Rate-Constrained Edge Metadata for Sender–Receiver Generative Video Super-Resolution

Jiaqi Guo[1] **Mingzhen Li**[1] **Haohong Wang**[2] **Aggelos K. Katsaggelos**[1]
[1]Department of Electrical and Computer Engineering
Northwestern University, Evanston, Illinois, USA
{jiaqi.guo, mingzhen.li, a-katsaggelos}@northwestern.edu
[2]TCL
haohong.wang@tcl.com

Abstract

Real-world television and streaming pipelines often deliver compressed HD video to increasingly capable displays, making video super-resolution (VSR) both practically important and fundamentally ambiguous. Many generative VSR methods are formulated as receiver-only reconstruction, requiring the model to infer missing high-frequency structures from degraded pixels alone. We propose MetaVSR, a rate-accounted sender–receiver collaborative framework that transmits compact structural edge metadata as bitrate-constrained side information for generative VSR. MetaVSR builds on a one-step video diffusion transformer and encodes both the low-quality video and transmitted metadata through the native 3D VAE, enabling token-level fusion within the original DiT attention blocks without additional control branches. We further introduce a rate-aware Canny edge metadata selection strategy that allocates metadata bits to structurally challenging regions, and evaluate reconstruction quality under matched total transmission budgets that include both compressed video and metadata. Across five public VSR benchmarks, MetaVSR shows that budgeted sender-side structural metadata can improve reconstruction over the receiver-only DOVE-5B baseline, increasing PSNR by 1.03 dB and SSIM by 0.055 on average while using a smaller 2B backbone. On MetaSPS, our rate–distortion evaluation shows approximately 20% rate saving under light-noise degradation and more than 50% under heavy-noise degradation compared with a no-metadata generative baseline. These results support bitrate-efficient VSR with Canny edge metadata within the evaluated DOVE/CogVideoX receiver family; the allocation benefit depends on the total budget and degradation.

1 Introduction

Modern video delivery creates an information asymmetry between sender and receiver: the sender can access high-quality source frames before compression, bitrate adaptation, and platform-specific formatting, while the receiver observes a degraded low-resolution stream. This asymmetry is relevant to television and streaming pipelines, where high-resolution displays motivate enhancement of bandwidth-constrained video [1, 2]. Our experiments evaluate 720p reconstruction, with 4K serving only as application motivation.

This setting suggests that video enhancement should not be modeled purely as receiver-side post-processing. We ask a transmission-aware question: under a fixed total bitstream budget, should all bits be spent on the compressed base video, or should some be reserved for compact sender-side structural metadata? We formulate VSR as a sender–receiver collaborative reconstruction problem, as illustrated in Figure 1. The sender extracts and codes Canny edges before transmission, while the receiver combines decoded edges with the LQ stream to guide high-resolution reconstruction. The

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

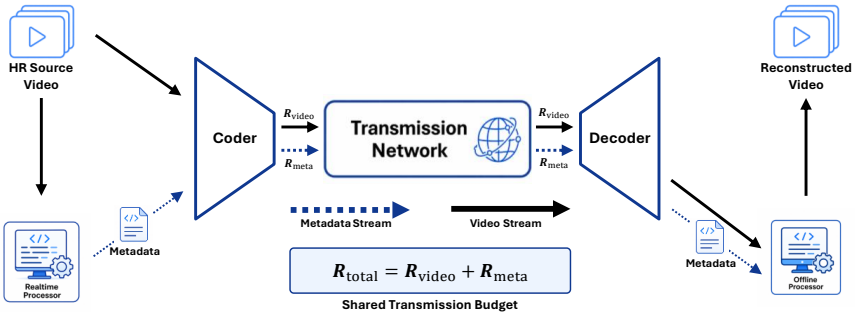


Figure 1: Sender–receiver system. The sender codes the LQ base stream and selected Canny edges extracted from HR frames. The receiver decodes both streams to reconstruct HR video. The independently decodable base stream and metadata share $R_{\text{total}} = R_{\text{video}} + R_{\text{meta}}$.

conventional base remains independently decodable, and metadata is an optional enhancement stream whose coded size is included in the total budget.

Most existing VSR methods formulate reconstruction as a receiver-only inference task, where all missing details must be inferred from degraded pixels alone. Although recent generative video models provide strong priors for plausible textures and motion, this formulation remains intrinsically ambiguous under compression and downsampling. Fine edges, small text, object boundaries, repeated patterns, and temporally varying structures may be suppressed before reconstruction begins. As a result, a powerful generator may synthesize visually realistic details, but can also hallucinate incorrect structures or introduce temporal inconsistency.

Under the sender–receiver perspective, the question is how to allocate a limited transmission budget between the compressed video and auxiliary metadata. Informative metadata can reduce reconstruction ambiguity by providing structural constraints that are difficult to recover from degraded pixels alone. However, metadata consumes bits otherwise available to the base video, so its benefit must be evaluated under the same total transmission budget. We focus on Canny edges as a compact and efficiently compressible structural signal, enabling a controlled analysis of metadata cost versus reconstruction gain.

Based on this motivation, we introduce MetaVSR, a metadata-guided generative VSR framework for sender–receiver collaborative reconstruction. MetaVSR builds on a one-step video diffusion backbone and treats metadata as a first-class side-information stream rather than a free auxiliary input. The receiver encodes both the low-quality video and decoded Canny edge metadata through the native 3D VAE/tokenization pathway of the video diffusion transformer, allowing metadata tokens to interact with video tokens through the original DiT attention blocks. This design avoids a separate control branch. Empirical validation is restricted to Canny edges and the DOVE/CogVideoX family; other modalities or backbones generally require adaptation. We evaluate MetaVSR under a rate–distortion framework that explicitly accounts for both compressed video bitrate and transmitted metadata bitrate.

Our core contribution is a rate-accounted sender–receiver generative VSR formulation and controlled evaluation of video–metadata allocation under a shared transmission budget.

Our main contributions are:

- **Budgeted sender–receiver VSR formulation.** We formulate generative VSR as a collaborative reconstruction problem under a matched total bitstream budget, where the receiver reconstructs HR video from both a compressed base stream and bitrate-constrained sender-side metadata.
- **Generative receiver implementation.** We implement metadata-conditioned VSR that encodes both low-quality video and decoded metadata through the native 3D VAE and fuses them as tokens inside the original DiT attention blocks, without adding a separate control branch.

- **Budgeted Canny edge allocation.** We study a sender-side edge selection strategy that allocates metadata bits to regions where the compressed base stream is least informative, making metadata a controllable part of the transmission budget.
- **Rate–distortion evaluation.** We evaluate MetaVSR under matched budgets where $R_{\text{total}} = R_{\text{video}} + R_{\text{meta}}$, showing that compact structural metadata can improve reconstruction quality and rate efficiency compared with receiver-only generative VSR.
- **Targeted MetaSPS benchmark.** We provide a sports-video benchmark and aligned metadata resources for studying transmission-aware reconstruction; Canny edges are the experimentally validated modality.

2 Related Work

2.1 Video Super-Resolution

Video super-resolution (VSR) extends image SR to spatio-temporal reconstruction, where models must recover high-frequency details while maintaining temporal coherence across frames. Classical deep VSR methods improve reconstruction through temporal propagation, motion alignment, recurrent aggregation, sliding-window fusion, or transformer-based modeling [3–6]. Real-world VSR further addresses complex degradations using paired real data, synthetic degradation pipelines, or architecture-level robustness [7–9]. However, most existing methods remain receiver-only: all missing details must be inferred from the degraded low-resolution stream after compression and downsampling. Recent generative VSR methods introduce diffusion priors to improve perceptual realism under severe degradation, including Upscale-A-Video, MGLD-VSR, STAR, and the one-step DOVE framework [10–14]. While these methods can synthesize visually plausible textures, the reconstruction remains fundamentally ambiguous when transmission artifacts remove fine edges, small text, object boundaries, and temporal details before recovery. In contrast, our work formulates VSR as a sender–receiver collaborative reconstruction problem, where compact metadata extracted before transmission is delivered as side information under a matched total bitrate budget, reducing reconstruction uncertainty without replacing the standard low-resolution video stream.

2.2 Conditioned Restoration and Metadata-Guided Reconstruction

Auxiliary information has been widely used to reduce the ambiguity of super-resolution when low-resolution pixels alone are insufficient. Prior image SR methods exploit edge maps to sharpen boundaries [15, 16], semantic maps to modulate class-aware textures [17], and reference images to transfer external high-frequency details [18, 19]. DiffBIR [20] studies blind restoration with a generative diffusion prior, and FullDiT [21] uses full attention for multi-modal video control. Native token fusion is therefore an implementation choice rather than a new conditioning principle. More recently, diffusion conditioning frameworks such as ControlNet and T2I-Adapter have shown that heterogeneous spatial controls, including edges, depth, and segmentation, can effectively guide pretrained generative models [22, 23]. However, these methods typically assume that the guidance signal is locally available and free to use, rather than treating it as a compressed, temporally aligned, and transmittable side-information stream. In contrast, our work studies metadata-guided VSR in a sender–receiver setting, where compact structural metadata is transmitted together with the compressed video and fused at the receiver under a matched total bitrate budget, connecting conditional reconstruction with practical rate–distortion-constrained delivery.

2.3 Rate-Aware VSR and Generative Communication

Rate-aware sender–receiver VSR already exists: LiveNAS [24], SRVC [25], DeepStream [26], and STDO [27] use content-adaptive SR, online learning, training patches, or model-update payloads. Generative semantic communication transmits compact semantic information for diffusion reconstruction [28, 29]. SPIC [30] conditions diffusion image reconstruction on an LR image and segmentation map; DiSCo [31] codes degraded video, text, and optional sketches or poses. Thus neither collaborative SR nor transmitted generative guidance is new. MetaVSR studies their joint use in fixed-scale generative VSR: an independently decodable LR base stream, optional HR-derived aligned Canny metadata, explicit accounting of both coded sizes, and controlled all-video versus video-plus-metadata

allocation within one receiver family. Direct-HR coding and learned codecs provide complementary system-level references; same-protocol learned-codec comparisons are not completed here.

2.4 Efficient Diffusion Backbones for Video Restoration

Diffusion models offer strong generative priors for restoration, but their iterative sampling process introduces substantial latency for video enhancement. Existing acceleration strategies reduce this cost through faster samplers and solvers [32–34], or by training few-step and one-step generators via distillation or consistency learning [35–38]. In VSR, DOVE recently adapts CogVideoX into a one-step real-world VSR model through latent–pixel training [13, 14]. Our work adopts this efficient one-step generative backbone as the receiver-side engine, but focuses on a different problem: how compact transmitted metadata can reduce reconstruction ambiguity under a matched bitrate budget.

