

GauMVC: Generative Decoupled Gaussian Representation for Human-centric Multi-view Video Compression

Supplementary Material

Table 1. Perceptual quality comparison on *AvatarRex-lbn1*.

Method	BPP ↓	VIFp ↑	DSS ↑	LPIPS ↓	DISTS ↓	PieAPP ↓
MV-HEVC [1]	0.006	0.329	0.628	0.372	0.109	1.302
HUST-4DGS [4]	0.006	0.396	0.435	0.332	0.117	1.424
ADC-GS [3]	0.010	0.246	0.340	0.412	0.174	1.872
E-D3DGS [2]	0.197	0.325	0.505	0.347	0.115	1.495
Ours	0.001	0.398	<u>0.555</u>	0.286	0.103	0.858

Table 2. Perceptual quality comparison on *AvatarRex-lbn2*.

Method	BPP ↓	VIFp ↑	DSS ↑	LPIPS ↓	DISTS ↓	PieAPP ↓
MV-HEVC [1]	0.006	0.332	0.637	0.349	<u>0.082</u>	<u>1.326</u>
HUST-4DGS [4]	<u>0.006</u>	0.380	0.394	0.340	0.116	1.611
ADC-GS [3]	0.012	0.228	0.257	0.422	0.202	1.987
E-D3DGS [2]	0.198	0.320	0.436	0.354	0.116	1.697
Ours	0.001	0.389	<u>0.517</u>	0.280	0.079	0.913

Table 3. Perceptual quality comparison on *AvatarRex-zxc*.

Method	BPP ↓	VIFp ↑	DSS ↑	LPIPS ↓	DISTS ↓	PieAPP ↓
MV-HEVC [1]	0.007	0.311	0.622	0.347	0.078	1.356
HUST-4DGS [4]	<u>0.006</u>	<u>0.352</u>	0.403	0.357	0.133	1.515
ADC-GS [3]	0.009	0.203	0.284	0.447	0.197	1.923
E-D3DGS [2]	0.184	0.274	0.423	0.383	0.133	1.687
Ours	0.001	0.366	<u>0.475</u>	0.278	<u>0.082</u>	0.964

Table 4. Perceptual quality comparison on *AvatarRex-zzr*.

Method	BPP ↓	VIFp ↑	DSS ↑	LPIPS ↓	DISTS ↓	PieAPP ↓
MV-HEVC [1]	0.006	0.310	0.625	0.437	0.126	1.460
HUST-4DGS [4]	<u>0.006</u>	0.383	0.383	0.365	0.130	1.610
ADC-GS [3]	0.010	0.227	0.276	0.443	0.197	2.041
E-D3DGS [2]	0.204	0.313	0.448	0.376	0.123	1.690
Ours	0.001	0.405	<u>0.459</u>	0.242	0.088	0.992

1. Extended Perceptual Quality Comparison

As established in the main paper, perceptual metrics (VIFp, DSS, LPIPS, DISTS, PieAPP) serve as the primary performance indicators for our generative compression approach. Tables 1 through 4 provide an extended, per-dataset breakdown of these metrics alongside the effective bitrate (BPP). A core advantage of our framework is its efficiency, achieved by leveraging human semantic priors. Our method consistently achieves an extremely low bitrate of **BPP = 0.001** across all datasets. This represents a substantial improvement over competing methods: it is **6 to 10 times lower** than the next-best coding approaches (MV-HEVC and HUST-4DGS), and up to **200 times lower** than other generative approaches like E-D3DGS (BPP \approx 0.20). Despite this massive reduction in data volume, our framework

consistently records the best results in the majority of crucial perceptual metrics (e.g., lowest LPIPS, DISTS, and PieAPP).

Note that the specific metric values presented here may show minor deviations from the aggregated results in Figure 4 of the main paper, as these tables report the average metric over the entire video sequence, while Figure 4 selectively presented results from a single frame. This comprehensive multi-dataset evaluation reinforces the conclusion that our semantics-aware generative compression achieves superior perceived visual quality while operating at a fraction of the bitrate required by both traditional and generative competitors. Additional video demonstrations are available on our website.

