et al., 2023), MambaByte (Wang et al.), and bGPT (Wu et al., 2024b) demonstrate that meaningful representations can be learned directly from byte-level inputs, reducing dependence on conventional decoding. Recent studies further extend this direction to corrupted bitstreams, where semantic information must be recovered from incomplete or damaged encoded data. Cibic (Liang et al., 2026) explores semantic extraction and caption generation from corrupted image bitstreams, while BLMSP (Huang et al., 2026) investigates semantic priors from corrupted video bytes for multiple downstream tasks. CBSU-ALLM (Wu et al., 2026) further treats corrupted image bitstream fragments as a byte-language modality for semantic understanding without successful pixel decoding. Together, these studies show that corrupted bitstreams can retain useful semantic information beyond what is recoverable from decoded pixels.

# 3 METHODOLOGY

## 3.1 FRAMEWORK OVERVIEW

We propose **Bitstream Native Restorer (BiNar)** for bitstream-corrupted image restoration (BCIR). BiNar exploits fine-grained bitstream-native semantics together with available visual information and corruption localization to guide generative restoration. As shown in Fig. 2, **Bitstream Semantic Modeling (BSeM)** directly models the corrupted byte sequence to extract Plot and Scene semantics, where Scene is characterized by Locale and Vibe. **Corruption-adaptive Multimodal Semantic Aggregation (CMSA)** aggregates these bitstream-native semantics with useful visual cues from the decoded corrupted image to produce a fine-grained description. In parallel, **Blind Corruption Localization (BCoL)** identifies regions requiring recovery. The resulting semantic description and corruption regions jointly guide a pretrained diffusion model for image restoration.

## 3.2 BITSTREAM SEMANTIC MODELING

BSeM models the corrupted byte sequence from two complementary semantic perspectives, **Plot** and **Scene**. Plot captures the main subjects, their appearance, actions, and interactions, while Scene describes the broader context through **Locale** and **Vibe**. Together, they provide fine-grained bitstream-native semantics for restoration.

**Plot Semantic Modeling (PSM).** Given a corrupted bitstream $\tilde{\mathbf{b}}$, PSM employs a ByteFormer encoder $E_p$ to obtain its bitstream representation. Since the resulting representation is variable in length, a Query Adapter $A$ with a fixed set of learnable queries aggregates it into compact semantic tokens,

$$\mathbf{Z}_p = A\big(E_p(\tilde{\mathbf{b}})\big). \tag{1}$$

The semantic tokens are projected into the embedding space of a language decoder and used to condition autoregressive generation. The decoder produces a Plot description $\mathbf{c}_p$ containing detailed subject, appearance, action, and interaction information.

**Scene Semantic Modeling (SSM).** SSM models the broader scene context through **Locale** and **Vibe**. Locale describes the type of place or scene, while Vibe characterizes its overall environmental attributes. Two independently trained ByteFormer-based classifiers estimate their category distributions,

$$\mathbf{p}_l = C_l\big(E_l(\tilde{\mathbf{b}})\big), \qquad \mathbf{p}_v = C_v\big(E_v(\tilde{\mathbf{b}})\big), \tag{2}$$

where $E_l$ and $E_v$ are the corresponding bitstream encoders, and $C_l$ and $C_v$ denote the classification heads.

Rather than retaining only the Top-1 prediction, SSM preserves the Top-$K$ Locale and Vibe candidates together with their confidence scores, denoted by $\mathcal{C}_l$ and $\mathcal{C}_v$. This retains multiple plausible scene hypotheses under corruption for subsequent aggregation.

**Training Objectives.** PSM is optimized with an autoregressive generation loss $\mathcal{L}_{\text{cap}}$, a token-level alignment loss $\mathcal{L}_{\text{align}}$, and a batch-wise contrastive loss $\mathcal{L}_{\text{con}}$,

$$\mathcal{L}_{\text{PSM}} = \mathcal{L}_{\text{cap}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{3}$$

In the first training stage, the Query Adapter and the final ByteFormer blocks are optimized while the language decoder remains frozen. In the second stage, ByteFormer is frozen, while the Query Adapter and low-rank adapters in the language decoder are jointly optimized. The same objectives are used in both stages. SSM is trained separately for Locale and Vibe prediction using classification supervision.

The Plot description $\mathbf{c}_p$, Locale candidates $\mathcal{C}_l$, and Vibe candidates $\mathcal{C}_v$ are subsequently passed to CMSA.