3 Method

3.1 Overview and one-step VSR backbone

We use LQ for the degraded input video, including its LR spatial resolution; HR denotes the target high-resolution video. MetaVSR uses CogVideoX-2B as the receiver-side generative VSR backbone and follows the one-step adaptation strategy introduced in DOVE [13]. CogVideoX consists of a 3D causal VAE, with encoder $\mathcal{E}$ and decoder $\mathcal{D}$, and a Transformer velocity predictor v_θ operating in the latent space. Given a degraded low-quality video sequence x_{lq}, we resize it to the target spatial resolution and encode it as $z_{\mathrm{lq}} = \mathcal{E}(x_{\mathrm{lq}})$. Following the one-step generative super-resolution formulation, z_{lq} is treated as a noisy latent at timestep t, and the restored latent is estimated by

$$z_{\mathrm{sr}} = \sqrt{\bar{\alpha}_t}\, z_{\mathrm{lq}} - \sqrt{1-\bar{\alpha}_t}\, v_\theta(z_{\mathrm{lq}}, c, t), \tag{1}$$

where t is the diffusion timestep, $\bar{\alpha}_t$ is the cumulative diffusion noise-schedule coefficient, θ denotes the velocity-predictor parameters, and c contains text prompts and transmitted metadata. The final output is decoded as $x_{\mathrm{sr}} = \mathcal{D}(z_{\mathrm{sr}})$.

During training, we adopt the two-stage optimization strategy of DOVE [13]: the model is first adapted in the latent space by matching the predicted latent to the high-resolution latent, and is then refined in the pixel space using reconstruction, perceptual, and video temporal-consistency losses. Details of the latent–pixel training strategy are provided in Appendix B.

3.2 Unified Token Representation for Metadata Conditioning

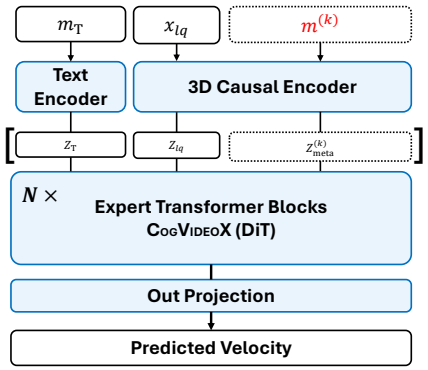


Figure 2: Unified metadata conditioning by encoding visual and metadata inputs into a shared latent token space.

To incorporate transmitted side information into generative reconstruction, we represent each metadata stream $m^{(k)}$ as an image-like tensor and encode it with the same 3D VAE encoder used for the visual input. Here k indexes the K metadata streams; this work uses Canny edge metadata ($K = 1$):

$$z^{(k)}_{\mathrm{meta}} = \mathcal{E}(m^{(k)}), \qquad k = 1, \ldots, K. \tag{2}$$

This maps the metadata into the same latent space as the visual input. Here lowercase z denotes a VAE latent and uppercase Z denotes a DiT token sequence. We apply the native patch embedding to z_{lq} and $z^{(k)}_{\mathrm{meta}}$ to obtain token sequences, and concatenate textual, visual, and metadata tokens as:

$$\mathbf{Z} = [Z_{\mathrm{T}}; Z_{\mathrm{lq}}; Z_{\mathrm{meta}}], \tag{3}$$

where Z_{T} denotes the text tokens, Z_{lq} denotes the low-quality video tokens, and $Z_{\mathrm{meta}} = [Z^{(1)}_{\mathrm{meta}}; \ldots; Z^{(K)}_{\mathrm{meta}}]$ contains all available metadata tokens. The concatenated sequence is processed by the original DiT backbone:

$$\mathbf{H}^{(\ell+1)} = \mathrm{DiTBlock}^{(\ell)}(\mathbf{H}^{(\ell)}; t), \tag{4}$$

where $\mathbf{H}^{(0)} = \mathbf{Z}$, ℓ indexes DiT layers, and $\mathbf{H}^{(\ell)}$ denotes the hidden state. Thus, metadata conditioning is realized through native joint self-attention, allowing structural side information to interact directly with visual tokens without additional conditioning branches or external fusion networks.

Position-aware metadata fusion. CogVideoX-2B uses fixed 3D sinusoidal positional encodings for visual tokens. Since the transmitted metadata is spatially and temporally aligned with the decoded video, each metadata token is assigned the same positional index as its paired visual token. This position-sharing design preserves correspondence between the low-quality input and metadata, enabling the DiT backbone to fuse side information at the correct spatial-temporal locations.

3.3 Rate-Aware Edge Selection for Super-Resolution

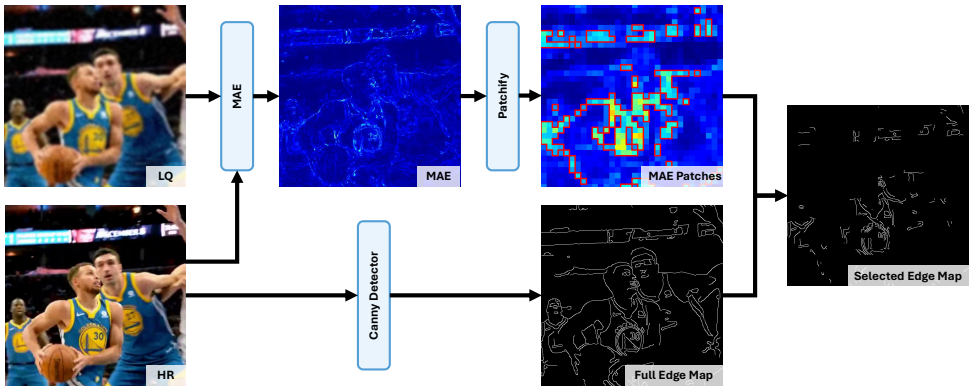


Figure 3: Illustration of the proposed rate-aware Canny edge metadata selection strategy for selecting transmitted Canny edge metadata under a constrained transmission budget.

Metadata-free VSR tends not to do well in regions where the low-quality input loses structural information, such as object boundaries, fine textures, and motion-dependent details. We therefore avoid transmitting dense Canny edge metadata over the entire frame and instead allocate metadata bits to spatial regions where the upsampled low-quality input differs most from the high-resolution source, as illustrated in Fig. 3. Let X_t and Y_t denote the HR and LQ frames at time t. We first bicubically upsample Y_t to the HR resolution and compute a single-channel error map E_t using the mean absolute RGB difference between X_t and the upsampled LQ frame. Here (u, v) indexes pixels, H, W are HR height and width, and P is patch side length. The error map is partitioned into non-overlapping $P \times P$ patches $\{\Omega_k\}$, with $|\Omega_k|$ pixels in patch k and mean mismatch score

$$s_t(k) = \frac{1}{|\Omega_k|} \sum_{(u,v)\in\Omega_k} E_t(u, v). \tag{5}$$

Given a selection ratio $\rho \in (0, 1]$, we rank all patches by $s_t(k)$ and select the top ρ fraction as structurally important. The selected patches are expanded into a binary spatial mask $M_t^\rho \in \{0, 1\}^{H\times W}$. We then extract a Canny edge map C_t from the HR frame using fixed hyperparameters and obtain the transmitted sparse Canny edge metadata by

$$\widetilde{C}_t^\rho = C_t \odot M_t^\rho, \tag{6}$$

where $\odot$ denotes element-wise multiplication. This operation preserves edge information only in high-mismatch regions, concentrating the metadata budget on areas where the LQ input is least informative. The selection ratio ρ provides a direct bitrate-control knob: smaller ρ yields sparser metadata and lower transmission cost, while $\rho = 100\%$ corresponds to transmitting the dense Canny edge map.

3.4 MetaSPS Dataset and Metadata Pipeline

We develop a multi-modality metadata extraction pipeline to build MetaSports (MetaSPS), a targeted benchmark of 49 sports clips spanning basketball, American football, tennis, and combat sports,

each standardized to 32 frames at 8 fps and 1280×720 pixel resolution. MetaSPS tests whether coded source-derived structure is worth its cost under a shared budget, rather than providing a scale contribution or a universally harder VSR dataset. Appendix F explains the benchmark motivation and its role relative to conventional VSR datasets. The associated extraction pipeline produces aligned skeleton, 3D pose, ROI contour, depth, and edge metadata for direct benchmark reuse and transfer to similar datasets. In this paper, we use Canny edge metadata as the primary validated metadata channel, while the remaining metadata streams are benchmark resources for future metadata-guided restoration research. Detailed extraction steps are provided in Appendix H and summarized in Fig. 12.

4 Experiments

Training setting. We fine-tune CogVideoX-2B on a mixed training corpus constructed from DIV2K [39] and HQ-VSR [13]. All experiments are implemented in PyTorch and trained on four NVIDIA A100 GPUs. Training and evaluation settings, including dataset preparation, optimizer settings, schedules, batch size, and degradation settings, are provided in Appendix A.

Simulation of practical video transmission. To model practical transmission-oriented degradation, we compress low-quality video inputs with H.264 [40], a widely used non-learning-based codec in streaming and broadcasting. We vary the quantization parameter (QP) to simulate different bitrate conditions for the transmitted low-quality video stream.

Canny edge metadata compression. We compress binary Canny edge metadata with JBIG2 [41], a standard codec for binary images. We control the metadata bitrate by varying the amount of selected Canny edge metadata as described in Sec. 3.3. JBIG2 also improves compression by exploiting repeated edge patterns across neighboring frames.

Transmission budget notation. We denote the total transmitted bitstream by

$$R_{\text{total}} = R_{\text{video}} + R_{\text{meta}}, \tag{7}$$

where R_{video} and R_{meta} denote the measured coded video and metadata sizes for the same evaluated sequences, reported in bytes (B).

4.1 Baseline Comparison

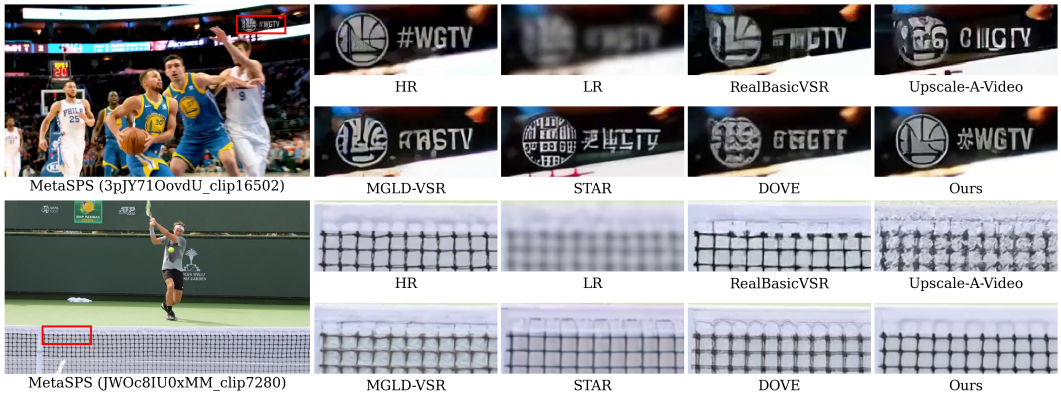


Figure 4: Comparison with Receiver-Only VSR Baselines under Budgeted Metadata.

