SAGE: SPARSE-VIEW HUMAN COMPRESSION VIA SOURCE-ANCHORED GAUSSIAN RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human multiview compression must preserve a person's appearance across camera queries while transmitting only a few captured views. A generative prior can synthesize missing viewpoints, but its inferred appearance and geometry may deviate from the captured subject. When decoded camera images and generated views jointly supervise a reconstruction, the more numerous generated views can dominate optimization and weaken captured detail. We propose **SAGE**, **S**ource-**A**nchored **G**aussian **E**ncoding, a sparse-view human compression framework that codes selected images and geometric conditions, constructs a reusable Gaussian representation with a fixed prior, and renders later viewpoints from that representation. Its key mechanism is source anchoring: decoded captured views receive more frequent and stronger photometric supervision, while generated views extend coverage. A later reduction in generated-view weight further steers optimization toward observed appearance. On the evaluated DNA-Rendering captures, SAGE achieves the smallest average complete package and leads the compared methods on all four quality measures. On calibrated THuman2.0 scans, it improves foreground fidelity and perceptual quality over size-aligned MV-HEVC under a common Gaussian receiver. These results show how compact human-view transmission can exploit generated coverage while preserving the distinctive appearance evidence in coded observations across camera queries.

1 INTRODUCTION

Viewing a person from different angles is a central capability of immersive visual media. Supporting this freedom calls for a compact representation that can render requested viewpoints. Multiview coding reduces redundancy among camera images (Tech et al., 2016). Three-dimensional human reconstruction offers another route: from selected coded views, the receiver builds a reusable representation and renders later queries without receiving a new image each time (Figure 1). This shifts the challenge from coding every requested image to reconstructing the person faithfully from sparse input.

That scene-based route must still explain parts of the person unseen by the selected cameras. Generalizable human renderers and human-specific diffusion models can infer complementary views (Zheng et al., 2024; Jin et al., 2025), supplying hypotheses for missing regions. Captured views, by contrast, carry measured appearance. Those views must also be compressed for transmission, so low rates diminish even their captured detail. The decoder must use generated coverage without losing what survives in the coded observations.

The tension between generated coverage and captured detail becomes concrete during Gaussian optimization, where decoded observations and generated support update the same scene (Kerbl et al., 2023). These support views are correlated predictions from the sparse input, not independent captures. Under uniform sampling, forty support views receive five of every six updates alongside eight coded sources; a shared prediction error can therefore spread through the reconstruction. The challenge is to gain coverage from these hypotheses without letting their volume dilute the influence of captured observations.

We address this challenge with SAGE, a **sparse-view human compression framework** built around source-anchored Gaussian reconstruction. The sender codes selected views and conditions; the receiver combines decoded images with a fixed generative prior to build a reusable Gaussian human. Source anchoring gives captured observations a stronger role: sampling governs their up-

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

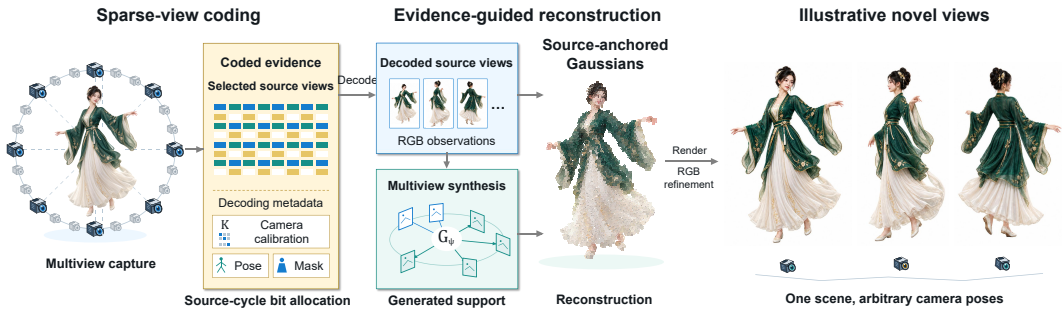


Figure 1: **From selected views to a reusable Gaussian human.** Coded RGB views and camera, pose, and mask conditions feed fixed multiview synthesis and source-anchored Gaussian reconstruction. Generated support extends coverage and enables optional RGB refinement of calibrated queries.

date frequency, and photometric weights set their influence. A decreasing weight on generated support further favors observed appearance in later updates. Source-cycle bit allocation complements this rule at the sender.

We first isolate the effect of source anchoring before evaluating complete compression systems. On DNA-Rendering, anchoring improves foreground-bounding-box (BBox) peak signal-to-noise ratio (PSNR) by 2.96 dB over uniform supervision. On THuman2.0, four calibrated scans gain 0.85 dB over size-aligned Multiview High Efficiency Video Coding (MV-HEVC) under a common Gaussian receiver, with lower perceptual distortion. Together, these comparisons separate scene supervision from source coding and connect the mechanism to low-rate human-view quality.

Our contributions are summarized as follows:

- We introduce **source anchoring**, a provenance-aware reconstruction principle that distinguishes transmitted observations from prior-generated hypotheses. Source-dependent sampling, photometric weighting, and scheduled support supervision implement it while preserving captured appearance in a reusable Gaussian scene.
- We develop **a complementary coding strategy** that uses source-cycle predictability to devote bits to captured views that neighboring observations recover poorly.
- We show that SAGE achieves the smallest complete packages and the best four-metric mean among the evaluated DNA-Rendering systems. On THuman2.0, it improves foreground fidelity and perceptual quality over size-aligned MV-HEVC with a common Gaussian receiver.

2 Related Work

Human novel-view synthesis. Neural radiance fields (NeRFs) model deforming humans through body and temporal conditioning (Mildenhall et al., 2020; Peng et al., 2021; Weng et al., 2022; Işık et al., 2023); Gaussians support animatable avatars and generalizable reconstruction (Kerbl et al., 2023; Hu et al., 2024; Zheng et al., 2024; Kwon et al., 2024). MVSGaussian combines multiview-stereo features with hybrid volume rendering (Liu et al., 2024). ReconFusion regularizes NeRFs using generated views (Wu et al., 2024), while ReconX applies confidence-aware Gaussian optimization after video generation (Liu et al., 2026). Diffuman4D supplies a human-specific multiview prior (Jin et al., 2025). SAGE uses provenance-aware supervision for rate-distorted observations and generated support, with fixed reconstruction budgets to isolate its effect.

Multiview and learned compression. MV-HEVC extends High Efficiency Video Coding (HEVC) across views (Tech et al., 2016); Versatile Video Coding (VVC) provides newer coding tools (Bross et al., 2021); and Moving Picture Experts Group (MPEG) Immersive Video codes texture–depth atlases (Boyce et al., 2021). Classical multiview allocation uses synthesized-view distortion to distribute bits (Velisavljević et al., 2011). For SAGE, generative leave-one-out predictability ranks a finite set of quantization-parameter (QP) allocations. Learned codecs optimize transforms, priors, and rate–perception objectives (Ballé et al., 2018; Blau & Michaeli, 2019; Mentzer et al.,

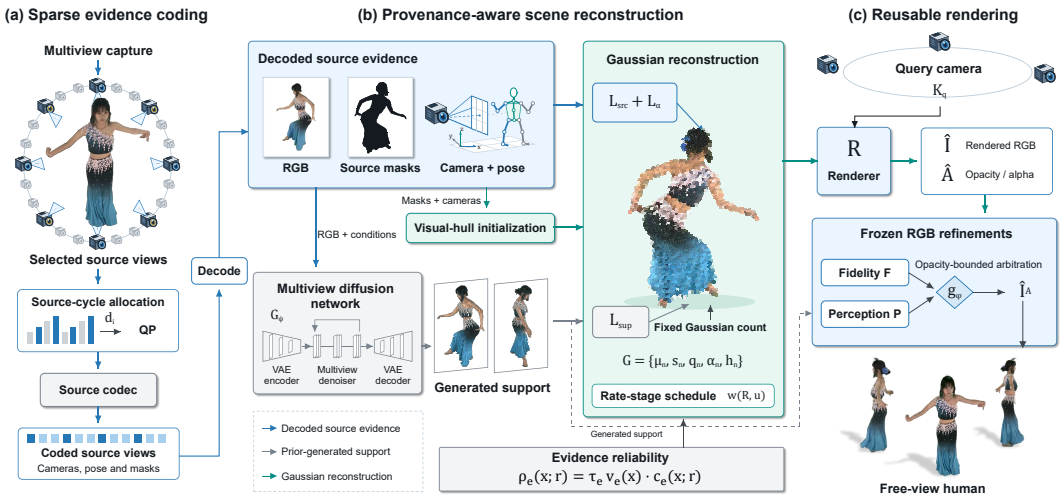


Figure 2: **SAGE framework.** (a) Selected coded views and geometric conditions provide source evidence. (b) Fixed multiview synthesis supplies support for Gaussian optimization; the reliability product summarizes evidence roles, and Gaussian count is fixed during terminal refinement. (c) The scene serves calibrated queries, with optional frozen, support-conditioned RGB refinement.

2020; Chen et al., 2026a;b), while video methods use motion–residual or conditional coding (Lu et al., 2019; Li et al., 2021).

Argus studies multiview streaming through sparse-view Gaussian reconstruction (Wang et al., 2026). We study iterative reconstruction from coded observations and generated support, separating scene preparation from rendering.

Gaussian compression. Compact Gaussian representations use structure, pruning, quantization, and entropy coding to reduce transmission cost (Lu et al., 2024; Lee et al., 2024; Niedermayr et al., 2024; Fan et al., 2024). Human-centric compression further exploits generative representations: HGC-Avatar uses hierarchical avatar coding (Tang et al., 2025), GauMVC studies generative decoupled Gaussian representations for multiview video (Yan et al., 2026), and Tang et al. (2026) investigate feedforward Gaussian generation for human-centric video compression. SAGE focuses on balancing decoded observations and generated support during iterative human reconstruction at the decoder. Our LightGaussian comparison separately examines self-contained scene transmission.