## 3.3 CORRUPTION-ADAPTIVE MULTIMODAL SEMANTIC AGGREGATION

CMSA aggregates the bitstream-native semantics produced by BSeM with useful visual cues from the decoded corrupted image. Given the Plot description $\mathbf{c}_p$, Locale candidates $\mathcal{C}_l$, Vibe candidates

$\mathcal{C}_v$, and the decoded image $\tilde{\mathbf{x}}$, the resulting fine-grained description is formulated as

$$\mathbf{c}_m = F_{\mathrm{CMSA}}\left(\mathbf{c}_p, \mathcal{C}_l, \mathcal{C}_v, \mathcal{V}(\tilde{\mathbf{x}})\right), \tag{4}$$

where $\mathcal{V}(\tilde{\mathbf{x}})$ denotes the available visual cues and $\mathbf{c}_m$ denotes the aggregated multimodal description.

When recognizable content remains in the decoded image, CMSA uses visual cues to complement and refine the bitstream semantics. As decoding artifacts become more severe, CMSA relies increasingly on the information extracted directly from the bitstream. When decoding fails, $\mathbf{c}_m$ is constructed solely from the BSeM outputs. The resulting description serves as the language prompt for restoration.

### 3.4 Blind Corruption Localization and Semantically Guided Restoration

BCoL identifies corrupted regions while preserving content that remains usable in the decoded image. Given $\tilde{\mathbf{x}}$, a frozen visual encoder $E_{\mathrm{vis}}$ extracts dense features, which are processed by a lightweight localization head $H_c$ to predict the corruption mask

$$\hat{\mathbf{m}} = \sigma\left(H_c\big(E_{\mathrm{vis}}(\tilde{\mathbf{x}})\big)\right). \tag{5}$$

BCoL requires no point, box, or mask prompts. During training, the localization head is supervised by the ground-truth corruption mask $\mathbf{m}$ using

$$\mathcal{L}_{\mathrm{BCoL}} = \mathcal{L}_{\mathrm{BCE}} + \mathcal{L}_{\mathrm{Dice}}, \tag{6}$$

while the visual encoder remains frozen.

The predicted mask and the fine-grained description produced by CMSA provide spatial and language guidance for a pretrained diffusion model,

$$\hat{\mathbf{x}} = G\left(\tilde{\mathbf{x}}, \hat{\mathbf{m}}, \mathbf{c}_m\right). \tag{7}$$

Here, $\hat{\mathbf{m}}$ specifies where recovery is required, while $\mathbf{c}_m$ provides semantic guidance for reconstruction. When decoding fails and no usable image is available, restoration falls back to the full image region using the semantics extracted by BSeM.

## 4 Experiment

### 4.1 Experimental Setup

**Dataset and Corruption.** We construct the training and evaluation sets based on ADE20K, which contains diverse visual scenes with annotations corresponding to the **Locale** and **Vibe** semantics modeled in BSM. Each image is resized to $256 \times 256$, center-cropped to $224 \times 224$, and encoded into an image bitstream. We consider two corruption types, Bit-Flip and Byte-Loss, which randomly invert bits and remove byte information, respectively. For each corruption type, corruption is applied within four stream ranges, covering $25\%$, $50\%$, $75\%$, and $100\%$ of the bitstream. Each of the resulting eight corruption configurations is independently applied to the complete test split, producing eight corrupted test subsets. For evaluation, each of the eight corruption configurations is applied to the complete test split, producing a separate corrupted test subset for each setting. All methods are evaluated on the same corrupted inputs.

**Implementation Details.** For semantic extraction, BiNar directly takes the corrupted bitstream as input, while vision-language baselines operate on the corresponding decoded image. For restoration comparison, the restoration backbone and sampling settings are kept consistent whenever applicable. Clean images are used only as references for training or evaluation and are never available during inference. Additional training details are provided in the supplementary material.

**Evaluation Metrics.** For semantic extraction, we report BLEU-1, BLEU-4, ROUGE-L, CIDEr, Token-F1, and BERTScore-F1. For image restoration, we report eight complementary metrics, divided into objective and perceptual categories. The objective metrics include PSNR, SSIM, MAE, and VFID, which evaluate pixel-level fidelity, structural consistency, reconstruction error, and distributional similarity, respectively. The perceptual metrics include DreamSim, LPIPS, CLIP-I, and DINO-I, which assess high-level visual similarity, perceptual feature consistency, and semantic or representation-level alignment between restored and reference images.