We first compare MetaVSR with representative state-of-the-art VSR methods on five public benchmarks, including UDM10 [4], SPMCS [3], YouHQ40 [10], RealVSR [7], and MVSR4× [9], together with our proposed MetaSPS dataset. All coded sizes reported in Table 1 are in units of 10^3 bytes. For MetaSPS, we use the light-noise low-quality setting. The compared baselines include the reconstruction-oriented RealBasicVSR [8] and the generative methods Upscale-A-Video [10], MGLD-VSR [11], STAR [12], and DOVE-5B [13]. Metadata cost is counted, but external methods use released LQ inputs and official checkpoints without common-corpus retraining. This off-the-shelf comparison is not the primary causal evidence for metadata; the controlled 2B experiments below share training and degradation protocols. Accordingly, we ensure that the compressed metadata

Table 1: Quantitative comparison with current state-of-the-art methods on UDM10, SPMCS, YouHQ40, RealVSR, MVSR4×, and MetaSPS.

Dataset	Metric	No-meta allocation **Total Rate** (Video / Meta)	RealBasic VSR [8]	Upscale-A-Video [10]	MGLD-VSR [11]	STAR [12]	DOVE-5B [13]	MetaVSR allocation **Total Rate** (Video / Meta)	MetaVSR-2B (ours)
UDM10	PSNR ↑	**2687** (2687 / 0)	24.13	21.72	24.23	23.47	26.48	**2238** (2097 / 142)	28.17
	SSIM ↑		0.6801	0.5913	0.6957	0.6804	0.7827		0.8301
	LPIPS ↓		0.3908	0.4116	0.3272	0.4242	0.2696		0.2236
	DISTS ↓		0.2067	0.2230	0.1677	0.2156	0.1492		0.1198
	CLIP-IQA ↑		0.3494	0.4697	0.4557	0.2417	0.5107		0.4061
SPMCS	PSNR ↑	**1470** (1470 / 0)	22.17	18.81	22.39	21.24	23.11	**1377** (1147 / 230)	22.67
	SSIM ↑		0.5638	0.4113	0.5896	0.5441	0.6210		0.6868
	LPIPS ↓		0.3662	0.4468	0.3263	0.5257	0.2888		0.2454
	DISTS ↓		0.2164	0.2452	0.1960	0.2872	0.1713		0.1502
	CLIP-IQA ↑		0.3513	0.5248	0.4348	0.2646	0.5690		0.4682
YouHQ40	PSNR ↑	**4413** (4413 / 0)	22.39	19.62	23.17	22.64	24.30	**4235** (3445 / 790)	25.82
	SSIM ↑		0.5895	0.4824	0.6194	0.6323	0.6740		0.7506
	LPIPS ↓		0.4091	0.4268	0.3608	0.4600	0.2997		0.2441
	DISTS ↓		0.1933	0.2012	0.1685	0.2287	0.1477		0.1155
	CLIP-IQA ↑		0.3964	0.5258	0.4657	0.2739	0.4985		0.4277
RealVSR	PSNR ↑	**32178** (32178 / 0)	22.12	20.29	22.02	17.43	22.32	**25759** (25331 / 428)	22.83
	SSIM ↑		0.7163	0.5945	0.6774	0.5215	0.7301		0.7460
	LPIPS ↓		0.1870	0.2671	0.2182	0.2943	0.1851		0.1993
	DISTS ↓		0.0983	0.1425	0.1169	0.1599	0.0978		0.1017
	CLIP-IQA ↑		0.2905	0.4855	0.4510	0.3641	0.5207		0.4666
MVSR4×	PSNR ↑	**44854** (44854 / 0)	21.80	20.42	22.77	22.42	22.42	**36284** (35597 / 687)	24.29
	SSIM ↑		0.7045	0.6117	0.7418	0.7421	0.7523		0.8233
	LPIPS ↓		0.4235	0.4717	0.3568	0.4311	0.3476		0.2485
	DISTS ↓		0.2498	0.2673	0.2245	0.2714	0.2363		0.1672
	CLIP-IQA ↑		0.4118	0.6106	0.3769	0.2674	0.5453		0.4796
MetaSPS	PSNR ↑	**185** (185 / 0)	25.32	23.52	25.55	23.91	26.73	**179** (91 / 87)	26.83
	SSIM ↑		0.8175	0.7312	0.7959	0.7844	0.8610		0.8654
	LPIPS ↓		0.2345	0.2946	0.2415	0.2327	0.1568		0.1589
	DISTS ↓		0.1301	0.1387	0.1263	0.1067	0.0915		0.0910
	CLIP-IQA ↑		0.4407	0.4329	0.4958	0.3234	0.3776		0.3708

Note: Best and second-best metric values are highlighted in red and blue, respectively. Allocation cells report total size on the first line and (video / metadata) on the second, in 10^3 B. Rates are rounded independently to the nearest integer, so displayed components may differ from their total. No-meta allocations assign zero bytes to metadata. These operating points are not exact matched-rate pairs.

can be accommodated by the bitrate saved from the compressed low-quality video stream across all evaluated datasets, rather than being treated as free auxiliary information; detailed rate allocations are reported in Appendix Table 4. We also provide representative qualitative comparisons in Fig. 4, where MetaVSR produces significantly more accurate object contours and better visual quality than all other baselines on MetaSPS. Additional light-noise and heavy-noise qualitative sweeps under varying LQ compression and Canny edge metadata selection rates are provided in Appendix Figs. 9 and 8.

Table 1 shows that MetaVSR achieves strong reconstruction quality across both distortion and perceptual metrics. We defer additional experimental results, implementation notes, and dataset details to the appendix. Compared with DOVE-5B, the most directly related one-step generative VSR baseline, MetaVSR improves PSNR by 1.69 dB on UDM10, 1.52 dB on YouHQ40, 0.51 dB on RealVSR, and 1.87 dB on MVSR4×, while using a smaller 2B backbone. Averaged over the five public datasets, MetaVSR improves PSNR by approximately 1.03 dB and SSIM by 0.055 over DOVE-5B. Although DOVE-5B obtains slightly higher PSNR on SPMCS, MetaVSR still achieves better SSIM, LPIPS, and DISTS, indicating stronger structural and perceptual reconstruction quality. The improvements are particularly consistent on perceptual metrics. MetaVSR reduces LPIPS and DISTS over DOVE-5B on four out of five public datasets, with especially large gains on MVSR4×, where LPIPS decreases from 0.3476 to 0.2485 and DISTS decreases from 0.2363 to 0.1672. These

results suggest that the transmitted metadata provides compact but task-relevant structural cues, helping the receiver recover object boundaries, high-frequency details, and perceptually meaningful textures that are difficult to infer from the compressed video stream alone. MetaVSR does not always achieve the highest CLIP-IQA score, which is reasonable because CLIP-IQA is a no-reference metric that may favor visually sharp or semantically plausible generations over faithful reconstruction.

4.2 Rate–Distortion Evaluation under Matched Total Bitstream Budget

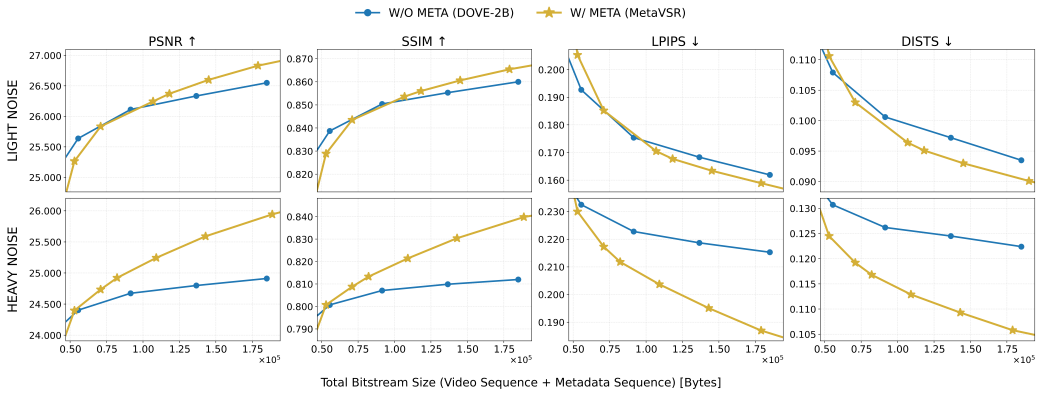


Figure 5: Rate–distortion (R–D) curves under light-noise (LN) and heavy-noise (HN) degradations, showing the effect of Canny edge metadata under matched total transmission budgets.

To further evaluate the bitrate efficiency of metadata-guided reconstruction, we conduct a rate–distortion evaluation on MetaSPS under light-noise and heavy-noise degradations. Unlike the fixed-budget comparison, this experiment explicitly varies the total transmission budget and reports complete rate–distortion curves. In Fig. 5, the horizontal axis denotes the total bitstream size, where the transmitted metadata is included in the overall budget for MetaVSR. We compare two settings: W/O META (DOVE-2B) and W/ META (MetaVSR). W/O META denotes our separately trained DOVE-style CogVideoX-2B no-metadata checkpoint (abbreviated DOVE-2B), while W/ META denotes our MetaVSR model, where compact Canny edge metadata is transmitted together with the compressed video stream. We evaluate PSNR and SSIM for distortion-oriented fidelity, and LPIPS and DISTS for perceptual quality. The two rows in Fig. 5 correspond to the light-noise and heavy-noise degradation settings. In this R–D evaluation, base-video compression is varied within each noise setting to sweep the transmission budget, rather than fixed to the conditioning comparison.

The R–D curves show favorable regions for metadata-assisted reconstruction under both degradation settings. Across most bitrate settings, MetaVSR obtains comparable or higher PSNR/SSIM while reducing LPIPS and DISTS, showing that even when metadata transmission cost is counted, allocating part of the bitrate to compact structural metadata can be more effective than using the entire budget for the video stream alone. The gain becomes more pronounced as degradation increases: in the light-noise setting, MetaVSR reaches a similar PSNR level with approximately 1.4×10^5 bytes, compared with nearly 1.8×10^5 bytes required by DOVE-2B, corresponding to about 20% rate saving; under heavy noise, MetaVSR reaches the best PSNR level of DOVE-2B using only about 0.8×10^5 bytes, compared with around 1.8×10^5 bytes, indicating more than 50% rate saving. These are approximate equal-quality savings read from the evaluated curves, rather than BD-rate estimates. The allocation benefit depends on the total budget and degradation; metadata is not beneficial at every operating point. Repeating the allocation study with only the base codec changed gives approximately 18% total-rate saving with HEVC/H.265 and 16% with VVC/H.266 at matched PSNR; metadata remains JBIG2-coded. Appendix Fig. 10 reports the R–D curves for H.264, HEVC/H.265, and VVC/H.266 under both LN and HN. A separate direct-HR VVC reference is reported in Appendix G.

4.3 Ablation Study on Edge Selection

Selection policy. To test whether prioritizing high-mismatch regions is effective, we compare forward, reverse, and random patch selection with the LQ video fixed as uncompressed. The forward policy selects patches with the largest LQ–HR mismatch (Sec. 3.3), the reverse policy selects the

lowest-mismatch patches first, and the random policy samples patches uniformly. We use identical selection ratios and plot PSNR, SSIM, LPIPS, and DISTS against the measured metadata sizes in Fig. 6. Forward selection generally performs best; random selection lies between forward and reverse on PSNR, LPIPS, and DISTS, with SSIM an exception. This comparison supports the tested mismatch-guided policy, rather than establishing its optimality.