3 METHOD

Given selected coded views of a person and their geometric conditions, SAGE builds a Gaussian representation for later camera queries (Figure 2). The decoder recovers captured views from the payload and uses a fixed generative prior to synthesize views at other cameras. Source anchoring treats supervision according to its provenance: decoded observations transmit captured appearance, while prior-generated hypotheses provide complementary coverage. The framework connects *sparse evidence coding*, *provenance-aware reconstruction*, and *reusable rendering*. Our main results come from direct Gaussian rendering; we study image-space refinement separately.

In Figure 2, blue paths carry decoded observations, gray paths carry generated support, and teal paths indicate Gaussian reconstruction and rendering; snowflakes identify frozen networks. SAGE uses a fixed Diffuman4D prior (Jin et al., 2025) and anisotropic Gaussians (Kerbl et al., 2023).

Transmission, scene preparation, and repeated viewing are distinct stages. The sender transmits the selected views and conditions once. The receiver generates support on the requested camera grid and jointly optimizes one Gaussian scene against the generated and decoded images. Later Gaussian queries reuse that scene. Source anchoring therefore changes what the representation learns during scene preparation, not each rendered image independently.

3.1 TASK FORMULATION

For a synchronized capture with V cameras indexed by $\mathcal{V} = \{1, \dots, V\}$, the sender compresses a selected source set $\mathcal{S} \subset \mathcal{V}$ and the decoder reconstructs a disjoint target set $\mathcal{T}$. Payload b contains source images I_i, silhouettes M_i, and conditions $\mathcal{C}$:

$$b = \mathcal{E}_{\mathrm{src}}(\{I_i, M_i\}_{i\in\mathcal{S}}, \mathcal{C}; r)\,, \qquad R(b) = B_{\mathrm{RGB}} + B_{\mathrm{cond}} + B_{\mathrm{meta}} + B_{\mathrm{repr}} + B_{\mathrm{ctr}}, \tag{1}$$

where $\mathcal{E}_{\mathrm{src}}$ is the encoder, r its coding setting, and $R(b)$ the complete size in bytes. The five disjoint terms count coded color, conditions, metadata, scene attributes, and container overhead. The Gaussian scene is constructed at the decoder, so $B_{\mathrm{repr}} = 0$. Conditions $\mathcal{C} = \{K_j, P_j\}_{j\in\mathcal{V}}$ contain calibrated cameras K_j (intrinsics and extrinsics) and pose projections P_j. In the evaluated capture protocol, pose is triangulated from the calibrated rig, projected to the camera grid, and coded with the selected source silhouettes; the decoder initializes its hull from those silhouettes.

The decoder recovers source RGB I_i^{dec} and source silhouettes, then invokes the fixed generator G_ψ to obtain support $\tilde{I}_j$ on the prescribed camera grid. The parameters ψ belong to a variational autoencoder (VAE) and a multiview latent denoiser. Decoded observations retain captured appearance after compression; generated views provide complementary coverage. These roles motivate source-dependent supervision rather than treating all training images as interchangeable observations.

The evaluation utility Q comprises PSNR, the structural similarity index (SSIM) (Wang et al., 2004), learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), and deep image structure and texture similarity (DISTS) (Ding et al., 2022). Training uses the scalar objectives below; evaluation retains all four components of Q for the same output.

3.2 PROVENANCE-AWARE SCENE RECONSTRUCTION

To initialize geometry from transmitted observations, we use the visual hull of the decoded source silhouettes and their cameras. The scene $\mathcal{G} = \{\mu_n, s_n, q_n, \alpha_n, h_n\}_{n=1}^{N_G}$ contains N_G primitives with position, anisotropic scale, rotation quaternion, opacity, and spherical-harmonic appearance, respectively. Rendering produces RGB $\hat{I}_j$ and foreground alpha $\hat{A}_j$ at camera j:

$$\big(\hat{I}_j, \hat{A}_j\big) = \mathcal{R}(\mathcal{G}; K_j), \tag{2}$$

where $\mathcal{R}$ denotes the differentiable rasterizer, distinct from byte count R. Rendered alpha $\hat{A}_j$ describes ray occupancy, while observation type supplies source provenance. The query-camera index in K_q is separate from the primitive rotation quaternion q_n.

Initialization and appearance supervision serve different purposes. The source hull supplies an initial spatial extent consistent with the transmitted silhouettes. Optimization can then refine the primitive positions, shapes, opacity, and appearance to explain the images. Generated support contributes additional camera constraints, whereas decoded sources anchor those updates to captured detail. Keeping the initial state fixed in paired controls lets us examine this supervision choice independently of a more favorable starting geometry.

The scene is shared across cameras: an update from one view can alter renderings at many others. Consequently, a plausible but incorrect support texture can propagate beyond the view that introduced it. Generated views supply constraints where sources are sparse, while reconstruction must preserve the appearance measured by the source cameras. We balance these influences during construction of the shared representation, before query-specific rendering.

The scene must preserve observed appearance while explaining regions supplied by the prior. We therefore distinguish the photometric constraints from decoded sources and generated support:

$$\ell_{\mathrm{photo}}(X, Y) = 0.8\|X - Y\|_1 + 0.2[1 - \mathrm{SSIM}(X, Y)],$$
$$\mathcal{L}_{\mathrm{src}} = \mathbb{E}_{i\in\mathcal{S}}\ell_{\mathrm{photo}}(\hat{I}_i, I_i^{\mathrm{dec}}), \qquad \mathcal{L}_{\mathrm{sup}} = \mathbb{E}_{j\in\mathcal{U}}\ell_{\mathrm{photo}}(\hat{I}_j, \tilde{I}_j), \tag{3}$$

where $\mathcal{U}$ indexes the generated support views used for scene optimization. In the principal protocol, $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$; source-count controls retain a common support set while varying the sources. Expectations are uniform within each set. For RGB images X, Y, L_1 averages absolute error over the canvas; SSIM compares local windows.

3.3 Source and Generated Supervision Balance

The two sets of images are complementary, but their counts need not reflect their reliability. If every image is sampled equally, generated support can supply most updates even where decoded observations contain the more faithful appearance. Source anchoring separates update frequency from loss strength, allowing support to extend coverage without giving it the same influence as captured evidence.

Sampling and photometric influence. Each update samples one camera with source probability p_s. Photometric coefficients λ_s, λ_g then give

$$\mathbb{E}[\ell_{\text{photo}}^{(u)}] = p_s \lambda_s \mathcal{L}_{\text{src}} + (1 - p_s)\lambda_g \mathcal{L}_{\text{sup}}, \tag{4}$$

where u indexes updates. A source update compares the rendered view with its decoded RGB; a support update compares it with generated RGB. Repetition changes how often existing source slots are drawn, not the number of transmitted observations. Sampling and weighting jointly affect the update trajectory, including Gaussian growth and pruning. For the evaluated eight-source configuration, uniform sampling gives $p_s = 8/48$. Repeating each source four times gives $p_s = 32/72$; with $\lambda_s = 1.8$ and $\lambda_g = 1$, the expected source photometric coefficient share becomes

$$\pi_s = \frac{p_s \lambda_s}{p_s \lambda_s + (1 - p_s)\lambda_g} = \frac{0.8}{0.8 + 5/9} \simeq 0.5902. \tag{5}$$

The share describes expected photometric coefficients; gradient magnitudes also depend on image content and optimization. For $N_s = |\mathcal{S}|$ sources, $N_u = |\mathcal{U}|$ support views, and k source copies, $p_s = kN_s/(kN_s + N_u)$; the source-count study fixes $k = 4$, $N_u = 40$, and the photometric coefficients. In the eight-source setting, expected source and support coefficients are 0.8 and $5/9$, versus $1/6$ and $5/6$ under uniform supervision. The coefficients enter the sampled loss without normalization, preserving their scale relative to auxiliary losses. A coefficient-matched control tests whether sampling trajectory matters beyond expected supervision mass.

Scene optimization and support schedule. Photometry anchors appearance, while silhouettes constrain foreground occupancy and terminal opacity stabilization preserves the accumulated scene. These roles enter the sampled objective as

$$\ell^{(u)} = \lambda_{e_u} \ell_{\text{photo}}(\hat{I}_{e_u}, I_{e_u}^{\text{ref}}) + \eta_\alpha(u) \ell_{\text{silhouette}}(\hat{A}_{e_u}, M_{e_u}; c_{e_u}) + \eta_o(u) \ell_{\text{opacity}}^{(u)}, \tag{6}$$

where e_u is the sampled view, $I_{e_u}^{\text{ref}}$ its decoded or generated reference, and λ_{e_u} its photometric coefficient. Source masks are decoded; support masks are derived from the corresponding generated RGB. Fixed contour confidence c_{e_u} reduces the weight of uncertain generated boundaries. Schedules $\eta_\alpha(u)$ and $\eta_o(u)$ weight silhouette and opacity terms. The silhouette term balances binary cross-entropy over foreground, background, and contour regions. Importantly, rendered alpha describes foreground occupancy, not whether a surface was seen in a source image; provenance follows the reference used for that update.

Reconstruction uses 30,000 base updates for scene growth and 1,000 terminal updates with fixed Gaussian count. The base stage densifies and prunes against a white background, introducing silhouette supervision near its end. The terminal stage uses a black background and penalizes opacity drift from the base checkpoint. The constant-coefficient controls retain the weights above, whereas the main native-resolution rate study decreases generated photometric weight from 1 to 0.25 after the first 5,000 updates, reaching 0.25 by update 28,000. This shifts the scene from broad prior-supported coverage toward observed detail without discarding support. Figure 2 denotes the rate- and update-indexed coefficient tuple by $w(R, u)$; measured-rate variants select frozen settings. *Appendix A.2.3* gives the phase-specific losses and boundaries. Paired supervision studies keep the same initialization, support, population policy, and endpoint.

The same provenance principle also guides coding at the sender. A source-cycle prediction measures how much captured evidence is difficult to recover from the other selected views. Views that are harder to predict contain detail less readily recovered from their neighbors, so the coding search directs more bits to them. The measured package can thus devote more bits to distinctive captured appearance, making those observations more useful when they later anchor the shared Gaussian scene.