### 4.2 Experimental Results

Since fine-grained semantic descriptions serve as the language guidance for restoration in BiNar, their quality directly affects the fidelity of the reconstructed content. We therefore first evaluate whether BiNar can obtain more accurate and robust semantics under bitstream corruption, and further examine whether this semantic advantage translates into better restoration quality. We then compare the complete BiNar framework with existing image restoration methods.

#### 4.2.1 Evaluations on Bitstream-native Semantics

**Fine-grained Semantic Description Quality.** We compare the fine-grained semantic descriptions produced by BiNar with those generated by six representative vision-language models, including GLM-4.6V, InternVL3, Qwen3-VL, GPT-5.6, Gemini-3.8-Flash, and Claude Opus 4.6. The competing models infer semantics from decoded corrupted images, whereas BiNar combines bitstream-native semantics with useful visual cues to construct fine-grained descriptions. As shown in Table 1, BiNar achieves the best performance across all reported metrics. The examples in Fig. 3 further show that BiNar preserves image content, Locale, and Vibe more faithfully under bitstream corruption, demonstrating the quality and robustness of the resulting fine-grained descriptions.

**Robustness of Bitstream-native Semantics.** We further evaluate the robustness of bitstream-native semantics against semantics extracted from decoded corrupted images. The two variants use identical corrupted inputs, localization masks, diffusion models, and sampling settings, differing only in the semantic source. As shown in Fig. 4, both variants degrade as corruption increases, but restoration guided by bitstream-native semantics degrades more slowly. Consequently, the performance gap consistently widens from $25\%$ to $100\%$ corruption across DreamSim, LPIPS, CLIP-I, and DINO-I. In particular, the gap increases from $0.009$ to $0.160$ on DreamSim and from $0.027$ to $0.265$ on DINO-I. These results demonstrate that bitstream-native semantics remain more robust to increasing corruption and provide more stable restoration guidance when semantics inferred from decoded images become unreliable.

#### 4.2.2 Comparison with SOTA Image Restoration models

We compare BiNar with eight representative restoration methods, covering recent generative restoration approaches and methods designed for bitstream corruption. All methods are evaluated on the same corrupted test sets. Since our test images contain no embedded thumbnails, methods requiring thumbnail guidance use a downsampled version of the decoded corrupted image as auxiliary input, following their original preprocessing without introducing clean-image information. For video-based baselines, each corrupted image is treated as a single-frame sequence. When a fixed temporal length is required, the same corrupted frame is replicated along the temporal dimension, providing no additional temporal information.

**Quantitative Results.** Table 2 reports the average performance across all bitstream corruption configurations. BiNar achieves the best results on seven of the eight metrics, showing consistent improvements in reconstruction fidelity, perceptual quality, and high-level content consistency. Compared with the strongest competing results, BiNar improves PSNR from $15.2171$ to $15.8629$ dB and reduces MAE from $36.1204$ to $31.5975$. It further reduces VFID, DreamSim, and LPIPS to $14.8910$, $0.3680$, and $0.4231$, respectively. These improvements indicate that the restored images are not only closer to the clean references at the pixel level, but also exhibit better perceptual quality.

The advantage is more pronounced on high-level representation metrics. BiNar improves CLIP-I from $0.6029$ to $0.7210$ and DINO-I from $0.5920$ to $0.7103$, corresponding to gains of $0.1181$ and $0.1183$, respectively. This suggests that fine-grained bitstream-native semantics help preserve the underlying image content when decoded pixels alone provide insufficient or misleading information. BiNar also substantially improves over the corrupted input itself, confirming that the gains arise from effective reconstruction rather than simply preserving the damaged observation. Although B2SCVR obtains the highest SSIM, BiNar performs better across all other reported metrics, demonstrating a stronger overall balance between low-level fidelity and perceptual consistency.