2. Rate-Distortion Analysis

The total bitrate of our dynamic scene representation consists of three components: the static 3D Gaussian Splatting (GS) scene (RateGS), the human key-viewpoint images, and the human body parameters (e.g., SMPL). Among these, RateGS overwhelmingly dominates the overall bitrate. As a result, our compression strategy and rate-control behavior are primarily determined by how we optimize the parameters of the static GS scene.

2.1. R-D Curve Construction Methodology

To evaluate the full compression behavior beyond the single operating point presented in Section 4.1 of the main paper, we construct a complete Rate-Distortion (R-D) curve by systematically varying the structural pruning ratio τ . We examine four discrete pruning levels $\tau \in 0.05, 0.2, 0.3, 0.4$, where each value indicates the proportion of Gaussians removed from the original scene representation.

For all R-D points, we fix a baseline quantization configuration $\mathcal{B}_{\text{Base}}$ with bit-depth assignments of (16, 6, 4, 4, 8, 8) bits for Position, Scale, Rotation, Opacity, DC color, and higher-order spherical harmonics, respectively. After pruning, we perform an iterative refinement stage to re-optimize the remaining Gaussians. This optimization allows the surviving primitives to redistribute geometric and photometric attributes, compensating for the loss of representation capacity caused by sparsification. Through this process, each point on the R-D curve accurately reflects a stable trade-off between bitrate and reconstruction fidelity, enabling a rigorous assessment of compression behavior.