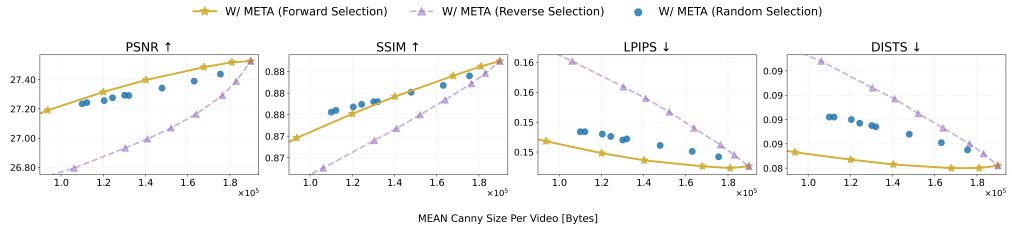


Figure 6: Forward, random, and reverse edge selection with fixed uncompressed LQ video, plotted against measured metadata sizes.

Sparse versus full-map metadata. To test whether the metadata budget is better spent on selected local edges than a rate-controlled full-frame edge map, we fix the base video and approximately match JBIG2-coded metadata sizes, with a maximum size mismatch of approximately 3.3% across the tested pairs. The full-map baseline uses Gaussian pre-smoothing before Canny extraction instead of mismatch-guided patch selection. Figure 7 shows that sparse selection improves all four metrics at the tested points: at 1%–10% selection, PSNR increases by 0.18–0.33 dB, SSIM by 0.0061–0.0111, and LPIPS/DISTS decrease by 0.0294–0.0427/0.0179–0.0250. At 20%, the PSNR difference is only 0.01 dB and does not establish statistically significant superiority. Pre-smoothing changes structural content as well as coded size, so these results support sparse selection against this specific full-map baseline, not every dense representation.

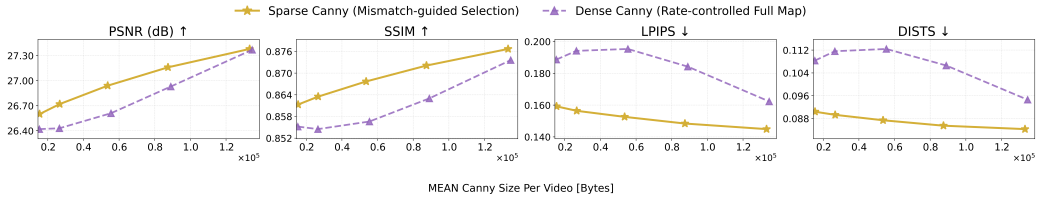


Figure 7: Sparse versus rate-controlled full-map Canny on a fixed base stream with approximately matched metadata sizes.

4.4 Controlled Conditioning Comparison

To compare different conditioning mechanisms, we evaluate receivers using the same LQ inputs and full Canny edge maps at two specific operating points, defined as follows:

Setting	Gaussian noise σ	Gaussian blur (k, σ)	Video stream	Metadata stream
Light degradation (LD)	10	$(5, 0.8)$	No compression	Full Canny edge map
Heavy degradation (HD)	20	$(7, 1.2)$	H.264, $\approx 10\%$ original size	Full Canny edge map

Within each setting, all receivers use the same LQ input, and all metadata-conditioned variants receive the same full Canny edge map. All variants in Table 2 use the same 2B initialization, training data, degradation pipeline, two-stage objectives, optimization schedule, and evaluation protocol. Cross-attention uses shared encoder/projection inputs and one-way rather than bidirectional joint attention. The T2I-Adapter-style branch has four factorized 3D residual blocks; the ControlNet-style implementation copies the first eight CogVideoX blocks and uses zero-convolution residual injection. Native fusion performs best among these tested implementations, gaining 0.98 dB PSNR under LD and 2.14 dB under HD over no-metadata DOVE-2B. Branch capacity and injection depth differ, so this is not a strictly parameter- or compute-matched comparison; counts and measured receiver timing are in Appendix E.

Table 2: Conditioning comparison on MetaSPS. All conditioned variants use full Canny input.

Receiver	Light degradation (LD)	Heavy degradation (HD)
	PSNR ↑ / SSIM ↑ / LPIPS ↓ / DISTS ↓	
DOVE-2B (no metadata)	26.55 / 0.8600 / 0.1619 / 0.0935	22.78 / 0.7552 / 0.2962 / 0.1558
Cross-attention	26.52 / 0.8563 / 0.1659 / 0.0922	22.96 / 0.7588 / 0.2891 / 0.1472
T2I-Adapter	26.64 / 0.8641 / 0.1526 / 0.0869	22.92 / 0.7622 / 0.2793 / 0.1435
ControlNet	26.79 / 0.8659 / 0.1656 / 0.1003	23.55 / 0.7848 / 0.2721 / 0.1590
MetaVSR (Ours)	27.53 / 0.8813 / 0.1426 / 0.0842	24.92 / 0.8240 / 0.2168 / 0.1179

Table 3: Metadata-robustness comparison on MetaSPS.

Condition	Light degradation (LD)
	PSNR ↑ / SSIM ↑ / LPIPS ↓ / DISTS ↓
DOVE-2B (no metadata)	26.55 / 0.8600 / 0.1619 / 0.0935
Clean metadata	27.53 / 0.8813 / 0.1426 / 0.0842
BAC (mild)	27.44 / 0.8767 / 0.1467 / 0.0866
BAC (severe)	27.38 / 0.8756 / 0.1471 / 0.0858
Shift (2 pixels)	27.28 / 0.8747 / 0.1476 / 0.0870
Shift (4 pixels)	27.26 / 0.8741 / 0.1483 / 0.0872
Random patch loss (20%)	27.44 / 0.8795 / 0.1442 / 0.0855
Random patch loss (50%)	27.29 / 0.8765 / 0.1470 / 0.0875

4.5 Robustness to Imperfect Metadata

To assess sensitivity to imperfect guidance, we perturb decoded metadata at inference under the LD operating point defined in Sec. 4.4. The sender has clean HR video and transmits its extracted edges with the LQ stream; extracting guidance from already degraded LQ video is outside this setting. Table 3 tests binary asymmetric channel (BAC) corruption, shifts of the full edge clip by 2 or 4 pixels in a random direction, and random loss of 20%/50% of the edge-selection-grid patches. Each patch-loss rate is averaged over five independent masks; patches are discarded randomly rather than by mismatch score. BAC independently flips each decoded edge-map pixel: $p_{0\to1}$ is the false-edge insertion probability and $p_{1\to0}$ is the true-edge deletion probability. Mild/severe settings use $(p_{0\to1}, p_{1\to0}) = (0.001, 0.05)$ and $(0.005, 0.20)$, respectively; this tests decoded-map corruption rather than errors in the JBIG2 or video bitstream.

All tested perturbations outperform our no-metadata DOVE-2B receiver on all four metrics. A 4-pixel shift retains gains of 0.71 dB PSNR and 0.0141 SSIM, with LPIPS/DISTS lower by 0.0136/0.0063. From the unrounded five-mask means, increasing patch loss from 20% to 50% reduces PSNR/SSIM by 0.1446 dB/0.0030 and increases LPIPS/DISTS by 0.0028/0.0020. At 50% loss, the gains over DOVE-2B remain 0.74 dB PSNR and 0.0165 SSIM, with LPIPS/DISTS lower by 0.0149/0.0060. These tests cover binary corruption, spatial misalignment, and random patch loss, not temporal misalignment, burst loss, retransmission, or protocol overhead.

5 Conclusions

We presented MetaVSR, a budgeted sender–receiver generative VSR framework examining whether a fixed transmission budget should go entirely to the compressed base stream or partly to compact structural metadata. MetaVSR encodes low-quality video and decoded metadata through the native 3D VAE and fuses them in the original DiT attention blocks, using sender-side edge metadata as a lightweight uncertainty-reduction signal. Rate–distortion experiments show that, with metadata cost explicitly counted, compact structural side information can improve reconstruction quality and bitrate efficiency over receiver-only generative VSR, especially under stronger degradation.

Scope and limitations. Validation is limited to Canny edges and the DOVE/CogVideoX family. Other modalities and backbones require adaptation; MetaVSR-5B, 4K reconstruction, and end-to-end commercial or real-time deployment were not evaluated. Sparse selection is supported only against the tested dense baseline, and metadata is not beneficial at every budget. Same-protocol learned-codec integrations were not compared.

Fallback and deployment. The base stream remains independently decodable, and MetaVSR can still operate without metadata by using an all-zero edge map. Under the LD conditioning setting, it achieves PSNR/SSIM/LPIPS/DISTS of 26.76/0.8644/0.1577/0.0920, comparable to 26.55/0.8600/0.1619/0.0935 for our separately trained no-metadata 2B receiver. The small differences reflect variation across training runs and checkpoints.Therefore, when metadata is unavailable, MetaVSR can simply use an all-zero edge map. For 9-frame 720p batch-1 FP16 inference on an A100-SXM4-80GB, MetaVSR takes 14.78 s/clip versus 10.78 s for DOVE-2B (about 37% overhead); both use 30.8 GB peak memory. These receiver timings exclude end-to-end deployment latency. Future work will explore model compression techniques, including quantization, knowledge distillation, and model pruning, to reduce deployment cost.

Funding

This work was supported by research funding from TCL.

Competing Interests

Other than the funding disclosed above and the affiliations listed in the author information, the authors declare no additional competing financial interests during the 36 months prior to submission.

References

[1] Consumer Technology Association. New 4k and 8k ultra hd television resources unveiled for integrators, retailers and sales associates to navigate features. `https://www.cta.tech/press-releases/new-4k-and-8k-ultra-hd-television-resources-unveiled-for-integrators-retailers-and-sales-associates-to-navigate-features`, August 2024. Accessed: 2026-04-28.

[2] Roku Developer. Streaming specifications. `https://developer.roku.com/docs/specs/media/streaming-specifications.md`, 2026. Accessed: 2026-04-28.

[3] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4472–4480, 2017.

[4] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3106–3115, 2019.

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[6] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. VRT: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. doi: 10.1109/TIP.2024.3372454.

[7] Xi Yang, Wangmeng Xiang, Hui Zeng, and Lei Zhang. Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4781–4790, 2021.

[8] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5962–5971, 2022.

[9] Ruohao Wang, Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chun-Mei Feng, Lei Zhang, and Wangmeng Zuo. Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1168–1177, 2023.

[10] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2535–2545, 2024.

[11] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European conference on computer vision*, pages 224–242. Springer, 2024.

[12] Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17108–17118, 2025.

[13] Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *Advances in Neural Information Processing Systems*, 2025.