Table 1: Quantitative comparison of fine-grained semantic descriptions under bitstream corruption. Higher is better for all metrics.

| Method | BLEU-1 | BLEU-4 | ROUGE-L | CIDEr | Token-F1 | Micro-F1 | BERTScore-F1 |
|---|---|---|---|---|---|---|---|
| GLM-4.6V | 0.1797 | 0.0505 | 0.1917 | 0.2707 | 0.2774 | 0.1969 | 0.6329 |
| InternVL3 | 0.1065 | 0.0242 | 0.1649 | 0.0958 | 0.2411 | 0.1555 | 0.6323 |
| Qwen3-VL | 0.0243 | 0.0061 | 0.1512 | 0.0236 | 0.2141 | 0.1405 | 0.6330 |
| GPT-5.6 | 0.2794 | 0.0693 | 0.1953 | 0.1440 | 0.3463 | 0.1939 | 0.9719 |
| Gemini-3.8-Flash | 0.0677 | 0.0169 | 0.1201 | 0.0186 | 0.1932 | 0.1388 | 0.9705 |
| Claude Opus 4.6 | 0.2593 | 0.0648 | 0.2093 | 0.1627 | 0.3130 | 0.1530 | 0.9698 |
| **BiNar (Ours)** | **0.4541** | **0.2248** | **0.4367** | **0.2934** | **0.5028** | **0.3435** | **0.9776** |

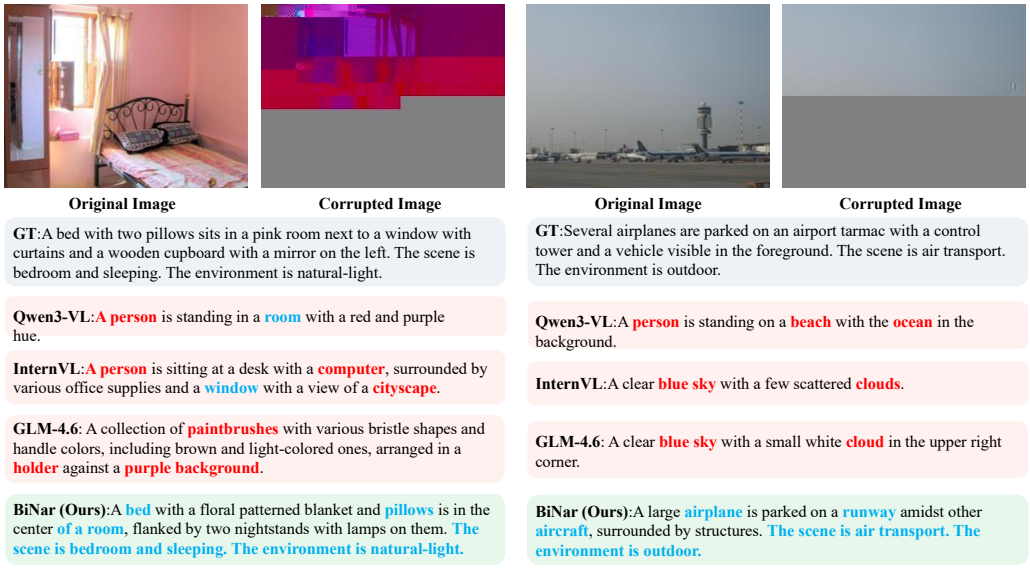


Figure 3: Qualitative comparison of fine-grained semantic descriptions under bitstream corruption. BiNar preserves image content and scene semantics more faithfully than vision-language models relying on decoded corrupted images.

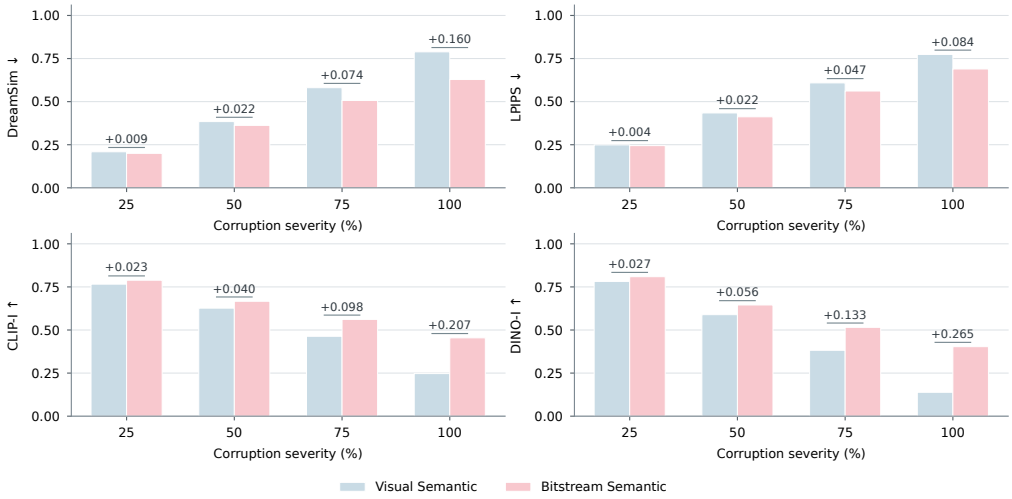


Figure 4: Restoration quality using visual and bitstream-native semantics across different corruption ranges. Results are averaged over Bit-Flip and Byte-Loss.