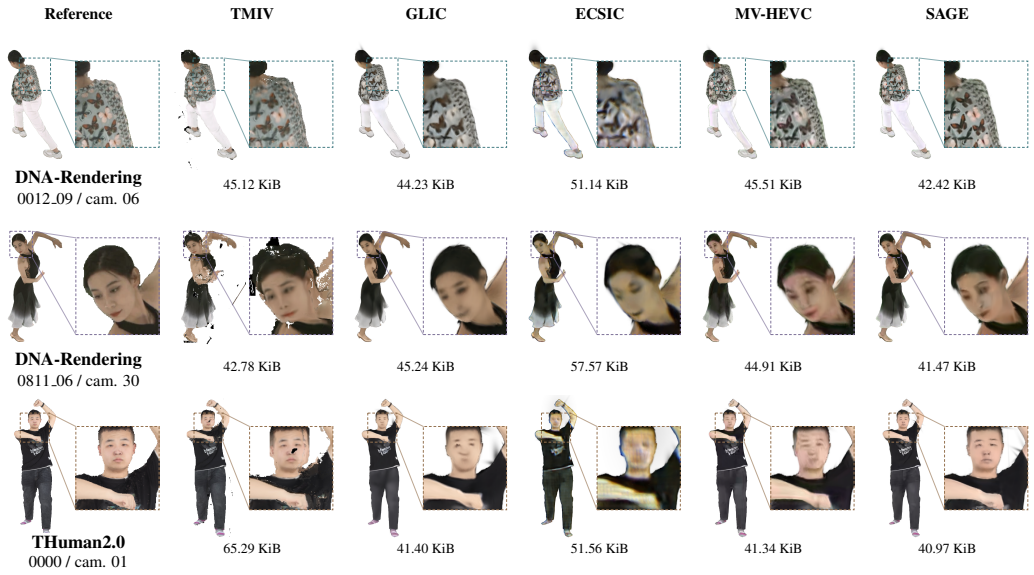


Figure 3: **Low-rate human reconstruction.** At matched viewpoints, SAGE retains facial and clothing detail using the smallest complete packages in each example.

For each source, the predictor withholds its RGB while retaining its camera and pose conditions. Within a fixed candidate set, less predictable views receive lower quantization parameters (QPs); the sender selects a candidate using measured complete bytes and decoded-source distortion. The selected streams are decoded afresh before support generation and scene construction. Thus, coding changes the evidence available to the decoder, whereas source anchoring controls its influence relative to generated support. *Appendix B.1.5* specifies the predictor and candidate search.

Scene construction is paid once per capture. After optimization, a Gaussian-only query consumes the shared state $\mathcal{G}$ and camera K_q in Equation (2), including for cameras outside the support grid. These direct renderings underlie the main comparisons. For a queried camera with generated support, we additionally study optional Full RGB refinement (Figure 2). This query-time operation changes the rendered image, not the Gaussian state, so subsequent Gaussian queries reuse the same reconstruction. Its inputs and cost appear in *Appendix B.4*.

4 Experiments

4.1 Experimental Setup

Data and evaluation. Controlled studies use three DNA-Rendering identities (Cheng et al., 2023); system tests use seven calibrated THuman2.0 scans (Yu et al., 2021). Each task codes eight RGB views and geometric conditions, with pose from the 48-camera rig, and scores forty targets. System metrics cover the black canvas; mechanism studies use a shared crop around the person (bounding box, BBox). Complete package sizes are reported in kibibytes (KiB). Higher PSNR/SSIM and lower LPIPS/DISTS indicate better quality. The former use 1024×1024 images and the latter 256×256 resizes.

Baselines. MV-HEVC codes the same eight sources. Epipolar Cross Attention for Stereo Image Compression (ECSIC) (Wödlinger et al., 2024) and Graph-based Learned Image Compression (GLIC) (Chen et al., 2026a) use frozen official weights; each decoded input generates its own support for the common Gaussian receiver. Main comparisons use direct Gaussian rendering. Query-time RGB refinement is evaluated separately as an optional output enhancement. The Test Model for Immersive Video (TMIV) uses its texture/silhouette/depth decoder. Figure 3 shows matched views, Figure 4 traces the measured size–quality settings, and Table 1 summarizes complete packages across identities.

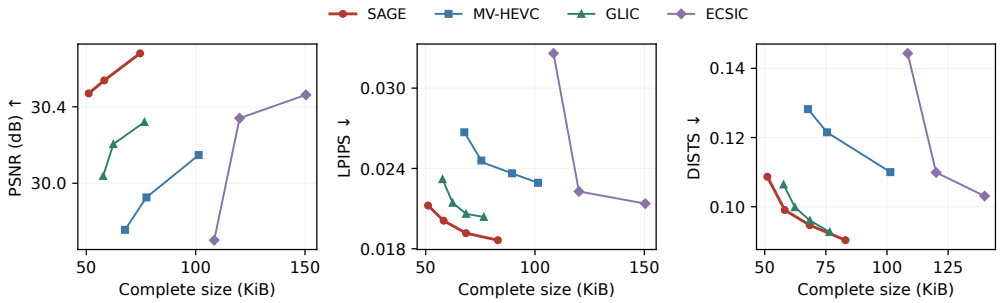


Figure 4: **Rate–distortion comparison.** At compact package sizes, SAGE combines higher PSNR with lower LPIPS and DISTS; curves trace metric-specific envelopes of measured Gaussian outputs.

Table 1: **Complete-package comparison.** Across three DNA-Rendering subjects, SAGE transmits the smallest payload and leads all four quality metrics. Non-learned describes the source coder; GS denotes the common Gaussian receiver.

Method	Front end	Non-learned	GS	KiB ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
TMIV	Texture/depth	✓	×	41.00	22.10	0.9434	0.0815	0.1641
ECSIC	Stereo	×	✓	54.63	28.96	0.9647	0.0476	0.1605
MV-HEVC	MV-HEVC	✓	✓	43.40	<u>30.44</u>	<u>0.9728</u>	<u>0.0194</u>	<u>0.1089</u>
GLIC	Image	×	✓	44.49	30.02	0.9714	0.0259	0.1299
SAGE	VVC	✓	✓	40.32	**30.53**	**0.9751**	**0.0185**	**0.1037**

4.2 Main Results

To test how coded detail survives scene construction, Figure 3 compares complete low-rate packages at matched cameras and crops. SAGE and MV-HEVC share the compact pose condition; MV-HEVC uses its nearest larger measured package. GLIC's 512-pixel sources and ECSIC's rectified half-resolution pairs are restored to the calibrated canvas before reconstruction.

Each row fixes the camera and crop across methods. Across four THuman2.0 scans, SAGE gains 0.85 dB in mean BBox PSNR over size-aligned MV-HEVC, with lower mean LPIPS and DISTS; the displayed scan is one of these four.

Table 1 averages forty target views within each of three DNA-Rendering identities, then weights the identities equally. It separates source coding from the Gaussian receiver shared by four methods; TMIV uses texture/depth. SAGE has the smallest package among Gaussian-receiver configurations on all three captures, and among all methods on two. Relative to MV-HEVC, it improves SSIM and LPIPS at fewer bytes throughout; DISTS improves on two captures. TMIV has the smallest package on *0345_08*, while SAGE leads its four quality metrics.

4.3 Rate–Distortion Analysis

A single operating point cannot show how quality changes when the transmitted package grows. We therefore sweep native-resolution coding settings on a common capture, keeping the forty target cameras, Gaussian-only output, and 31k reconstruction budget fixed. Figure 4 places pixel fidelity beside two perceptual metrics so their rate-dependent trends can be read together. The curves are metric-specific envelopes drawn from 26 measured settings (construction in *Appendix C*); numerical comparisons retain each setting's complete metric vector and actual transmitted bytes.

At near-equal sizes, SAGE (70,089 bytes) improves all four canvas and BBox metrics over GLIC (70,124 bytes), including 0.37 dB higher canvas PSNR. Beyond this pair, *Appendix C* integrates shared, locally invertible quality intervals and reports negative BD-rate estimates against MV-HEVC, GLIC, and ECSIC.

Table 2: **Source anchoring.** Person-crop scores on DNA-Rendering. Combining source repetition and weighting improves all four metrics over either change alone.

Supervision	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Uniform	21.93	0.8805	0.1050	0.1536
Repetition	24.21	0.8964	0.0839	0.1449
Weighting	22.29	0.8859	0.0988	0.1505
Source anchored	**24.89**	**0.9004**	**0.0734**	**0.1399**

Table 3: **Evidence trade-off.** Person-crop scores on THuman2.0. Anchoring improves on uniform supervision; source-only and support-mask variants favor different metrics.

Reconstruction	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Source-only	<u>28.60</u>	0.9089	<u>0.0809</u>	**0.1508**
Uniform	25.44	0.8870	0.1351	0.1756
Source anchoring	28.36	**0.9153**	0.1005	0.1624
Anchored, support masks	**28.90**	<u>0.9123</u>	**0.0807**	<u>0.1527</u>

4.4 Component Analysis

Source anchoring. We next isolate supervision at fixed inputs and 31k updates. Paired controls also share streams, support, initialization, seeds, and auxiliary losses. Table 2 separates source sampling frequency (Repetition) from source loss strength (Weighting), averaging three identities equally over 120 targets at 69.09–74.49 KiB.

Repetition alone improves quality more than weighting alone, showing that update frequency matters within the finite optimization budget. Yet a coefficient-matched uniform sampler, using source/support weights 4.8 and 2/3, reaches 24.85 dB BBox PSNR, close to the joint setting. The near match suggests that expected source photometric mass drives much of the benefit beyond the particular repetition schedule.

Observed and generated constraints. To separate the roles of observed and generated evidence, Table 3 compares four reconstruction rules on scans *0000*, *0100*, and *0200*, averaging each scan's forty targets equally. Source-only uses decoded RGB and silhouettes; uniform adds generated views and silhouettes; anchoring strengthens sources within that input set. The anchored support-mask variant retains generated silhouettes but removes their RGB photometry. All rows share streams, initialization, cameras, and 31k updates. Bold and underline mark first and second per metric.