[14] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[15] Wenhan Yang, Jiashi Feng, Jianchao Yang, Fang Zhao, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep edge guided recurrent residual learning for image super-resolution. *IEEE Transactions on Image Processing*, 2017. doi: 10.1109/TIP.2017.2750403. URL https://arxiv.org/abs/1604.08671.

[16] Kwanyoung Kim and Se Young Chun. Sredgenet: Edge enhanced single image super resolution using dense edge detection network and feature merge network. arXiv preprint, 2018. URL https://arxiv.org/abs/1812.07174.

[17] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Wang_Recovering_Realistic_Texture_CVPR_2018_paper.pdf.

[18] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Zhang_Image_Super-Resolution_by_Neural_Texture_Transfer_CVPR_2019_paper.pdf.

[19] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Yang_Learning_Texture_Transformer_Network_for_Image_Super-Resolution_CVPR_2020_paper.pdf.

[20] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. DiffBIR: Towards blind image restoration with generative diffusion prior, 2024. URL https://arxiv.org/abs/2308.15070.

[21] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. FullDiT: Video generative foundation models with multimodal control via full attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15737–15747, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ju_FullDiT_Video_Generative_Foundation_Models_with_Multimodal_Control_via_Full_ICCV_2025_paper.html.

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.pdf.

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2024.

[24] Jaehong Kim, Youngmok Jung, Hyunho Yeo, Juncheol Ye, and Dongsu Han. Neural-enhanced live streaming: Improving live video ingest via online learning. In *Proceedings of ACM SIGCOMM*, 2020. URL https://ina.kaist.ac.kr/projects/livenas/.

[25] Mehrdad Khani, Vibhaalakshmi Sivaraman, and Mohammad Alizadeh. Efficient video compression via content-adaptive super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Khani_Efficient_Video_Compression_via_Content-Adaptive_Super-Resolution_ICCV_2021_paper.html.

[26] Hadi Amirpour, M. Ghanbari, and Christian Timmerer. DeepStream: Video streaming enhancements using compressed deep neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (4):3786–3797, 2025. doi: 10.1109/TCSVT.2022.3229079. URL https://ieeexplore.ieee.org/document/9984695/.

[27] Gen Li, Jie Ji, Minghai Qin, Wei Niu, Bin Ren, Fatemeh Afghah, Linke Guo, and Xiaolong Ma. Towards high-quality and efficient video super-resolution via spatial-temporal data overfitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10259–10269, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_Towards_High-Quality_and_Efficient_Video_Super-Resolution_via_Spatial-Temporal_Data_Overfitting_CVPR_2023_paper.html.

[28] Eleonora Grassucci, Sergio Barbarossa, and Danilo Comminiello. Generative semantic communication: Diffusion models beyond bit recovery, 2023. URL https://arxiv.org/abs/2306.04321.

[29] Hang Yin, Li Qiao, Yu Ma, Shuo Sun, Kan Li, Zhen Gao, and Dusit Niyato. Generative video semantic communication via multimodal semantic fusion with large model. *IEEE Transactions on Vehicular Technology*, 75(1):1701–1706, 2026. doi: 10.1109/TVT.2025.3595688. URL https://pure.bit.edu.cn/en/publications/generative-video-semantic-communication-via-multimodal-semantic-f/.

[30] Francesco Pezone, Osman Musa, Giuseppe Caire, and Sergio Barbarossa. Semantic-preserving image coding based on conditional diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 13501–13505, 2024. URL https://iris.uniroma1.it/handle/11573/1726376.

[31] Lingdong Wang, Guan-Ming Su, Divya Kothandaraman, Tsung-Wei Huang, Mohammad Hajiesmaili, and Ramesh K. Sitaraman. Low-bitrate video compression through semantic-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Wang_Low-Bitrate_Video_Compression_through_Semantic-Conditioned_Diffusion_CVPRF_2026_paper.html.

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint, 2021. URL https://arxiv.org/abs/2010.02502.

[33] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.00927.

[34] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2302.04867.

[35] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint, 2022. URL https://arxiv.org/abs/2202.00512.

[36] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. arXiv preprint, 2023. URL https://arxiv.org/abs/2303.01469.

[37] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint, 2023. URL https://arxiv.org/abs/2310.04378.

[38] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2024.

[39] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017.

[40] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the h. 264/avc standard. *IEEE Transactions on circuits and systems for video technology*, 17(9):1103–1120, 2007.

[41] International Telecommunication Union (ITU-T). Itu-t recommendation t.88: Information technology – lossy/lossless coding of bi-level images (jbig2). Standard, 2000. URL https://www.itu.int/rec/dologin_pub.asp?id=T-REC-T.88-200002-S!!PDF-E&lang=e&type=items.

[42] Zhenwei Zhang, Fanhua Shang, Hongying Liu, Liang Wan, Wei Feng, and Yanming Hui. QBasicVSR: Temporal awareness adaptation quantization for video super-resolution. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4449. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c17f3035898a7044194fe45c5df6c6e8-Abstract-Conference.html.

[43] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[44] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[46] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[47] Zujin Guo, Size Wu, Zhongang Cai, Wei Li, and Chen Change Loy. Controllable human-centric keyframe interpolation with generative prior. *arXiv preprint arXiv:2506.03119*, 2025.

[48] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 4210–4220, 2023.

[49] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025.

[50] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

A Additional Implementation and Compute Details

Compute resources. All experiments were implemented in PyTorch and conducted on four NVIDIA A100 GPUs. Model fine-tuning used fp16 mixed precision, DeepSpeed ZeRO-2, and gradient checkpointing, with per-GPU batch size 1 and gradient accumulation 1 for an effective batch size of 4. Evaluation was conducted on the same GPU platform. The main hardware and runtime settings are summarized below for reproducibility.

- GPU memory: 80GB A100
- GPU configuration: 4 $\times$ NVIDIA A100, per-GPU batch $= 1$, gradient accumulation $= 1$, effective batch $= 4$
- Mixed precision / distributed training: fp16 with DeepSpeed ZeRO-2
- Training time for MetaVSR: 4 days
- Metadata extraction time: ≈ 6 s/clip (32f, 1280×720, CPU); edge extraction + JBIG2 compression.

Training and evaluation details. We summarize the main training and evaluation settings used for MetaVSR and the no-metadata baseline, including backbone, optimization, input configuration, degradation pipeline, evaluation metrics, and the dataset/checkpoint conventions used in Table 1.

- Backbone: CogVideoX-2B fine-tuned in two stages (Stage 1 pre-training + Stage 2 supervised fine-tuning (SFT))
- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$)
- Learning rate: Stage 1 uses 2×10^{-5}; Stage 2 uses 5×10^{-6}
- Batch size: per-GPU batch $= 1$, gradient accumulation $= 1$, effective batch $= 4$
- Training steps / epochs: Stage 1 runs for 7,500 steps and Stage 2 runs for 1,000 steps;
- Weight decay: 10^{-4}
- Gradient clip: max grad norm $= 0.1$
- Scheduler: `constant_with_warmup` with 100 linear warmup steps
- Input frames: Stage 1 uses 25 frames at 320×640 and 17 frames at 720×1280, following the resolution schedule below; Stage 2 uses 2 frames
- Resolution during training: Stage 1 video schedule is $25 \times 320 \times 640 \rightarrow 17 \times 720 \times 1280$ at steps 0 / 5000; Stage 2 uses video resolution $2\times320\times640$ and image resolution $1\times320\times640$ with `image_ratio`$= 0.8$
- The 25-frame 320×640 and 17-frame 720×1280 settings are separate training configurations; fewer frames at higher spatial resolution fit the GPU memory/activation budget. LQ and HR clips have identical frame counts, with no temporal subsampling or interpolation.
- Canny edge detector parameters: thresholds 50/150 with Gaussian kernel 5×5 and $\sigma = 1.4$
- Training-time data augmentation pipeline: Real-ESRGAN-style [43] pipeline defined in `configs/degradation_image_video.yaml` (random blur + resize + Gaussian/Poisson noise + JPEG + MPEG, followed by a second-order degradation with sinc filtering)
- MetaSPS evaluation degradation purpose: simulate realistic "capture-side" degradation layered atop bitrate-controlled compression artifacts.
- MetaSPS light-noise / heavy-noise settings: per-pixel grayscale Gaussian additive noise with $\sigma \in \{10, 20\}$ for LN / HN, respectively, clipped to the valid dynamic range; separable Gaussian blur ("ISO blur") with $(k, \sigma) = (5, 0.8)$ for LN and $(7, 1.2)$ for HN, where k is the kernel size in pixels.
- Metrics implementation: PSNR, SSIM, LPIPS, DISTS, CLIP-IQA

B Backbone Adaptation

For completeness, we summarize the DOVE-style backbone adaptation [13] objective used to fine-tune the CogVideoX receiver. In the first stage, the high-resolution target x_{hr} is encoded into the latent space as $z_{\mathrm{hr}} = \mathcal{E}(x_{\mathrm{hr}})$. The denoised latent z_{sr} from Eq. (1) is optimized to match the target latent by minimizing a normalized latent reconstruction loss:

$$\mathcal{L}_{\mathrm{s1}} = \frac{1}{\|z_{\mathrm{hr}}\|_2^2} \left\|z_{\mathrm{sr}} - z_{\mathrm{hr}}\right\|_2^2. \tag{8}$$

During this stage, the VAE remains frozen, and only the DiT denoiser is updated.

In the second stage, we further optimize the model in the pixel domain to improve visual fidelity and recover fine-scale details. For image samples, the pixel-space objective is defined as

$$\mathcal{L}_{\mathrm{s2,img}} = \|\hat{x}_{\mathrm{sr}} - \hat{x}_{\mathrm{hr}}\|_2^2 + \lambda_1 \, \mathcal{L}_{\mathrm{per}}(\hat{x}_{\mathrm{sr}}, \hat{x}_{\mathrm{hr}}), \tag{9}$$

Here $\hat{x}_{\mathrm{sr}}$ and $\hat{x}_{\mathrm{hr}}$ denote single-frame reconstructed and target images; unhatted $x_{\mathrm{sr}}, x_{\mathrm{hr}}$ below denote video sequences. $\mathcal{L}_{\mathrm{per}}$ is perceptual loss, and λ_1, λ_2 are its and the temporal loss's weights. n is the number of video frames, i their index, and Δ the adjacent-frame difference. For video sequences, we additionally impose a temporal consistency term based on frame differences:

$$\mathcal{L}_{\mathrm{s2,vid}} = \|x_{\mathrm{sr}} - x_{\mathrm{hr}}\|_2^2 + \lambda_1 \, \mathcal{L}_{\mathrm{per}}(x_{\mathrm{sr}}, x_{\mathrm{hr}}) + \lambda_2 \, \mathcal{L}_{\mathrm{frame}}(x_{\mathrm{sr}}, x_{\mathrm{hr}}), \tag{10}$$

$$\mathcal{L}_{\mathrm{frame}} = \frac{1}{n-1} \sum_{i=2}^{n} \left\|\Delta x_{\mathrm{sr}}^{(i)} - \Delta x_{\mathrm{hr}}^{(i)}\right\|_1, \quad \Delta x^{(i)} = x^{(i)} - x^{(i-1)}. \tag{11}$$

C Additional Visual Results

C.1 Visual Illustration of Reconstruction Quality

We complement quantitative benchmarks with qualitative visualizations that isolate two tightly coupled controls in our pipeline: the bitrate of the low-quality video stream and the Canny edge metadata supplied alongside it. Figures 8 and 9 provide additional examples under heavy-noise and light-noise settings, respectively, covering different sports categories, LQ compression rates, and Canny edge metadata selection rates.