Table 2: Average quantitative comparison with image restoration methods across all bitstream corruption configurations.

| Method | Objective Metrics | | | | Perceptual Metrics | | | |
|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | MAE ↓ | VFID ↓ | DreamSim ↓ | LPIPS ↓ | CLIP-I ↑ | DINO-I ↑ |
| Corrupted Input | 12.6055 | 0.5179 | 46.7493 | 18.0492 | 0.5228 | 0.5679 | 0.5072 | 0.4676 |
| LD-RPS (ICCV 2025) | 8.9032 | 0.1571 | 79.2295 | 36.2306 | 0.7982 | 0.7768 | 0.2992 | 0.0734 |
| LucidFlux (ICLR 2026) | 11.9684 | 0.4075 | 51.1149 | 18.9616 | 0.4770 | 0.5294 | 0.5613 | 0.5210 |
| SeeSR (CVPR 2024) | 12.5476 | 0.5012 | 47.3506 | 16.1839 | 0.4327 | 0.4786 | 0.6018 | 0.5682 |
| HYPIR (SIGGRAPH 2025) | 12.2851 | 0.4710 | 49.1705 | 17.4295 | 0.4404 | 0.5049 | 0.5862 | 0.5566 |
| DATPRL-IR (ICLR 2026) | 11.2825 | 0.4942 | 56.6913 | 16.2785 | 0.4135 | 0.4785 | 0.6029 | 0.5920 |
| M-GDM (CVPR 2025) | 12.9040 | 0.4505 | 44.3498 | 18.9229 | 0.4977 | 0.4816 | 0.5325 | 0.4717 |
| B2SCVR (ACM MM 2025) | 15.2171 | **0.5683** | 36.1204 | 16.1131 | 0.4015 | 0.4533 | 0.5871 | 0.4886 |
| Two-ACIR (CVPR 2023) | 12.6263 | 0.5156 | 46.7522 | 32.0256 | 0.4316 | 0.4704 | 0.5803 | 0.5732 |
| **BiNar (Ours)** | **15.8629** | 0.5198 | **31.5975** | **14.8910** | **0.3680** | **0.4231** | **0.7210** | **0.7103** |

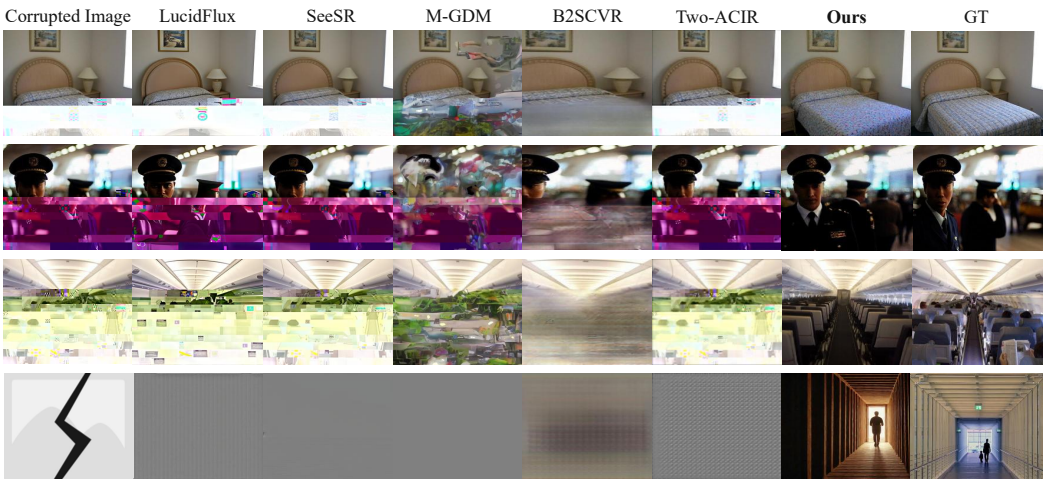


Figure 5: Qualitative comparison with image restoration methods under bitstream corruption. BiNar produces more faithful restoration results.