Table 3 reveals an evidence trade-off. Source-only avoids generated RGB and has higher PSNR and lower DISTS than the original anchored rule; anchoring retains generated views for coverage and reaches the highest SSIM. It also improves all four means over uniform supervision. Keeping support silhouettes but omitting their RGB photometry yields another operating point with the highest PSNR and lowest LPIPS. The system comparisons use the original anchored rule; *Appendix A.2.3* examines staged support supervision.

Visual effects of source anchoring. Figure 5 shows generated support beside four reconstructions of the same target: uniform supervision, repetition, weighting, and joint source anchoring. A reference image, matched head crops, and full-body BBox scores locate the visual and measured differences. Across twelve paired VVC/GLIC comparisons, decreasing support weight also improves mean canvas, BBox, and head metrics, connecting this example to the broader supervision test.

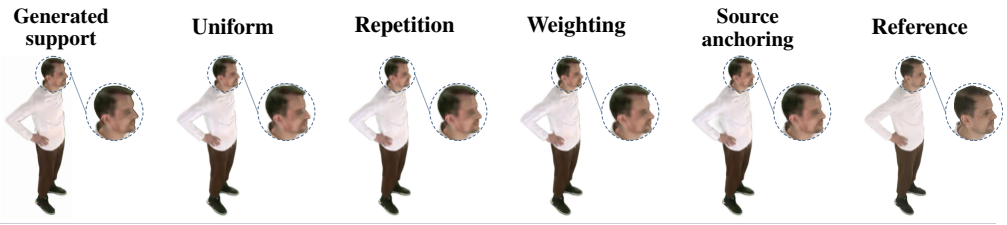


Figure 5: **Source anchoring in one view.** The combined rule lowers LPIPS and DISTS below either single change and better preserves the face in the matched head crop.

Table 4: **Source-view scalability.** The source-anchored rule improves all four person-crop metrics from two to eight transmitted views; its PSNR gain grows with the source set.

Sources	KiB	Rule	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
2	15.75	Uniform	19.94	0.8186	0.2052	0.2156
		Anchored	**20.71**	**0.8246**	**0.1994**	**0.2124**
4	24.70	Uniform	23.24	0.8677	0.1596	0.1859
		Anchored	**25.42**	**0.8846**	**0.1364**	**0.1767**
6	32.60	Uniform	24.39	0.8765	0.1451	0.1790
		Anchored	**27.16**	**0.8998**	**0.1243**	**0.1687**
8	40.97	Uniform	25.34	0.8863	0.1337	0.1740
		Anchored	**29.10**	**0.9168**	**0.0962**	**0.1606**

4.5 Scalability and Transfer

Source-view scaling under a fixed rule. We vary nested source sets from two to eight views on scan *0000*, retaining forty target/support cameras and VVC QP 37 (Table 4). Coverage and complete size both change. Each paired control shares inputs and 31k updates; anchoring retains four repetitions, source weight 1.8, and unit support weight across counts.

With two views, limited captured coverage constrains the benefit of anchoring. Its BBox PSNR advantage rises from +0.76 to +3.76 dB as the source set grows; the expected source photometric share increases from 0.265 to 0.590. Camera coverage and supervision mass therefore change together, rather than source count alone determining the gain.

Transfer to new scans without retuning. Table 5 evaluates scans *0400*, *0450*, and *0500* with source selection, QP, prior, supervision, initialization, and endpoint frozen. Uniform and anchored branches share VVC inputs and support; MV-HEVC uses the same anchored receiver, its own decoded inputs and support, and the nearest larger measured package.

Table 5: **Transfer to new human scans.** With its reconstruction rule fixed, SAGE leads the paired alternatives across all four person-crop metrics on three additional scans.

Method	KiB range	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
MV-HEVC	35.44–41.64	26.06	0.8857	0.1288	0.1797
Uniform	34.96–41.24	24.14	0.8848	0.1372	0.1728
Source anchoring	34.96–41.24	**26.72**	**0.9116**	**0.1004**	**0.1572**

The fixed rule gains PSNR and reduces LPIPS against uniform supervision on every scan. Against MV-HEVC with the same Gaussian receiver, its mean BBox PSNR is 0.66 dB higher. The paired uniform comparison ties the gain to source supervision rather than scan-specific retuning.

5 Conclusion

We present **SAGE, a source-anchored strategy for sparse-view human compression**. It gives decoded observations and generated support distinct roles in Gaussian optimization: source-dependent sampling and photometric weighting protect captured appearance, while generated views fill missing coverage. A staged support schedule steers later updates toward observed detail, leaving a reusable scene for calibrated queries. At matched inputs and budgets, SAGE improves DNA-Rendering BBox PSNR by 2.96 dB over uniform supervision. Complete-package THuman2.0 comparisons favor SAGE over size-aligned MV-HEVC with a common receiver, and a frozen-rule test extends anchoring to further scans. Together, these findings make image provenance a useful design signal for generative reconstruction. The current framework uses a fixed human prior and calibrated sparse views; adapting to broader capture conditions and learned rate control is a promising next step.

STATEMENTS

AI USE STATEMENT

AI tools assisted with finding related literature, exploring research ideas, writing experimental code, interpreting and visualizing measured results, and drafting or polishing parts of this paper. They also helped execute and monitor some experiments. Human authors set the research questions, chose the protocols, reviewed the outputs, and made the final scientific and editorial decisions. Every reported measurement comes from executed experiments and saved data rather than an AI-generated estimate. The authors take responsibility for the implementation, results, citations, and manuscript, including all AI-assisted work.

ETHICS STATEMENT

This work uses existing research datasets of synchronized human captures and conducts no new human-subject study. Human imagery contains identifiable appearance, clothing, and motion; dataset access, redistribution, and release of derived artifacts are subject to the original licenses and consent conditions. Free-viewpoint human reconstruction and compression can support telepresence and immersive communication, while unauthorized use can enable impersonation, non-consensual reconstruction, or surveillance. Practical deployment therefore requires subject authorization, provenance disclosure, access control, retention and removal mechanisms, and safeguards against deceptive use. The experiments focus on reconstruction quality and coding efficiency rather than demographic fairness or biometric recognition.

REPRODUCIBILITY STATEMENT

The main paper specifies the task, available observations, rate definition, baselines, and metrics. *Appendices A.1 and A.2* describe source/target splits, Gaussian optimization, trust construction, and training budgets. *Appendix B* provides additional comparisons, controlled studies, and statistical analysis. Companion materials retain per-view measurements and arithmetic-reproduction scripts, with implementation provenance and execution records kept outside the manuscript. The document source package includes the bibliography and visual assets for an independent PDF build. Repeating the experiments additionally requires the experimental implementation, pretrained weights, and legally acquired datasets, which are not all bundled with the document sources.

REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational Image Compression with a Scale Hyperprior. In *International Conference on Learning Representations*, 2018.

Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, International Telecommunication Union, Video Coding Experts Group, 2001. URL https://www.itu.int/wftp3/av-arch/video-site/0104_Aus/VCEG-M33.doc.

Yochai Blau and Tomer Michaeli. Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In *International Conference on Machine Learning*, pp. 675–685, 2019.

Jill M. Boyce, Renaud Doré, Adrian Dziembowski, Julien Fleureau, Joel Jung, Bart Kroon, Basel Salahieh, Vinod Kumar Malamal Vadakital, and Lu Yu. MPEG Immersive Video Coding Standard. *Proceedings of the IEEE*, 109(9):1521–1536, 2021.

Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

Yunuo Chen, Bing He, Zezheng Lyu, Hongwei Hu, Qunshan Gu, Yuan Tian, and Guo Lu. Adaptive Learned Image Compression with Graph Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12150–12161, 2026a.

Yunuo Chen, Zezheng Lyu, Bing He, Hongwei Hu, Qi Wang, Yuan Tian, Li Song, Wenjun Zhang, and Guo Lu. Content-Aware Mamba for Learned Image Compression. In *International Conference on Learning Representations*, 2026b.

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-Centric Rendering. In *IEEE/CVF International Conference on Computer Vision*, pp. 19982–19993, 2023.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejia Xu, and Zhangyang Wang. LightGaussian: Unbounded 3D Gaussian Compression with 15x Reduction and 200+ FPS. In *Advances in Neural Information Processing Systems*, 2024.

Yujie Guo, Yudong Jin, Lingteng Qiu, Zehong Shen, Zhen Xu, Jing Zhang, Xianchao Shen, Hujun Bao, Sida Peng, and Xiaowei Zhou. PointSplat: Compact Gaussian Splatting via Human-Centric Prediction. *arXiv preprint arXiv:2606.32036*, 2026.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 634–644, 2024.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics*, 42(4):1–12, 2023.

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4D: 4D Consistent Human View Synthesis from Sparse-View Videos with Spatio-Temporal Diffusion Models. In *IEEE/CVF International Conference on Computer Vision*, pp. 11047–11057, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, 2024.

YoungJoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo J. Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable Human Gaussians for Sparse View Synthesis. In *European Conference on Computer Vision*, 2024.

Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3D Gaussian Representation for Radiance Field. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21719–21728, 2024.

Jiahao Li, Bin Li, and Yan Lu. Deep Contextual Video Compression. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct any scene from sparse views with video diffusion model. *IEEE Transactions on Image Processing*, 35:2305–2319, 2026. doi: 10.1109/TIP.2026.3666733.

Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. MVSGaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *European Conference on Computer Vision*, 2024.

Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. DVC: An End-to-End Deep Video Compression Framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11006–11015, 2019.

Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20654–20664, 2024.

Fabian Mentzer, George D. Toderici, Michael Tschannen, and Eirikur Agustsson. High-Fidelity Generative Image Compression. In *Advances in Neural Information Processing Systems*, 2020.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, 2020.

Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3D Gaussian Splatting for Accelerated Novel View Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10349–10358, 2024.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Haocheng Tang, Ruoke Yan, Xinhui Yin, Qi Zhang, Xinfeng Zhang, Siwei Ma, Wen Gao, and Chuanmin Jia. HGC-Avatar: Hierarchical gaussian compression for streamable dynamic 3D avatars. In *ACM International Conference on Multimedia*, pp. 8125–8134, 2025. doi: 10.1145/3746027.3755317. URL https://arxiv.org/abs/2510.16463.

Haocheng Tang, Ruoke Yan, Xiaolong Zhang, Jiaqi Zhang, Qi Wang, Lili Chen, and Siwei Ma. Efficient feedforward human-centric video compression via 3D gaussian generation. In *Data Compression Conference*, pp. 468, 2026. doi: 10.1109/DCC66757.2026.00093.

Gerhard Tech, Ying Chen, Karsten Müller, Jens-Rainer Ohm, Anthony Vetro, and Ye-Kui Wang. Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(1):35–49, 2016.

Vladan Velisavljević, Gene Cheung, and Jacob Chakareski. Bit allocation for multiview image compression using cubic synthesized view distortion model. In *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2011.

Yizong Wang, Hongbo Ning, Haohua Wang, Yutao Yuan, Yue Ling, Dong Zhao, Siwei Ma, and Wen Gao. Argus: Bandwidth-Efficient Live Multiview Video Streaming via Sparse-View Gaussian Reconstruction. In *IEEE Conference on Computer Communications*, pp. 1–10, 2026. doi: 10.1109/INFOCOM59046.2026.11571422.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-Viewpoint Rendering of Moving People from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16210–16220, 2022.

Matthias Wödlinger, Jan Kotera, Jan Xu, and Robert Sablatnig. SASIC: Stereo image compression with latent shifts and stereo attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 661–670, 2022.

Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, and Robert Sablatnig. ECSIC: Epipolar Cross Attention for Stereo Image Compression. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3436–3445, 2024.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. ReconFusion: 3d reconstruction with diffusion priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21551–21561, 2024.

Ruoke Yan, Mingjia Yang, Xinfeng Zhang, Haocheng Tang, Qian Yin, Zhipin Deng, Kai Zhang, Li Zhang, and Siwei Ma. GauMVC: Generative decoupled gaussian representation for human-centric multi-view video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4963–4972, 2026.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4D: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5746–5756, 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. GPS-Gaussian: Generalizable Pixel-wise 3D Gaussian Splatting for Real-time Human Novel View Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19680–19690, 2024.

Appendix

Sparse-View Human Compression via Source-Anchored Gaussian Reconstruction

Anonymous submission to ICLR 2027

APPENDIX CONTENTS

A Method details 14
A.1 Task and rate accounting 14
A.2 Construction and optimization 16
A.3 Source-cycle correction 18
B Additional experiments 18
B.1 Core component studies 18
B.2 Generalization and robustness 23
B.3 THuman2.0 system transfer 24
B.5 Compression baselines 26
B.6 Reconstruction and query cost 28
B.7 Human-region analysis 30
B.8 Identity-level effects 30
B.9 Gaussian refinement studies 30
C Analysis and outlook 31
C Scope and deployment 31
C Measured points and envelopes 31
C.3 Input-rate interaction 33

Guide to the appendix. The sections below group the method, measurements, and their interpretation:

Start with the method. Appendix A defines the available observations, reconstruction objective, source trust, and decoder inference.

For the mechanisms. Appendices B.1 and B.9 distinguish source supervision, appearance correction, and the mechanisms of Gaussian refinement.

Compare the results. Appendix B groups measured comparisons; Appendices B.9 and B.9 report repeated low-rate Gaussian reconstruction.

For interpretation. Appendices B.8 and C explain statistical uncertainty, deployment costs, and the scope of the evidence.

Reading conventions. Figure and table numbering continues from the main paper. Rates include images, conditions, and packaging; all four metrics describe the same output and shared evaluation regions. Complete measurements accompany the paper, including the points underlying the geometric summaries in Appendix C.

A Method Details

A.1 Task Definition and Complete-Rate Accounting

A.1.1 Information access

To reconstruct forty target views, the decoder receives eight coded source images, cameras, poses, and source masks, together with shared networks. Pose extraction uses the full capture at the encoder; scene construction takes place at the decoder. Auxiliary-geometry studies are distinguished below.

A.1.2 Reconstruction inputs and conditions

The three-identity control shares eight-source visual hulls, initialization, support, and 30k+1k updates between uniform and anchored supervision. Each frontend generates its own support RGB and masks from decoded inputs. The five-unit Full study uses eight-source hulls with the construction and appearance recipe specified in Appendix B.4.

Pose conditions triangulate Sapiens detections (Khirodkar et al., 2024) and reproject the skeleton to the camera grid. The mapping uses 63 unique Goliath joints for 65 drawn landmarks, including duplicated wrists. Weighted triangulation and Huber refinement use confidence-filtered observations. Packages transmit coded projections or structured coordinates; the companion implementation notes give confidence and joint-mapping details.

Restricting pose triangulation to eight transmitted cameras on *0023_06*, times 72–78, gives 441 finite joint–time estimates. Median projection differences from 48-camera pose span 1.79–2.87

pixels. Three paired 31k runs at time 75 share decoded RGB, masks, calibration, an 8,646-point source hull, and forty 1023×1023 targets, regenerating support for each pose condition. Complete packages are 51,397/51,409 bytes for 8/48-camera pose. Their BBox PSNR/SSIM/LPIPS/DISTS means are $(25.42, 0.9178, 0.0659, 0.1222)$ and $(25.36, 0.9174, 0.0690, 0.1232)$. PSNR and LPIPS improve in all three paired runs; SSIM/DISTS directions vary. Paired PSNR standard deviation is 0.082 dB. Full canvas vectors and per-view records accompany the paper; the main system comparison retains full-capture pose.

Earlier three-identity controls use auxiliary geometry from 48 ground-truth masks and pre-compression support masks. Their archived Low/High and repetition/weighting studies describe a shared-geometry setting, separate from current source-hull controls. Historical trust-enhanced outputs also include image-space calibration.

A.1.3 Rate definition

For task a, we record

$$R_a = B_{a,\mathrm{RGB}} + B_{a,\mathrm{cond}} + B_{a,\mathrm{meta}} + B_{a,\mathrm{repr}} + B_{a,\mathrm{ctr}}, \tag{7}$$

where $B_{a,\mathrm{cond}}$ contains cameras, poses, and eight source masks; $B_{a,\mathrm{meta}}$ counts task metadata. Generated masks are computed at the decoder. Scene attributes cost $B_{a,\mathrm{repr}} = 0$ for shared-prior coding and their actual bytes for self-contained streams. Components are disjoint: metadata and packaging inside a condition archive are counted there once. RGB-only analyses remain separate. For target images of height H and width W, the task-normalized rate is

$$\mathrm{bpp}_{\mathrm{task}} = \frac{8R_a}{|\mathcal{T}|HW}, \tag{8}$$

where bpp denotes bits per pixel; H, W are target height and width, respectively, and $|\mathcal{T}| = 40$ for the principal eight-source tasks. Source-normalized complete and RGB rates are $8R_a / \sum_{i\in\mathcal{S}} H_i W_i$ and $8B_{a,\mathrm{RGB}} / \sum_{i\in\mathcal{S}} H_i W_i$, using the actual coded source dimensions. For eight sources and forty targets at equal resolution, complete-source bpp is five times task bpp. These are different normalizations of the same stream. The historical self-contained study uses $\mathrm{bpp}_{48} = 8R_a/(48HW)$. Our principal rate axis is actual complete KiB, independent of query count.

A.1.4 Payload composition and rate alignment

Conditions occupy a median 59.0% in the sixteen historical paired groups. Table 1 uses different packages: SAGE's RGB streams occupy 24,473–30,660 bytes and its compressed 3D pose 847–923 bytes, whereas GLIC/ECSIC transmit 21,723–23,677 bytes of coded skeleton images. All include cameras, source masks, metadata, and framing. The companion decomposes all 15 actual packages; their component sums reproduce the table. TMIV retains its joint texture/depth bitstream and query conditions. These are complete-configuration comparisons, not isolated RGB-codec effects.

The statistical comparison brackets each SAGE payload R with two real MV-HEVC payloads $R^- \le R \le R^+$ produced from the same source cameras and conditions. The original five-unit statistic uses linear-byte weights $\alpha = (R - R^-)/(R^+ - R^-)$; Appendix B.8 reports its log-byte sensitivity. Later reconstruction studies use log-byte weights $\alpha = (\log R - \log R^-)/(\log R^+ - \log R^-)$. Both protocols align each target before aggregation. Interpolation serves numerical comparison, while RD markers and SAGE scores correspond to actual measured reconstructions.

A.1.5 Captures and data splits

The Full-output study uses five captures from two identities: *0013_01* at times 75, 100, and 149, and *0023_06* at times 75 and 100. Schedule and visualization development uses *0023_06*/149 and the dense curve on *0013_01*/100. Five further identities—*0012_09*, *0235_11*, *0326_07*, *0345_08*, and *0811_06*—extend reconstruction comparisons to two complete rates. These development and descriptive evaluations retain their respective configuration-freeze histories. The base prior's identity overlap remains unverified.