For consistency, all examples follow the same visualization format. Each subfigure is organized as a 2×4 grid. The top row shows the ground-truth frame, the bilinearly upsampled LQ input, the MetaVSR reconstruction, and the reconstruction with the transmitted Canny edge map overlaid in red. The bottom row provides a 2×2 zoom-in mosaic of four non-overlapping 90×160 px crops selected from dense edge regions. Within each subfigure, the same frame index and crop coordinates are reused across all compared views, enabling direct pixel-level inspection of structural fidelity, boundary sharpness, and metadata-guided restoration behavior.

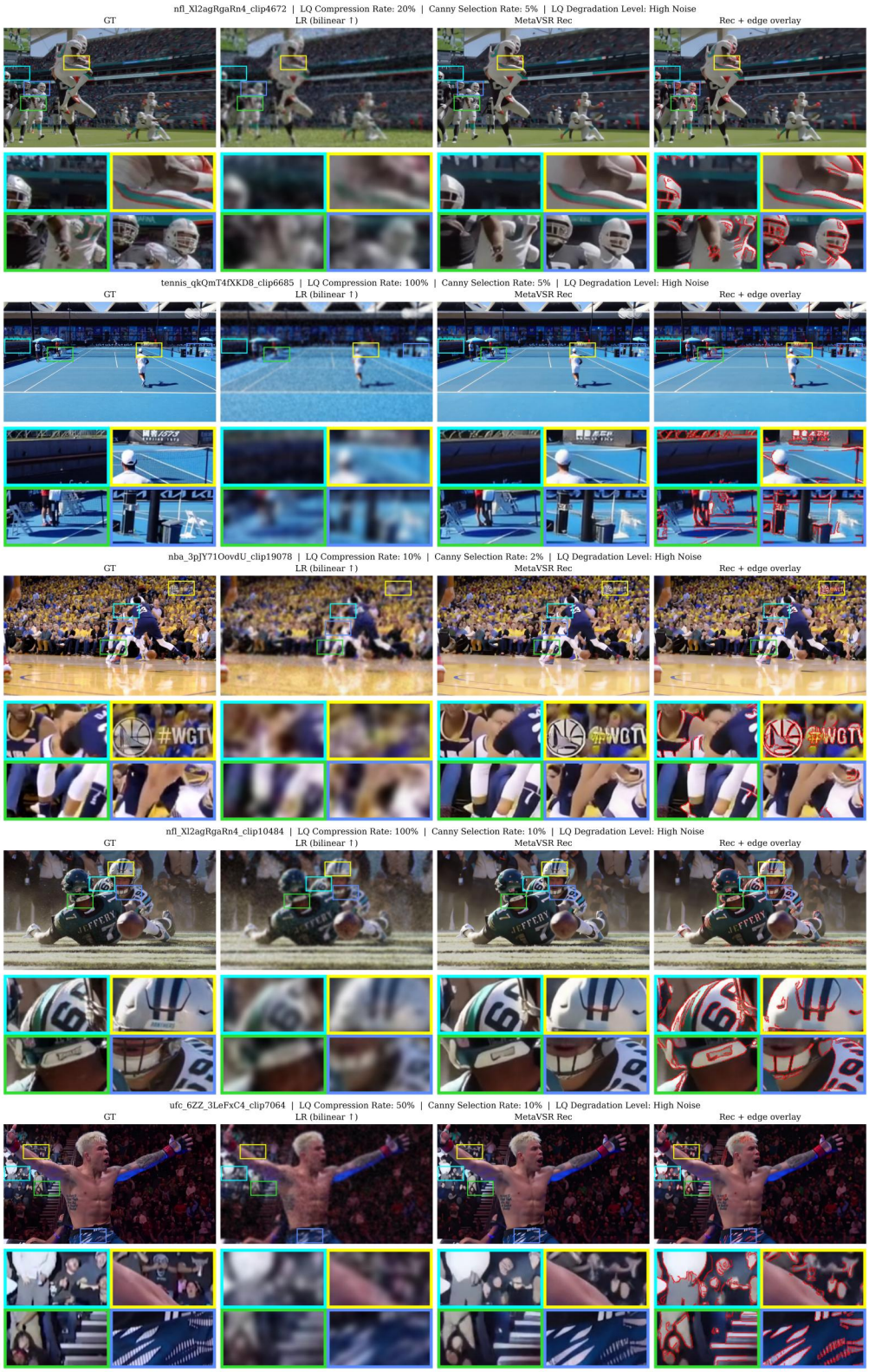


Figure 8: Additional MetaVSR qualitative results on MetaSPS under the heavy-noise setting. The examples cover different LQ compression rates and Canny edge metadata selection rates.

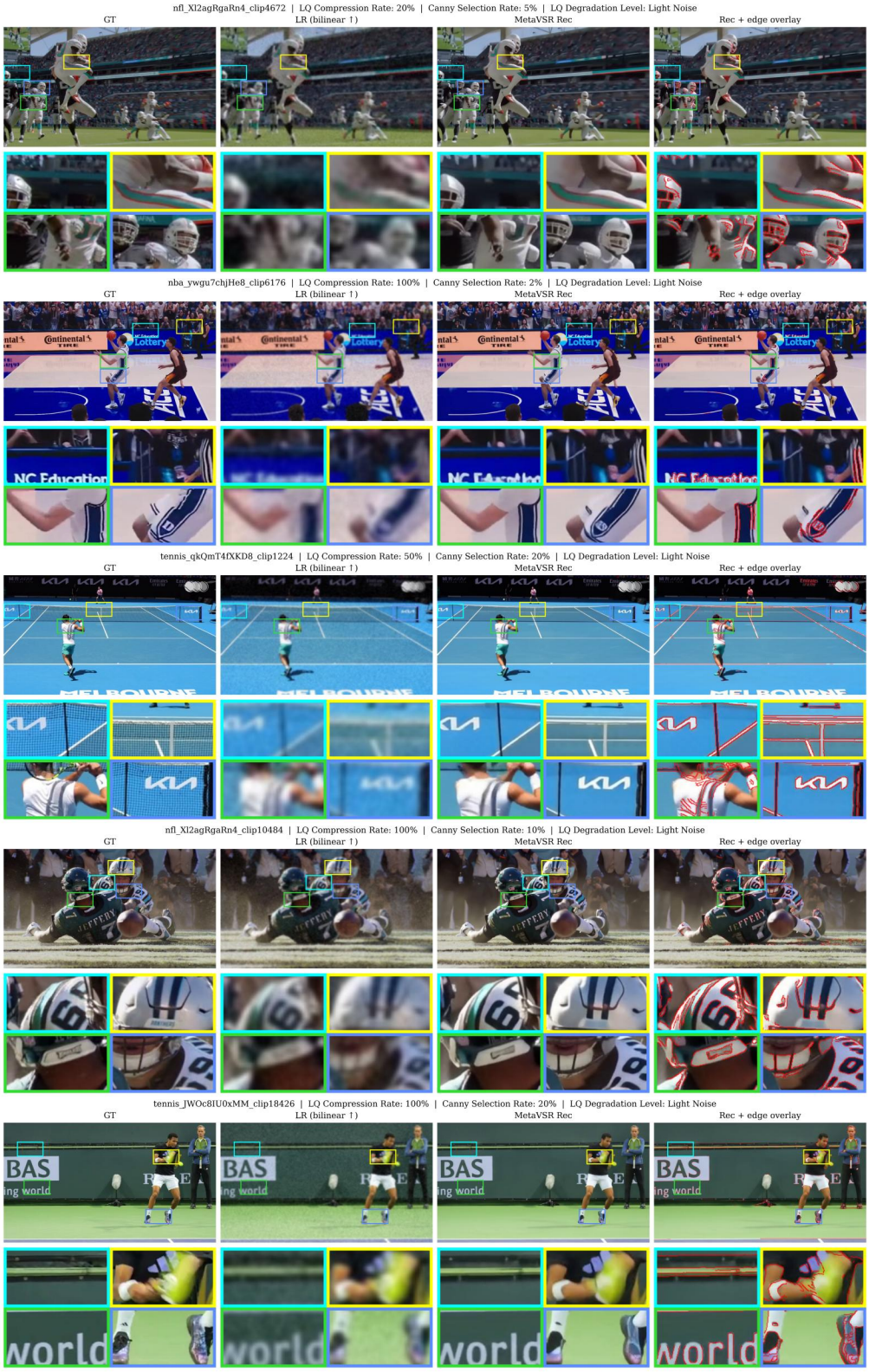


Figure 9: Additional MetaVSR qualitative results on MetaSPS under the light-noise setting. The examples cover different LQ compression rates and Canny edge metadata selection rates.

D Supplementary Materials for Baseline Comparison

To clarify the experimental protocol used for the results in Table 1, we emphasize the following setup choices:

1. For all public datasets used in Table 1, external receiver-only methods use the **released default LQ inputs**; MetaVSR uses the budgeted compressed stream plus coded metadata.
2. For all baseline methods in Table 1, external baseline results use **official public VSR checkpoints** without additional training. Our controlled DOVE-style 2B no-metadata checkpoint is separately trained, as distinguished in Table 5.

D.1 Sender-Side Bitrate Statistics

Table 4 summarizes the sender-side bitrate allocation under a shared transmission budget. Here, R_{ori} denotes the original low-quality video size, $R_{\text{ori}} - R_{\text{video}}$ denotes the bitrate saved by compressing the base stream, R_{meta} denotes the transmitted metadata size, and $R_{\text{total}}/R_{\text{ori}}$ reports the normalized total transmitted size after including both compressed video and metadata. Across all evaluated datasets, the transmitted metadata remains within the bitrate saved from the compressed base stream, confirming that the metadata channel is explicitly budgeted rather than treated as free auxiliary information.

Table 4: Sender-side bitrate statistics for the compressed base stream and transmitted metadata under a shared transmission budget.