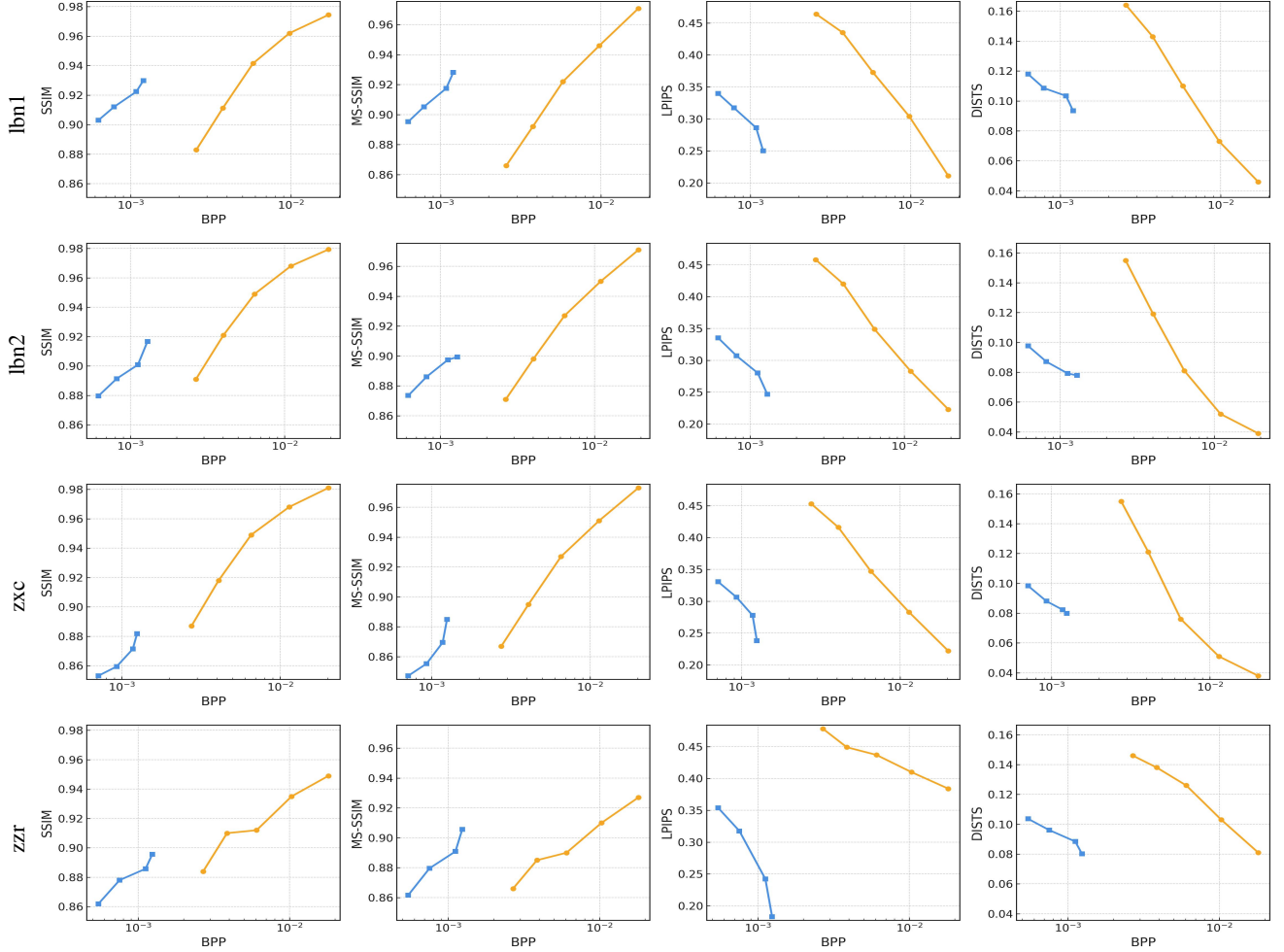


Figure 1. **Full Rate-Distortion (R-D) curves across four sequences of *AvatarRex*.** From left to right, the columns correspond to **SSIM** and **MS-SSIM** (higher is better), followed by **LPIPS** and **DISTS** (lower is better). Blue curves denote our method under different pruning ratios, while orange curves represent MV-HEVC evaluated at multiple CRF settings. The perceptual metrics (LPIPS and DISTS) highlight the clear low-bitrate advantage of our generative compression framework.

2.2. R-D Curve Analysis and Results

The complete Rate-Distortion (R-D) curves shown in Figure 1 provide a comprehensive characterization of the trade-off between bitrate and reconstruction fidelity. Our method produces four operating points (blue curves) by varying the structural pruning ratio $\tau \in \{0.05, 0.2, 0.3, 0.4\}$, while the MV-HEVC baseline (orange curves) is evaluated at CRF values $\{32, 36, 40, 44, 48\}$.

Across all datasets and evaluation metrics, our semantics-driven generative compression framework demonstrates consistently stronger R-D performance than MV-HEVC. The advantage is particularly pronounced on perceptual metrics such as LPIPS and DISTS, which emphasize high-level structural and texture fidelity. These metrics are highly sensitive to the over-smoothing and detail loss commonly observed in block-based codecs

at low bitrates. As illustrated in Figure 1, our method achieves substantially lower LPIPS and DISTS values even in bitrate regions where MV-HEVC exhibits clear degradation, indicating that our reconstruction preserves perceptual realism far more effectively under aggressive compression.

Most notably, for both LPIPS and DISTS, our method attains **comparable or clearly superior perceptual quality at bitrates that are often an order of magnitude ($10\times$) lower** than those required by MV-HEVC. This consistent margin highlights the strength of our human-centric, decoupled representation, which shifts compression away from low-level pixel redundancy toward high-level semantic priors, enabling robust perceptual reconstruction in the extreme low-bitrate regime.