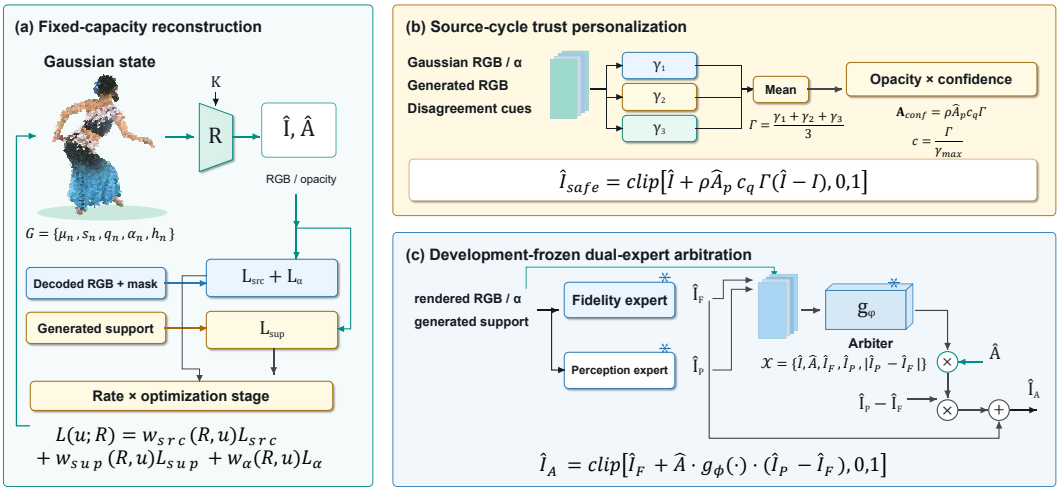


Figure 6: **Reconstruction and query refinement.** (a) Source and support losses update the late fixed-capacity Gaussian state. (b) The averaged source-cycle gate defines an opacity- and confidence-modulated RGB correction. (c) Frozen experts and an alpha-modulated mixer produce Full output. The displayed feature group $\mathcal{X}$ omits generated RGB, which supplies the remaining three mixer channels. Panels (b) and (c) are alternative query refinements.

A.2 Decoder Construction and Optimization

A.2.1 Sparse support generation

The frozen Diffuman4D generator (Jin et al., 2025) produces support from decoded sources and coded conditions. Target ground-truth RGB and masks are absent from the training manifest and replaced by zeros in the decoder path. Decoded source masks and cameras initialize the point cloud.

A.2.2 Reconstruction and output modes

Figure 6 expands the main overview into scene optimization and two appearance-refinement studies. Panel (a) shows late fixed-population refinement after densification and pruning. Panel (b) studies three-seed source-cycle correction; panel (c) shows the fixed experts and spatial mixing network used for Full output. The two appearance branches are alternatives, not successive stages.

Gaussian-only queries use the reconstructed scene and a calibrated camera. Full additionally consumes camera-specific generated support, two shared experts, and the spatial mixer. Its evaluated path uses the prescribed forty-camera support grid. A new camera requires a compatible support-generation path and a separate latency measurement.

Rendered alpha measures foreground occupancy along a Gaussian ray. Source provenance instead determines sampling and photometric coefficients. During scene construction, silhouette supervision differentiates through rendered alpha. The core recipe uses fixed mask-confidence constants and disables optional photometric confidence. During Full inference, alpha modulates the fidelity-to-perception increment in Equation (15); Gaussian parameters remain fixed.

The diagram summarizes the role of evidence item e at pixel x through provenance τ_e, foreground support $v_e(x)$, and confidence $c_e(x;r)$:

$$\rho_e(x;r) = \tau_e v_e(x) c_e(x;r). \tag{9}$$

These nonnegative factors organize the evidence roles; r is the coding setting. The actual optimization uses the sampled objective in Equations (4)–(6), with separate photometric and auxiliary coefficients.

In Figures 1–2, G_ψ is the prior, the diagram's G is scene $\mathcal{G}$, and K or K_q is a calibrated query camera. The axis triad shows three-dimensional (3D) coordinates; the VAE and denoiser form the multiview synthesis prior; d_i guides QP allocation. The R icon in Figures 2 and 6 denotes the renderer $\mathcal{R}$ from Equation (2), distinct from the byte count $R(b)$.

Panel (a) of Figure 6 uses source, support, and silhouette losses. The main figure's $w(R, u)$ is the tuple $(w_{\mathrm{src}}, w_{\mathrm{sup}}, w_\alpha)$ in Equation (11) at update u and byte anchor R. Alpha measures ray occupancy, not source visibility.

In panel (b), $\gamma_1, \gamma_2, \gamma_3$ are bounded spatial gates from three fixed seeds. The diagram's $\Gamma = (\gamma_1 + \gamma_2 + \gamma_3)/3$ equals $\bar{\gamma}_j$ in Equation (12); $c = \Gamma/\gamma_{\max}$ uses the gate upper bound $\gamma_{\max}$, where ρ is a scalar correction scale, distinct from conceptual evidence weight ρ_e. The parameters p and q are exponents: the diagram's opacity–confidence factor $A_{\mathrm{conf}} = \rho \hat{A}^p c^q \Gamma$ is the effective correction coefficient, not a camera or primitive index. $\tilde{I}$, $\hat{I}$, and $\hat{I}_{\mathrm{safe}}$ denote generated, Gaussian-rendered, and trust-corrected RGB. Clipping limits RGB to $[0, 1]$; alpha-zero pixels receive no trust correction.

Panel (c)'s $\hat{I}_F, \hat{I}_P, \hat{I}_A$ match the superscripts in Equation (15). Its displayed group $\mathcal{X} = [\hat{I}, \hat{A}, \hat{I}_F, \hat{I}_P, |\hat{I}_P - \hat{I}_F|]$ has thirteen channels. The implemented input $\mathbf{f}_j$ concatenates $\tilde{I}_j$ as a further three RGB channels, yielding sixteen. Multiplication nodes apply elementwise alpha and gate weights before addition to $\hat{I}_F$ and clipping. Panels (b,c) change query RGB; panel (a) updates scene parameters.

A.2.3 Reconstruction objective

Each update samples one camera as in Equations (4)–(5). The source coefficient applies to photometry; silhouette supervision is weighted separately and scale regularization is disabled. All metrics describe the same complete scene checkpoint.

Masks are downscaled to the current rendering resolution, thresholded at 0.5, and, if needed, resized to the alpha dimensions with nearest-neighbor interpolation. Define $F = \{x : M(x) > 0.5\}$ and $B = F^c$. Square max/min filters D_k, E_k use stride one and padding $\lfloor k/2 \rfloor$, ignoring out-of-image samples through max-pooling's infinite padding. The boundary is $\partial M = \{x : D_7(M)(x) - E_7(M)(x) > 0.5\}$. Thus 7 denotes a kernel size, not a band width; boundary pixels also belong to F or B. Generated confidence equals 0.15 on $U_{11} = \{x : D_{11}(M)(x) - E_{11}(M)(x) > 0.5\}$ and 1 elsewhere. Source confidence is 1 throughout. Define

$$h(x) = -c(x)\big[M(x)\log(\hat{A}(x) + \epsilon) + (1 - M(x))\log(1 - \hat{A}(x) + \epsilon)\big], \quad \epsilon = 10^{-6}. \tag{10}$$

The silhouette objective $\mathcal{L}_\alpha$ is $\frac{1}{4}\,\mathrm{mean}_F\, h + \frac{1}{4}\,\mathrm{mean}_B\, h + \frac{1}{2}\,\mathrm{mean}_{\partial M}\, h$. Empty foreground/background falls back to the whole-image mean; an empty boundary uses the balanced foreground/background mean. Its coefficient is 0.006 from base step 28,000. The masks and confidence constants are fixed, not normalized trainable weights; the optional photometric confidence branch is disabled. Generated masks come from that same frontend's support RGB.

The 1k terminal switches from white to black background, fixes the population, retains silhouette and source photometry, and adds $0.01\ \mathrm{mean}_n\, |\sigma(o_n) - \sigma(o_n^{(30,000)})|$, where o_n is an opacity logit, $o_n^{(30,000)}$ its frozen base value, and σ the logistic sigmoid. Dual-background regularization is disabled; perceptual expert losses belong to the separate query component.

The staged-support control starts from the same 30k checkpoint as the support-mask variant. Its final 1k updates restore generated RGB photometry at coefficient 0.25 and raise the SSIM fraction to 0.6; source sampling and multiplier, silhouettes, opacity regularization, and the update count are otherwise unchanged. This control changes the Gaussian state, not the query-time appearance module. Table 6 retains the complete metric vector for each measured setting; the staged setting has the highest PSNR and SSIM here, whereas the support-mask setting has the lowest LPIPS and source-only has the lowest DISTS.

A.2.4 Rate-conditioned schedule

The diagram denotes a frozen anchor-specific configuration by

$$\mathcal{L}(u; R) = w_{\mathrm{src}}(R, u)\mathcal{L}_{\mathrm{src}} + w_{\mathrm{sup}}(R, u)\mathcal{L}_{\mathrm{sup}} + w_\alpha(R, u)\mathcal{L}_\alpha, \tag{11}$$

where R selects a measured-rate anchor, u is the update index, and the three coefficients select source, generated-support, and silhouette supervision. The shorthand $w(R, u)$ denotes this coefficient tuple. The current core control instead uses the fixed recipe in Appendix A.2.3.

Table 6: **Staged-support control.** BBox means over the three THuman2.0 development scans, forty target views each. Every row uses the same coded inputs, initialization, and 31k updates. Bold identifies the best result per metric.

Reconstruction	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Source-only	28.60	0.9089	0.0809	**0.1508**
Uniform	25.44	0.8870	0.1351	0.1756
Source anchoring	28.36	0.9153	0.1005	0.1624
Anchored, support masks	28.90	0.9123	**0.0807**	0.1527
Staged support	**29.22**	**0.9186**	0.0832	0.1568

The historical 30k+1k rate study optimizes each anchor independently. Its M_2 configuration uses source weight 2.305, alpha trust 0.18, base/terminal SSIM weights 0.196/0.45, and terminal LPIPS weight decreasing from 0.002 to 0.0008.

Reconstruction selects one configuration from the finite measured-rate anchors before optimization; the renderer consumes the resulting scene rather than QP or task bpp.

A.3 Source-Cycle Appearance Correction

A.3.1 Source-Cycle Trust Personalization

To study whether generated appearance can usefully correct a fixed Gaussian rendering, we evaluate a bounded, source-calibrated mixture:

$$\hat{I}_j^{\text{safe}} = \text{clip}\left(\hat{I}_j + \rho \hat{A}_j^p \left(\frac{\bar{\gamma}_j}{\gamma_{\max}}\right)^q \bar{\gamma}_j(\tilde{I}_j - \hat{I}_j), 0, 1\right). \tag{12}$$

Gate $\bar{\gamma}_j$ is a personalized prediction or the mean of three models; alpha denotes foreground support. Figure 6 distinguishes this study from dual-expert Full output.