Dataset	UDM10	YouHQ40	SPMCS	RealVSR	MVSR4×	MetaSPS
Original Size, R_{ori} (B)	2,687,136	4,412,644	1,470,328	32,178,207	44,854,070	184,909
$R_{\text{ori}} - R_{\text{video}}$ (B)	590,374	967,464	323,517	6,846,888	9,257,346	93,422
R_{meta} (B)	141,595	789,513	229,816	428,031	687,286	87,423
R_{total} / R_{ori} (%)	83.3%	96.0%	93.6%	80.1%	80.9%	96.8%

D.2 Rate–Distortion Curves across Base-Video Codecs

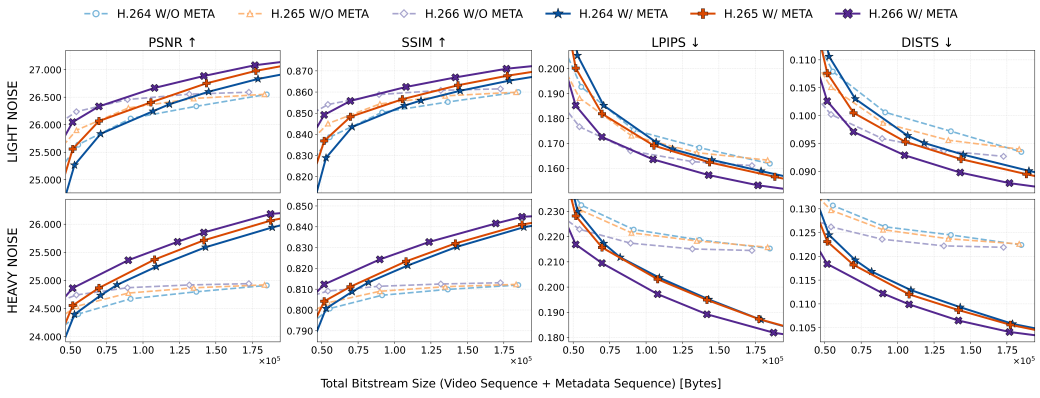


Figure 10: R–D curves on MetaSPS with H.264, HEVC/H.265, and VVC/H.266 base-video coding. Top: light-noise (LN); bottom: heavy-noise (HN). Columns report PSNR, SSIM, LPIPS, and DISTS. Dashed curves use the no-metadata DOVE-2B receiver; solid curves use MetaVSR with JBIG2-coded Canny metadata. The horizontal axis reports total transmitted size in bytes, including metadata where present.

To assess whether the metadata-allocation benefit is specific to H.264, Fig. 10 compares H.264, HEVC/H.265, and VVC/H.266 on MetaSPS under LN and HN. Only the base-video codec changes; the receiver models, test sequences, degradation settings, edge-selection procedure, and JBIG2 metadata coding remain unchanged. For each codec, W/O META denotes our no-metadata DOVE-2B receiver and W/ META denotes MetaVSR. The horizontal axis includes both video and metadata bytes for MetaVSR. At matched PSNR, the reported total-rate savings are approximately 18% with HEVC/H.265 and 16% with VVC/H.266. Thus allocating part of the shared budget to Canny metadata

remains beneficial with these stronger base codecs and is not specific to H.264. The benefit remains budget-dependent, rather than a uniform improvement at every operating point; these summaries are not BD-rate estimates.

E Receiver Provenance and Conditioning Capacity

The official DOVE-5B and our separately trained no-metadata 2B receiver are different checkpoints (Table 5). The latter and MetaVSR use the same CogVideoX-2B initialization and DOVE-style VSR adaptation. MetaVSR-5B was not trained; no scaling claim is made.

Table 5: Distinct model origins and experimental roles. Both controlled 2B models are VSR-adapted, not unadapted text-to-video checkpoints.

Model	Initialization	Adaptation / metadata	Experimental role
Official DOVE-5B	CogVideoX1.5-5B	Original two-stage DOVE VSR; no metadata	External official checkpoint in Table 1
Our no-metadata DOVE-2B	CogVideoX-2B	Author-trained DOVE-style VSR; no metadata	Controlled 2B receiver for ablation and R–D
MetaVSR-2B	CogVideoX-2B	Same VSR framework plus edge conditioning	Proposed metadata-assisted receiver

Conditioning capacity. T2I-Adapter and ControlNet freeze the 1.694B-parameter DOVE-2B backbone and train only their conditioning branches, with 7.394M and 487.5M trainable parameters, respectively. The frozen backbone remains active during inference, giving approximately 1.701B and 2.182B inference-active parameters. MetaVSR fine-tunes approximately 1.694B parameters and has the same inference-active count, without a separate control branch. Cross-attention adds less than 0.001B projection parameters at the reported precision. The comparison is not capacity-matched because trainable capacity, injection depth, and implementation differ; nevertheless, MetaVSR does not have the largest inference-active model, since ControlNet is larger.

Measured receiver cost. With 9-frame 720p clips, batch size 1, FP16, and one-step inference on an A100-SXM4-80GB, DOVE-2B and MetaVSR require 10.78 and 14.78 s/clip, respectively, with 30.8 GB peak memory for both. The longer metadata-conditioned token sequence adds approximately 37% latency. Cross-attention, T2I-Adapter, and ControlNet were not profiled under this runtime setting, so no latency comparison is made for those variants.

F MetaSPS Benchmark Role

MetaSPS is a targeted benchmark for asking whether sender-derived structural guidance remains useful after compression and whether its coded cost is justified under a shared video–metadata transmission budget. It complements conventional VSR benchmarks, which primarily evaluate reconstruction from paired HR/LQ videos, by providing aligned structural streams and controlled base-video compression, metadata selection, and coding. Its contribution is this evaluation setting and extraction/alignment pipeline, rather than dataset scale or a claim that no related benchmark exists.

Sports clips are suitable for this stress test because they combine rapid motion with small balls, object boundaries, field markings, scoreboard text, and dense crowd texture. MetaSPS contains 49 clips spanning basketball, American football, tennis, and combat sports, enabling evaluation of structural preservation and video–metadata budget allocation in these scenarios. We do not claim that MetaSPS is universally harder than other VSR datasets; equal normalized file-size reductions do not imply equal QP or distortion, and rate sensitivity depends on content.

The benchmark provides aligned Canny edges, depth, 2D skeletons, 3D poses, and ROI contours. Only Canny edges are experimentally validated in this work. The remaining streams are resources for future metadata-guided restoration studies, not evidence that the current checkpoint generalizes to other modalities.

G Direct-HR VVC Reference

This single low-rate reference changes the transmitted resolution and should not be conflated with the noisy LQ-based allocation experiments. Direct-HR coding preserves high-frequency information at sufficiently high rates; its unmeasured crossover is not claimed here. MetaVSR also retains an independently decodable LQ base stream. Same-protocol learned ultra-low-bitrate codecs have not been integrated.

Table 6: Illustrative direct-HR VVC comparison on the same clip, without added transmission noise. MetaVSR size includes video and metadata. No full-range superiority is claimed.

System	Total (B)	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Direct-HR VVC	14,967	23.6	0.78	0.32	0.16
LQ+VSR (MetaVSR)	14,922	24.3	0.81	0.26	0.14

H Dataset Detail & Metadata Extraction

Dataset release, provenance, and access. MetaSPS is constructed from publicly available YouTube sports videos uploaded by official sports-media channels, covering basketball, American football, tennis, and combat-sports content. For each clip, we maintain a source manifest containing the YouTube video ID, source channel, clip start/end time, selected frame indices, frame rate, resolution, split assignment, and all preprocessing parameters/code needed to reproduce the benchmark from legally accessible source videos.

The full MetaSPS benchmark was not publicly released during review. For the review stage, we provide representative MetaSPS samples, aligned metadata examples, and the metadata extraction pipeline in anonymized supplementary material. The planned release includes the metadata extraction pipeline, benchmark construction protocol, benchmark manifest, evaluation code, and derived metadata where permitted, together with documentation, license information, and persistent hosting. We will also release sample clips for which redistribution rights have been verified. For source videos whose redistribution is restricted, we will instead provide clip identifiers and reconstruction instructions consistent with the original license terms.

Because the source videos are hosted on YouTube and may be subject to source-license and platform restrictions, we do not generally redistribute raw RGB videos, decoded RGB frames, or processed RGB clip tensors, including `.mp4`, `.mkv`, `.png`, or RGB `.npy` versions of the clips, except for sample clips with verified redistribution rights. For the remaining sources, the released code, manifest, and reconstruction instructions are designed to reproduce the benchmark from legally obtained local copies of the source videos. The released metadata and code are intended for non-commercial research on metadata-guided video super-resolution, bitrate-aware video restoration, and aligned conditioning-signal evaluation. MetaSPS is not intended for commercial rebroadcasting, identity recognition, athlete surveillance, or misleading media manipulation.

H.1 Dataset Composition

We curate a 49-clip benchmark of action-rich sports footage spanning four categories: basketball (NBA), American football (NFL), tennis, and combat sports (UFC). The per-category clip counts are 15, 10, 12, and 12, respectively. Every clip is a 32-frame segment recorded at 1280×720 pixels and 8 frames per second, encoded with H.264 and distributed in both Matroska (`.mkv`) and MPEG-4 (`.mp4`) containers as redundant copies. The dataset therefore contains $49 \times 32 = 1{,}568$ unique frames in total. For metadata extraction the videos are decoded once into per-clip directories of lossless 24-bit PNGs; these PNGs are the canonical RGB inputs every extractor reads.

Representative benchmark samples and their aligned metadata are shown in Fig. 11.

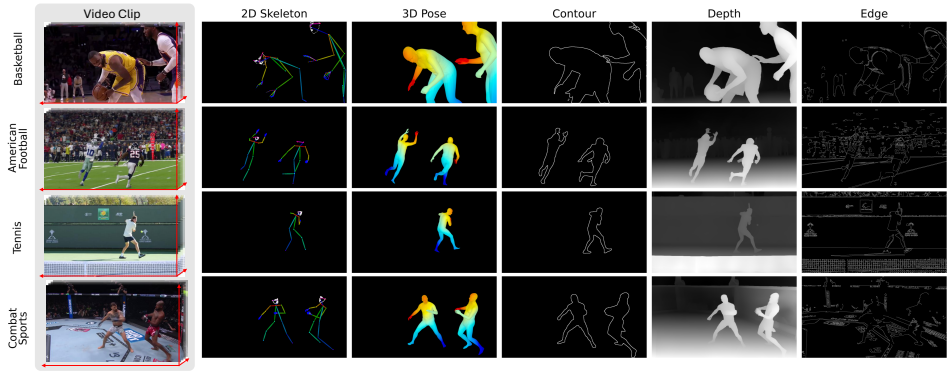


Figure 11: MetaSPS benchmark samples and aligned metadata visualizations.

H.2 Metadata Extraction Pipeline

For every frame we extract five aligned metadata streams: 2D whole-body skeletons, 3D body-surface pose renderings, sport-aware region-of-interest (ROI) contours, monocular depth maps, and Canny edge maps. SAM3-derived ROI masks are used as a hard prior for the 2D skeleton, 3D pose, and contour streams; the depth and edge maps are computed full-frame without ROI masking, so that background scene structure (court lines, ring posts, field markings) remains available as a conditioning signal for downstream models. All five streams are spatially aligned with the input frame at the original 720×1280 resolution. The overall extraction workflow is summarized in Fig. 12, which shows how each decoded RGB frame is routed through SAM3-guided ROI extraction, ROI-conditioned pose and contour generation, and full-frame depth and edge estimation.