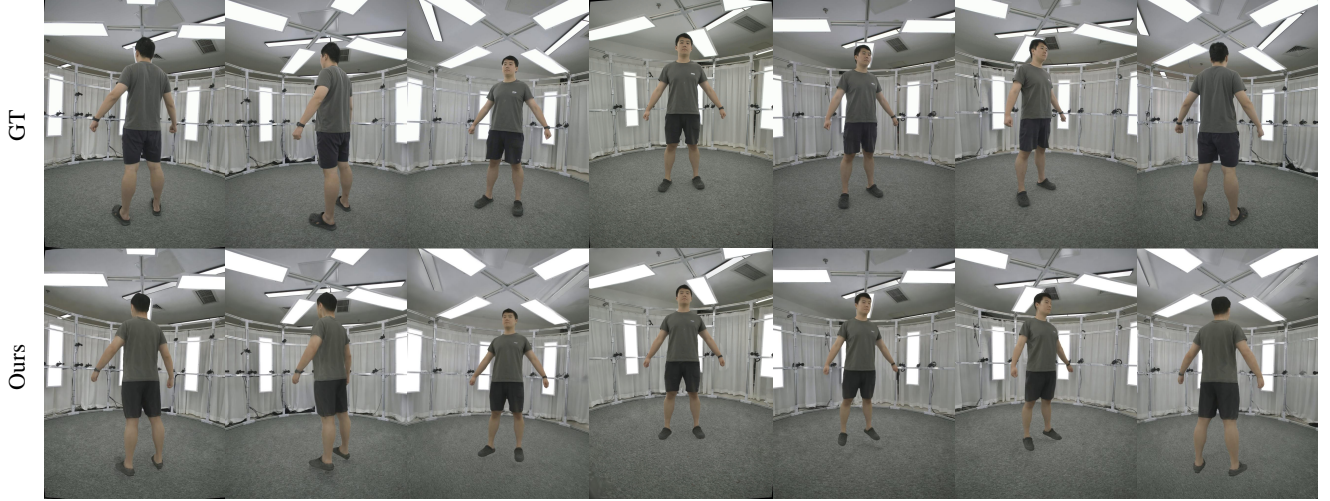


Figure 2. **Multi-View Reconstruction Fidelity.** Qualitative comparison of Ground Truth (GT, top) versus the reconstructed views (Ours, bottom) from our generative compression framework.

3. Multi-View Synthesis Qualitative Results

To evaluate the robustness and visual fidelity of our compression framework across the entire camera array, we present multi-view qualitative results reconstructed from the compressed representation. As shown in Figure 2 (top: GT, bottom: Ours), our method consistently reproduces high-quality appearances across a wide range of viewpoints.

The shown viewpoints cover frontal, side, and rear angles of the subject, representing a broad portion of the capture volume. Across all these perspectives, our method delivers uniform and high-fidelity reconstructions, indicating that only a small set of transmitted components (Background, Key Views, SMPL parameters) is sufficient for accurate multi-view rendering. These results validate the strength of our semantics-aware generative modeling pipeline for human-centric multi-view video compression.

4. Synthetic Dataset

Existing multi-view human datasets generally suffer from limited scale, insufficient scene diversity, and a lack of controllable capture conditions. These constraints hinder the development of models that require dense multi-view supervision or that must generalize across a wide range of geometric and photometric variations. To overcome these issues, we design a fully synthetic multi-view data generation pipeline that enables large-scale, diverse, and highly controllable video capture.

Our pipeline is implemented in the Unity 3D engine and provides fine-grained control over virtual humans, articulated motions, camera layouts, illumination, and background environments. A key design goal is to emulate realistic multi-camera capture systems while ensuring complete

spatial coverage. To this end, we place a total of 54 synchronized virtual cameras across three concentric elevation rings around the performer. These rings correspond to low-angle, eye-level, and high-angle viewpoints:

- **Low-angle ring:** Positioned below the subject, capturing upward-looking perspectives that highlight limb foreshortening and body proportion variations.
- **Eye-level ring:** Aligned with the performer’s mid-body height, providing canonical front, side, and oblique views commonly used in real-world capture studios.
- **High-angle ring:** Placed above the subject with a slight downward tilt, improving visibility of body articulation and reducing self-occlusion.

Each ring contains 18 cameras uniformly distributed along the azimuth, achieving continuous 360° coverage across multiple elevation levels. This layout ensures that the synthesized dataset reflects both practical studio configurations and challenging multi-view geometric conditions.

To enhance scene variability, we integrate a large collection of indoor and outdoor environments with diverse lighting conditions, material properties, and background structures. Automated batch rendering scripts handle asset loading, camera synchronization, and motion playback, enabling scalable data generation at high throughput. Furthermore, we apply a dedicated color-correction module that compensates for discrepancies between Unity’s linear rendering pipeline and the standard sRGB space, effectively mitigating overexposure, gamma inconsistencies, and unrealistic color responses. The resulting dataset provides photo-realistic multi-view sequences with consistent geometry and reliable correspondences, making it highly suitable for training and evaluating multi-view generative and neural rendering models.

References

- [1] Hvc test model (htm) reference software. <https://vcgit.hhi.fraunhofer.de/jvet/HTM>. 1
- [2] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 321–335, 2024. 1
- [3] He Huang, Qi Yang, Mufan Liu, Yiling Xu, and Zhu Li. Adcgs: Anchor-driven deformable and compressed gaussian splatting for dynamic scene reconstruction. In *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, pages 1179–1187, 2025. 1
- [4] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 20310–20320, 2024. 1