The 213,505-parameter gate takes Gaussian/generated RGB, their absolute difference, alpha, and local/high-frequency disagreement. A 48-channel stem and five residual blocks produce $\gamma = 0.125\,\text{sigmoid}(z)$ from logit z, with $\gamma_{\max} = 0.125$. Personalization trains on six decoded sources and selects on two. Shared consensus pools three captures (18/6 views), averaging three 4,000-update gates before $(\rho, p, q) = (0.6, 1, 3)$. The same-architecture student distills normalized gates and decoded-source predictions; validation selects step 2,000 from twelve checkpoints across three runs. Full architecture, operators, and losses remain in the companion.

Amplitude-matched constant and spatially shuffled controls yield 44.72 dB on 24 development source views, versus 44.88 dB for the learned gate. A gate-independent residual candidate has increasing benefit across confidence buckets. The same frozen predictions score 20.00 dB against decoded sources and 19.91 dB against pre-codec sources, illustrating the distinction between source consistency and original appearance.

On 120 targets, three-gate consensus improves PSNR by 0.05 dB and LPIPS by 0.0001; SSIM and DISTS intervals cross zero. Distillation gives similar mean effects with gate-only latency reduced from 49.10 to 16.37 ms. Full costs, including its separate experts and spatial mixer, are measured independently.

B Additional Experimental Results

B.1 Component Analysis

We first test Gaussian supervision and source coding with paired controls, holding other factors fixed within each study. Optional query-time appearance enhancement is analyzed separately below.

B.1.1 SOURCE-SUPERVISION DECOMPOSITION

To determine whether anchoring transfers across codecs, we cross the reconstruction rule with the coding frontend (Table 7). Each paired supervision control shares inputs, support, initialization, and updates; each frontend supplies its own decoded RGB and generated support.

Table 7: **Frontend and supervision controls.** Gaussian-only outputs use 31k updates, with three identities weighted equally. SSIM, LPIPS, and DISTS use the foreground BBox; Gaussian counts are measured after the base stage. Paired supervision settings share bytes within each frontend.

Frontend	Supervision	Canvas PSNR	BBox PSNR	SSIM	LPIPS	DISTS	Gaussians
VVC	Uniform	27.10	21.93	0.8805	0.1050	0.1536	27.5k
VVC	Source anchored	30.06	24.89	0.9004	0.0734	0.1399	35.6k
MV-HEVC	Uniform	27.13	21.96	0.8743	0.1089	0.1637	27.4k
MV-HEVC	Source anchored	29.96	24.81	0.8932	0.0790	0.1510	35.8k

Anchoring gains 2.96/2.84 dB BBox PSNR with VVC/MV-HEVC; all four BBox metrics improve for each identity within both frontends. This comparison isolates supervision within each coded input. The companion retains identity-wise results for time 30 and forty targets, including shared initialization, support, auxiliary losses, seeds, and 31k endpoints.

Expected photometric coefficients total 1.0000/1.0000/1.1333/1.3556 for uniform/repetition/weighting/joint; joint source share is 0.5902. Eighteen scenes yield 720 measurements over 120 unique references and three development identities, with one pipeline seed.

The visualizations in Figure 5 and Appendix B.1.2 composite predictions using their own alpha; quantitative evaluation retains the unmasked black-canvas outputs.

The VVC factorial interaction, joint minus repetition minus weighting plus uniform with improvement-oriented metrics, is 0.33 dB PSNR, -0.0014 SSIM, 0.0043 LPIPS, and 0.0018 DISTS. Interaction is therefore metric-dependent. MV-HEVC *0811_06* canvas LPIPS increases by 0.0033 despite improved BBox LPIPS, identifying a background-sensitive effect.

B.1.2 EXPECTED-COEFFICIENT MATCHING

To distinguish supervision balance from update frequency, we match the expected coefficients while retaining uniform sampling: $p_s = 1/6$, $\lambda_s = 4.8$, and $\lambda_g = 2/3$ give $p_s\lambda_s = 0.8$ and $(1 - p_s)\lambda_g = 5/9$. Table 8 fixes initialization, support, auxiliary weights, seeds, and 31k updates across three VVC captures. Sampling still affects stochastic gradients, silhouette updates, and Gaussian growth.

Table 8: **Photometric-supervision controls.** Three development identities share coded inputs and 31k updates across rules. Coefficient-matched and anchored rows share expected source/support coefficients. The generated-photometry ablation keeps anchored sampling and source weights; counts are final Gaussian populations.

Supervision	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	Gaussians
Uniform	21.93	0.8805	0.1050	0.1536	27,541
Coefficient matched	24.85	**0.9004**	0.0737	0.1405	36,989
Without generated photometry	**24.90**	0.8961	0.0749	0.1462	22,242
Source anchoring	24.89	**0.9004**	**0.0734**	**0.1399**	35,633

The coefficient-matched control recovers most of the anchoring gain. Anchoring changes BBox PSNR by -0.07, $+0.23$, and -0.02 dB relative to this control on *0012_09*, *0345_08*, and *0811_06*, respectively. Similar mean quality despite different primitive counts links the principal effect to supervision balance; the sampling implementation remains capture-dependent.

Generated appearance supervision. Setting generated photometry to zero retains anchored sampling, source weight, initialization, support, generated silhouettes, terminal opacity regularization, and the 31k schedule. Unused photometric steps are not reassigned. Table 8 therefore isolates

appearance supervision, while other prior-derived information remains active. Generated appearance supervision improves BBox DISTS on all three identities, SSIM and LPIPS on two, and PSNR on one. Its structural and perceptual means improve while mean PSNR remains comparable. The complete regional measurements accompany the supplement. Across the three identities in Table 8, source anchoring has the lowest LPIPS and DISTS and shares the highest displayed SSIM. Removing generated photometry changes mean PSNR by $+0.01$ dB but worsens the other three means. Effects differ by capture: on *0811_06*, the ablation improves all four canvas metrics, including 0.35 dB in PSNR. Thus the single-capture result and the three-identity mean answer different questions about generated appearance supervision. Transferred without further tuning to the three THuman2.0 scans of Table 3, the same ablation leads BBox PSNR and LPIPS, while the original anchored rule leads SSIM and source-only leads DISTS. These scans were already part of the analysis, so this transfer characterizes the supervision trade-off rather than an independent identity test. Across these six states, omitting generated photometry changes mean tensor storage from 8.02 to 5.01 MiB, loaded-query latency from 1.79 to 1.65 ms, and BBox DISTS from 0.1399 to 0.1462. Archived two-process Gaussian-construction times average 662.3 and 495.4 s, respectively, including process startup and optimization but excluding support generation. Query timings use an H800, FP32, batch one, ten warm-ups, and five passes over forty cameras with GPU-synchronized wall-clock timing. Quality, storage, and timing records refer to the same checkpoints. Generated silhouettes remain part of reconstruction.

Source-supervision atlas. Figure 7 compares supervision under fixed inputs, source initialization, support, and 31k updates across three development identities.

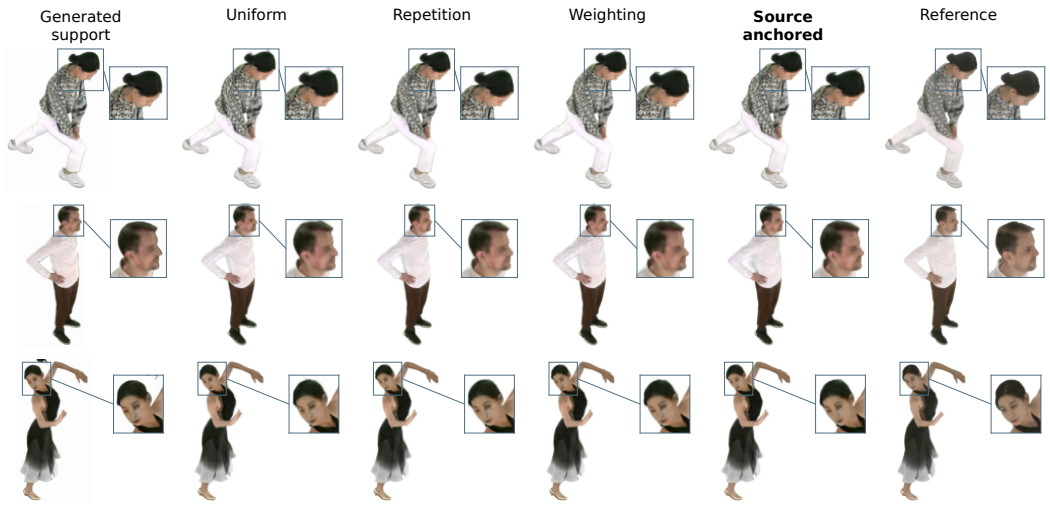


Figure 7: **Source-supervision atlas.** Rows show *0012_09*, *0345_08*, and *0811_06* at time 30, camera 30. Generated support is the fixed completion input; uniform, repetition, weighting, and anchored reconstruction share 31k updates. Insets use the same head region across columns.

B.1.3 Support Supervision over Optimization

To shift reconstruction from generated coverage toward observed detail, we retain source probability $4/9$ and coefficient 1.8 while multiplying generated-view photometry by

$$\eta(u) = 1 - 0.75 \operatorname{clip}\left(\frac{u - 5{,}000}{23{,}000}, 0, 1\right), \tag{13}$$

where u is the global update and clipping truncates to $[0, 1]$. Generated weight stays at 1 through 5k, decreases to 0.25 at 28k, and remains there through the terminal stage. Expected source coefficient share rises from 0.590 to 0.852. Initialization, support, auxiliary objectives, and the 30k+1k budget stay fixed; the schedule changes both evidence balance and total photometric scale relative to auxiliary losses.