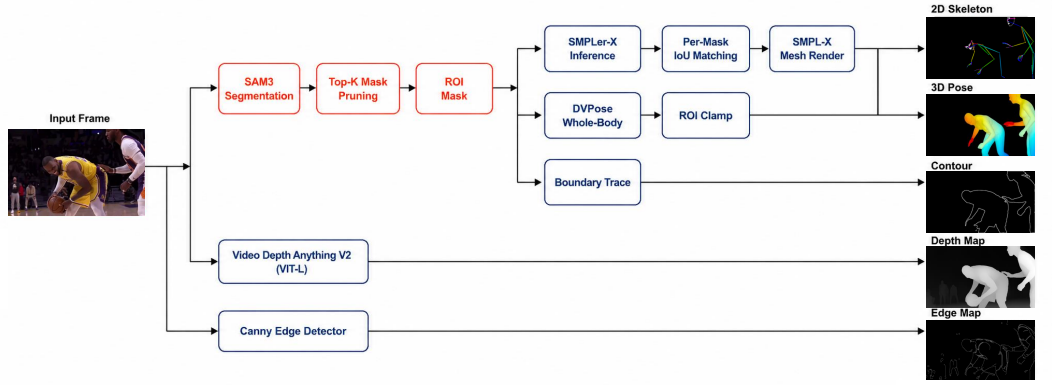


Figure 12: Overview of the metadata extraction pipeline used for MetaSPS. Each frame first passes through SAM3 to obtain sport-aware ROI masks, which are then used to guide SMPLer-X re-inference [46], DWPose-based 2D [48] skeleton rendering, SMPL-X surface rendering [45], and ROI contour extraction, while Video Depth Anything V2 and Canny edge extraction are applied to the full frame.

SAM3 segmentation with top-K pruning. For every frame we run SAM3 [44] with category-specific text prompts: "basketball player | basketball" (basketball), "player" (American football), "tennis player | tennis racket" (tennis), and "fighter" (combat sports). When two prompts are joined by "|" SAM3 is run in multi-prompt mode and we keep detections from either prompt. Detections with confidence below 0.75 are discarded. SAM3 generally over-segments in cluttered sports scenes (e.g. "player" returns the bench in football), so we apply a top-K filter that keeps only the K largest masks by foreground pixel count. Here, K is manually set after inspecting each video clip so that the retained segments correspond only to the foreground players with the heavy motion that we aim to model.

ROI-aware SMPLer-X re-inference on masked frames. SMPLer-X relies on a Faster R-CNN person detector that emits many false positives in dense sports scenes (referees, fans, partial-body detections); even when those false positives are filtered post-hoc by SAM3 IoU, the SMPL-X mesh regression has already burned compute and the resulting bbox list is hard to disambiguate. We instead apply the SAM3 ROI as a hard image-level prior *before* SMPLer-X is invoked. For each frame we compute the masked image $I_i^{\text{mask}} = I_i \odot M_i$, where M_i is the post-prune SAM3 union mask binarized to $\{0, 1\}$ broadcast across the three RGB channels and $\odot$ is element-wise multiplication. The masked PNGs are written to a staging directory and re-fed to SMPLer-X's whole-body inference (`smpler_x_h32`, ViT-Huge backbone) [46]. Because all non-athlete pixels are now identically zero, the bounding-box stage detects only the kept athletes, and SMPLer-X regresses a clean set of SMPL-X parameters (root pose, body pose, two hand poses, jaw and eye poses, shape betas, expression, translation) for each. The output annotation directory `smplx_annotations_roi/<clip>/` contains one `.npz` (mesh parameters) and one `.json` (per-detection camera intrinsics and bounding box) per athlete per frame.

2D whole-body skeletons (DWPose, ROI-filtered). We obtain 2D pose estimates with DWPose accessed through the `controlnet_aux` wrapper, which composes a YOLOX-Large person detector (mmdet 3.x) with an RTMPose-Large 384×288 whole-body pose estimator (mmpose 1.x) trained via the DWPose self-distillation recipe on COCO-WholeBody and UBody. Each frame yields up to 134 keypoints per detected person (17 body keypoints, plus a synthesized neck at the shoulder midpoint, plus 6 feet, 68 face, and 2×21 hand landmarks). Keypoints with confidence below 0.3 are discarded. We then assign each DWPose detection to the SAM3 person mask whose visible-keypoint bounding box has the highest IoU with the mask's bounding box; detections that fail to overlap any mask are dropped. For each matched detection we render the OpenPose-style skeleton (colored bone segments as thick line strokes, joint disks, hand chains, face point cloud) on a black canvas, then apply a pixel-level clamp using the corresponding SAM3 person mask so that any skeleton pixel outside the segment is zeroed. The frame's final 2D skeleton is the per-pixel maximum over the K ROI-clamped renders.

3D body-surface pose render (SMPL-X, IoU-matched). For each frame we render the colored SMPL-X surface for every athlete in the masked-frame annotation set from Step 2. We build a SMPL-X [45] `NEUTRAL` body model with face-contour vertices ($10{,}475$ vertices, $20{,}908$ triangular faces), apply the per-vertex color palette from `smplx_color.pt` introduced by PoseFuse3D-KI [47], forward the SMPL-X parameters on GPU to obtain world-space vertices, project them through the SMPLer-X-estimated camera (focal length and principal point read from the per-detection `.json`), and rasterize with pyrender's EGL backend at the original frame resolution.

To assign each SAM3 mask to the correct SMPLer-X detection, we perform a post-hoc IoU match against the raw mmdet bounding box. The `bbox_mmdet` field in each meta JSON is stored in (x, y, w, h) format; we convert to normalized (x_1, y_1, x_2, y_2) in $[0, 1]^2$ before computing IoU against the bounding box of the SAM3 mask ($\text{maskToBBox}(\cdot)$, normalized to the same coordinate frame). For each SAM3 mask we pick the unused SMPLer-X detection with the highest IoU; if no detection overlaps the mask at all we fall back to the nearest-center detection so that a SAM3 mask is never silently dropped. Each SMPLer-X detection is consumed at most once. The chosen per-athlete renders are merged into a single canvas by a non-zero overwrite. We deliberately do *not* apply a per-pixel SAM3 clamp to the rendered mesh, because the clamp truncates limbs (extended arms, kicking feet) that legitimately extend beyond the binary silhouette, and – in the multi-person composite – successive clamps would erase earlier athletes' contributions.

Sport-aware ROI contours. The contour metadata is the external boundary of the post-prune SAM3 masks. For each per-person mask we extract the boundary with `cv2.findContours(RETR_EXTERNAL, CHAIN_APPROX_NONE)` and rasterize it with `cv2.drawContours` at a stroke width of 2 pixels in white on a black canvas. The frame's contour image is the per-pixel maximum across the K per-person contour renders, producing a sparse structural cue that highlights human-and-equipment silhouettes without retaining background texture. When no per-person masks are available we fall back to the union mask. The output is a three-channel RGB PNG.

Monocular depth map (full-frame). Per-frame relative depth is estimated with Video Depth Anything V2 [49] using the ViT-Large encoder, an internal input size of 518 pixels, a maximum-resolution cap of 1280 pixels, and FP32 disabled (the default). Depth is computed over each clip's frame sequence in temporal order so that the model's temporal-consistency layers function as intended. The predicted relative-disparity output is normalized to the $[0, 255]$ range and saved as a single-channel 8-bit PNG at the original 720×1280 resolution. We do *not* apply the SAM3 ROI: full-scene depth preserves environmental geometry (depth gradients of the court, sideline, ring), which is useful both as a generative-model conditioning signal and as a benchmark resource.

Canny edge map (full-frame). Each RGB frame is converted to grayscale, smoothed by a Gaussian filter (kernel 5×5, σ=1.4), and processed with the Canny edge detector [50] using hysteresis thresholds T_1=50 and T_2=150. The output is a binary edge image at the original frame resolution. As with depth, no SAM3 ROI is applied here: the unfiltered Canny response retains background structural cues (court / field markings, ring posts, on-screen graphics) that downstream conditioning may exploit.

Pipeline orchestration and reproducibility. All extractors are deterministic given fixed input frames, model checkpoints, and the parameters listed above (SAM3 score threshold 0.75 and per-clip top-K; DWPose YOLOX-L / RTMPose-L checkpoints; SMPLer-X-H32 + Faster R-CNN COCO checkpoints; Video Depth Anything ViT-L; the Canny edge detector / Gaussian parameters listed above). The figure renderer and downstream training loaders read the same manifest to guarantee that what is shown in Fig. 11 is exactly what is consumed downstream.

I Broader Impacts

MetaVSR aims to improve bitrate-efficient video restoration for practical television and streaming scenarios. Its potential positive impacts include improving perceived video quality under bandwidth-limited delivery, reducing unnecessary transmission cost, and enabling better viewing experiences on high-resolution displays without requiring native high-resolution streams.

At the same time, improved generative video restoration may also raise risks if used to enhance misleading, manipulated, or copyright-sensitive video content. Because MetaVSR is designed for sender–receiver restoration rather than open-ended video generation, its intended use is limited to reconstructing videos with access to the original sender-side metadata. We will release the benchmark, code, and metadata pipeline with documentation, license information, and research-use guidance to encourage responsible use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly summarize the sender–receiver formulation, bitrate-constrained edge metadata, unified token-level conditioning, rate-aware edge selection, and MetaSPS benchmark. These claims are supported by the method, baseline comparison, RDO evaluation, and ablation study.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Sec. 5, noting that the current study focuses on Canny edge metadata. It also identifies richer metadata modalities, broader streaming evaluations, and more efficient receiver-side architectures as future directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results; metadata allocation is evaluated empirically.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The paper describes the architecture, training, rate accounting, and evaluation. It does not provide a complete per-experiment reproduction specification.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: During review, representative MetaSPS samples and aligned metadata were supplied in supplementary material; the full benchmark was not publicly released. The appendix describes the extraction pipeline and source-video redistribution restrictions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The shared training schedule and evaluation protocols are described in Appendix A and the experimental sections, rather than exhaustive configurations for each individual operating point.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: The paper reports quantitative results and five-mask averaging for random patch loss, but no confidence intervals, error bars, or significance tests. The 0.01 dB sparse/dense difference is not treated as statistically meaningful. Full repeated training is computationally expensive for large video diffusion models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the experiments compute resources in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work focuses on video restoration and metadata-guided reconstruction using public or standard research datasets. It does not involve private user data, human-subject experiments, or decision-making systems affecting individuals.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses broader impacts in Appendix I, including positive uses for bitrate-efficient video restoration and potential risks related to misleading video enhancement, authenticity, and copyright-sensitive content.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a general-purpose open-ended video generator or high-risk generative model. The proposed benchmark and metadata pipeline are designed for video restoration research.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: Existing datasets, models, and methods are credited. The appendix describes the benchmark source provenance and restrictions on redistribution of source videos.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The appendix documents MetaSPS categories, clip composition, resolution, frame rate, aligned metadata, and extraction pipeline. Representative examples accompanied the review supplement.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, subjective human evaluation, or research with human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human-subject experiments, crowdsourcing studies, or collection of participant data. Therefore, IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core method does not use LLMs as an important, original, or non-standard research component. Any use of LLMs for writing, editing, or formatting does not affect the methodology, experiments, or scientific claims.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.