Table 3: Ablation on Plot–Scene decomposition in Bitstream Semantic Modeling (BSeM).

| Method | BLEU-1 | BLEU-4 | ROUGE-L | CIDEr | Token-F1 | BERTScore-F1 |
|---|---|---|---|---|---|---|
| Unified Semantic Generation | 0.3026 | 0.0686 | 0.2901 | 0.0426 | 0.3335 | 0.9680 |
| **Plot–Scene Decomposition** | **0.3743** | **0.1611** | **0.3682** | **0.1169** | **0.4200** | **0.9718** |

Table 4: Ablation on Corruption-adaptive Multimodal Semantic Aggregation (CMSA).

| Method | BLEU-1 | BLEU-4 | ROUGE-L | CIDEr | Token-F1 | BERTScore-F1 |
|---|---|---|---|---|---|---|
| Visual Only | 0.2395 | 0.1295 | 0.3032 | 0.2634 | 0.3521 | 0.6354 |
| Bitstream Only | 0.3919 | 0.1826 | 0.4187 | 0.1091 | 0.4771 | 0.9716 |
| **CMSA** | **0.4541** | **0.2248** | **0.4367** | **0.2934** | **0.5028** | **0.9776** |

**Qualitative Results.** Figure 5 provides visual comparisons under different bitstream corruption conditions. Existing restoration methods may retain corruption artifacts or reconstruct content that is inconsistent with the original image when the decoded input contains substantial distortion or missing information. In contrast, BiNar produces more coherent structures and recovers image content that is better aligned with the clean reference. These visual results are consistent with the quantitative improvements, particularly on LPIPS, CLIP-I, and DINO-I, and further demonstrate the benefit of combining fine-grained bitstream-native semantics with spatial corruption localization.

### 4.3 Validation Experiments

We conduct controlled experiments to validate three key designs of BiNar, including the Plot–Scene decomposition in BSeM, corruption-adaptive multimodal semantic aggregation, and blind corruption localization.

**Ablation on Plot–Scene Decomposition in BSeM.** We examine the structured semantic modeling design in BSeM by comparing it with unified semantic generation. The unified variant directly generates a complete semantic description, while BSeM decomposes the bitstream semantics into Plot and Scene, with Scene further characterized by Locale and Vibe. As shown in Table 3, the proposed decomposition consistently improves all semantic description metrics, demonstrating the effectiveness of structured semantic modeling in BSeM.

**Ablation on Corruption-adaptive Multimodal Semantic Aggregation.** We compare visual-only, bitstream-only, and the full CMSA. As shown in Table 4, the bitstream-only variant outperforms the visual-only variant on most metrics, while CMSA further improves semantic quality by incorporating complementary visual cues when they remain informative.

**Ablation on Blind Corruption Localization.** We compare BCoL with a generic SAM-based localization baseline. BCoL keeps the pretrained visual encoder frozen and trains only a lightweight corruption localization head. It improves mIoU from $0.7957$ to $0.9925$ and Dice-F1 from $0.8765$ to $0.9962$, demonstrating substantially more accurate localization of bitstream-induced corruption.

## 5 Conclusion

In this work, we presented Bitstream Native Restorer (BiNar), a generative framework for bitstream-corrupted image restoration that exploits fine-grained bitstream-native semantics for image reconstruction. BiNar uses Bitstream Semantic Modeling (BSeM) to extract semantic information directly from corrupted bitstreams without relying on pixel decoding. Corruption-adaptive Multimodal Semantic Aggregation (CMSA) further aggregates useful visual cues with the bitstream-derived semantics, while Blind Corruption Localization (BCoL) identifies regions requiring recovery. The resulting semantic descriptions and spatial prompts jointly guide diffusion-based restoration. Extensive experiments demonstrate that bitstream-native semantics remain more informative than pixel-derived semantics under bitstream corruption and provide effective guidance for restoring both visual quality and semantic content. BiNar consistently achieves strong restoration performance across diverse corruption conditions, while ablation studies validate the effectiveness of its key components. These results highlight corrupted bitstreams as a valuable native source of semantic information for robust image restoration.

## References

Pascal Bourdon, Bertrand Augereau, Christian Chatellier, and Christian Olivier. A multi-resolution, geometry-driven error concealment method for corrupted jpeg color images. *Signal Processing: Image Communication*, 20(7):681–694, 2005.

Zhihao Duan, Zhan Ma, and Fengqing Zhu. Unified architecture adaptation for compressed domain semantic inference. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 4108–4121, 2023.

Michael Gormish. Icon based error concealment for jpeg and jpeg 2000 images. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 2, pp. II–207. IEEE, 2003.

Eric Hamilton. Jpeg file interchange format. 2004.

Maxwell Horton, Sachin Mehta, Ali Farhadi, and Mohammad Rastegari. Bytes are all you need: Transformers operating directly on file bytes. *arXiv preprint arXiv:2306.00238*, 2023.

Qiongyang Hu, Yi Wang, and Lap-Pui Chau. Restoration of bitstream-corrupted images: A mamba-based thumbnail-guided network. In *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2025.

Chaoran Huang, Fangcheng Li, Tianyi Liu, Wenyang Liu, and Kejun Wu. Towards bitstream-corrupted harsh visual understanding: Through bitstream language modeling as robust semantic priors. *arXiv preprint arXiv:2608.21837*, 2026.

Ruolei Ji and Lina J Karam. Compressed-domain vision transformer for image classification. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 14(2):299–310, 2024.

Shumin Jiang, Hao Qin, Tianyi Liu, and Yi Wang. Optimizing jpeg decoder for bitstream-corrupted image restoration. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2305–2310. IEEE, 2025.

Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12809–12819. IEEE, 2025.

Yu-Chun Kuo, Ruei-Fong Chiu, and Ren-Shuo Liu. Long-term jpeg data protection and recovery for nand flash-based solid-state storage. In *2019 35th Symposium on Mass Storage Systems and Technologies (MSST)*, pp. 141–147. IEEE, 2019.

Wai-Man Lam and Amy R. Reibman. An error concealment algorithm for images subject to channel errors. *IEEE Transactions on Image Processing*, 4(5):533–542, 1995. doi: 10.1109/83.382489.

Pei-Jun Lee and Liang-Gee Chen. Error concealment algorithm using interested direction for jpeg 2000 image transmission. *IEEE Transactions on Consumer Electronics*, 49(4):1395–1401, 2003.

Junnan Liang, Kejun Wu, Xuanwei Hu, and Tianyi Liu. Cibic: Pixel-free foundation model for robust corrupted image bitstream captioning. *Pattern Recognition*, pp. 114238, 2026.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*, pp. 430–448. Springer, 2024.

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*, 36:68420–68433, 2023a.

Tianyi Liu, Kejun Wu, Chen Cai, Yi Wang, Kim-Hui Yap, and Lap-Pui Chau. Towards blind bitstream-corrupted video recovery: A visual foundation model-driven framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7949–7958, 2025.

Wenyang Liu, Yi Wang, Kim-Hui Yap, and Lap-Pui Chau. Bitstream-corrupted jpeg images are restorable: Two-stage compensation and alignment framework for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9979–9988. IEEE, 2023b.

Zhenyang Liu, Gang He, Li Xu, Tong Qiao, and Kepeng Xu. Beyond missing holes: Taming feature leakage in mask-guided bitstream-corrupted video recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2092–2098, 2026.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas Schön. Controlling vision-language models for multi-task image restoration. In *International Conference on Learning Representations*, volume 2024, pp. 16226–16246, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3172–3182. IEEE, 2022.

Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. Mambabyte: Token-free selective state space model, 2024. *URL https://arxiv. org/abs/2401.13660*.

Shuyun Wang, Hu Zhang, Xin Shen, Dadong Wang, and Xin Yu. Blind bitstream-corrupted video recovery via metadata-guided diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22975–22984. IEEE, 2025.

Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7): 560–576, 2003.

Kejun Wu, Fangcheng Li, Wenyang Liu, Qiong Liu, and You Yang. Corrupted bitstream semantic understanding by adaptive-modal large language models. *Pattern Recognition*, 180:114151, 2026.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25456–25467. IEEE, 2024a.

Shangda Wu, Xu Tan, Zili Wang, Rui Wang, Xiaobing Li, and Maosong Sun. Beyond language models: Byte models are digital world simulators. *arXiv preprint arXiv:2402.19155*, 2024b.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25669–25680. IEEE, 2024.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 5505–5514. IEEE, 2018.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. Megabyte: Predicting million-byte sequences with multiscale transformers. *Advances in Neural Information Processing Systems*, 36:78808–78823, 2023.