Selection uses two GLIC development rates of *0811_06*; the frozen rule transfers to two captures, two further GLIC rates, and four VVC rates. Both frontends share the native source canvas and

condition representation, generating support from their own streams. A paired optimizer-seed repeat keeps prior outputs fixed. Table 9 reports within-stream effects. To locate appearance changes, Figure 8 compares the same head regions from three calibrated cameras. Head diagnostics use saved 8-bit renders and forty shared fixed boxes.

Table 9: **Support-annealing controls.** Gaussian-only outputs use 31k updates. Rows average forty views within rates, then rates equally. PSNR columns are absolute BBox scores; other columns give SSIM increases and LPIPS/DISTS reductions from fixed joint supervision. Seed A/B repeats the same two-rate comparison with different optimization seeds, not identities.

Control	Fixed PSNR	Variant PSNR	ΔSSIM	LPIPS reduction	DISTS reduction
Uniform, selection rates	26.38	24.75	−0.0143	−0.0153	−0.0127
Coefficient-matched, selection rates	26.38	26.40	+0.0004	+0.0005	−0.0001
Annealed, selection rates (seed A)	26.38	26.68	+0.0030	+0.0031	+0.0042
Annealed, selection rates (seed B)	26.37	26.69	+0.0032	+0.0040	+0.0045
GLIC, 0012_09 (one rate)	23.34	23.63	+0.0052	+0.0034	+0.0043
GLIC, 0345_08 (one rate)	25.43	25.54	+0.0007	+0.0054	+0.0056
GLIC, 0811_06 (four rates)	26.35	26.67	+0.0030	+0.0037	+0.0040
VVC, 0811_06 (four rates)	26.39	26.66	+0.0026	+0.0070	+0.0027

On the two selection rates, annealing changes mean BBox PSNR by +0.299 dB in seed A and +0.318 dB in seed B. The corresponding head-region PSNR changes are +0.300 and +0.238 dB, measured over all forty fixed head boxes. The other rows keep the selected schedule fixed across captures and coding frontends. These within-stream controls separate supervision changes from source-codec effects; all regional vectors and negative per-view changes are retained in the companion. With the same annealed receiver applied to GLIC, the 59,609-byte VVC point improves all four canvas means under log-byte alignment, including 0.175 dB in PSNR. Higher budgets show a fidelity–perception trade-off between these frontends. The companion retains both linear- and log-byte comparisons, marking budgets outside measured baseline brackets as unavailable.

A terminal control retains each 30k parent and population, increasing SSIM share from 0.2 to 0.6 for the final 1k updates. Selection compares 0.4/0.6 on one VVC and one GLIC development stream, then freezes 0.6. Canvas DISTS improves at all six VVC rates (mean reduction 0.00045), with PSNR decreasing 0.031 dB and LPIPS increasing 0.00004. Four GLIC rates improve mean DISTS by 0.00018, with one adverse endpoint. Figures 4 and 8 retain the original 0.2 terminal share.

Regional and color controls. Source-LPIPS supervision at weight 0.3 reduces VVC head DISTS by 0.00410, with 28 of forty boxes improving, but reduces canvas DISTS by just 0.00006; GLIC's canvas reduction is 0.00140. Terminal source-DISTS supervision instead worsens canvas DISTS on both streams. These controls distinguish regional gains from whole-image behavior; the companion reports the two tested weights, full vectors, population, and timing.

Multiplying nonconstant spherical-harmonic coefficients by 0.8 preserves geometry and bytes. Mean canvas DISTS decreases by 0.00138/0.00178/0.00076/0.00189 across SAGE/MV-HEVC/GLIC/ECSIC rates, respectively. Under common stabilization, SAGE's log-rate-aligned DISTS advantage grows by 0.00071 over three brackets, while GLIC with shared annealing remains ahead at two rates. SAGE head DISTS increases by 0.00038. Stronger degree-dependent attenuation worsens DISTS on all five tested states. Complete measurements and timing remain separate from the main RD configuration.

Annealing lowers head-crop LPIPS in all three displayed views. The paired images separate changes in facial appearance from the remaining hair and silhouette discrepancies; the complete regional measurements retain the other thirty-seven targets and the DISTS scores.

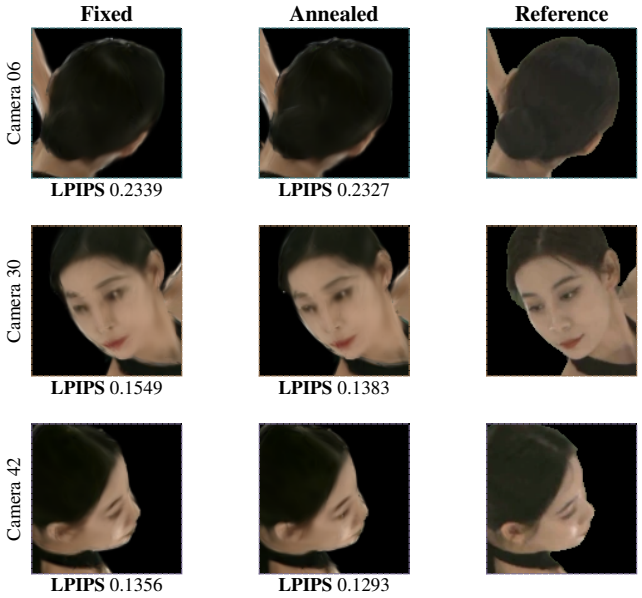


Figure 8: **Head appearance under fixed evidence.** Fixed and annealed supervision share the 68.48-KiB GLIC stream and 31k updates on *0811_06/30*. Cameras 06, 30, and 42 use identical head crops across columns. Displayed LPIPS scores describe these crops, not the full canvas.

B.1.4 GENERATED-SUPPORT PERTURBATION

To test how the two supervision rules respond to errors in synthesized views, we perturb generated RGB during Gaussian optimization while keeping decoded sources, silhouettes, cameras, initialization, and the 31k-update recipe fixed. Table 10 compares these conditions on the same forty targets for each of the three THuman2.0 scans in Table 3. The normal condition reuses those saved outputs. The moderate condition applies a 5×5 box blur and a four-pixel right shift; the strong condition uses a 13×13 blur and a twelve-pixel shift. Newly exposed pixels are white.

Table 10: **Sensitivity to generated-support RGB.** Scan-equal foreground-BBox means from direct Gaussian rendering. Each perturbed condition uses paired uniform and anchored reconstructions with unchanged coded inputs and optimization budget.

Support RGB	Rule	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
Normal	Uniform	25.44	0.8870	0.1351	0.1756
Normal	Source anchoring	28.36	0.9153	0.1005	0.1624
Moderate perturbation	Uniform	23.77	0.8549	0.2149	0.2249
Moderate perturbation	Source anchoring	25.45	0.8959	0.1532	0.1822
Strong perturbation	Uniform	19.55	0.7433	0.3966	0.3003
Strong perturbation	Source anchoring	24.30	0.8786	0.1934	0.1921

Anchoring improves all four scan-equal BBox metrics over uniform at both perturbation levels, with PSNR gaps of +1.68 and +4.75 dB. Relative to normal support, BBox PSNR changes by -1.67/-2.91 dB (uniform/anchored) under moderate perturbation and -5.89/-4.06 dB under strong perturbation. The relative sensitivity changes with perturbation strength. The conditions isolate sensitivity to deterministic support degradation within these inspected scans; the complete per-view metric vectors are retained with the experiment records.

B.1.5 SOURCE-CYCLE BIT ALLOCATION

Source-cycle predictability ranks views for a finite QP search. For $\mathcal{M} = \{\mathrm{PSNR}, \mathrm{SSIM}, \mathrm{LPIPS}, \mathrm{DISTS}\}$, let $h_{i,m} \in [0,1]$ be the normalized difficulty of predicting omitted source i under metric m, reversing PSNR and SSIM. The difficulty is

$$d_i = |\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} h_{i,m}. \tag{14}$$

With $\bar{d} = |\mathcal{S}|^{-1} \sum_i d_i$, mean-one weights ω_i define a mean squared error (MSE) proxy, $|\mathcal{S}|^{-1} \sum_i \omega_i \, \mathrm{MSE}_{420}(Y_i, Y_i^{\mathrm{dec}})$, where Y_i, Y_i^{dec} are original and decoded luma–chroma (YUV) 4:2:0 arrays; MSE averages luma and chroma samples. Every candidate is encoded: zero-sum QP offsets do not imply equal bytes. Each selected stream receives fresh support and reconstruction. The mean-one weights are $\omega_i = (0.5 + d_i/\bar{d})/[|\mathcal{S}|^{-1} \sum_{k \in \mathcal{S}} (0.5 + d_k/\bar{d})]$ for $\bar{d} > 0$; zero hardness fails the eligibility checks and retains the uniform reference.

Hardness uses within-capture min–max normalization, reversing PSNR/SSIM. Nonfinite or near-constant metrics fall back to uniform; ties follow camera index. The proxy averages 8-bit YUV 4:2:0 squared error, with luma/chroma sample ratio $4:1:1$, then applies mean-one source weights.

The frozen Diffuman4D predictor leaves out source i while retaining its camera and capture-pose condition. Eight independent half-precision (fp16) calls use guidance 3.0 and seed 20260820. Sliding twelve-view windows take two denoising steps each with unit stride, giving 24 steps per predicted view. Each call generates the 48-camera grid; only the omitted source is scored, using source silhouettes and LPIPS/DISTS at 256×256.

Descending-hardness QP offsets are $(-1,-1,-1,-1,1,1,1,1)$, $(-2,-2,-1,-1,1,1,2,2)$, and $(-2,-2,0,0,0,0,2,2)$ around base QP 33. An inverted first template is encoded as a separate diagnostic. Every candidate is independently decoded. Eligible templates satisfy complete-byte and source-distortion bounds across two development captures; selection minimizes aggregate weighted MSE, then size, then template order. The selected template is frozen for transfer; an empty eligible set retains uniform. Exact thresholds, slice-QP adjustments, and candidate records are retained in the companion allocation notes.

B.1.6 ALLOCATION WITH FRESH SCENE RECONSTRUCTION

Table 11 tests whether source predictability improves reconstruction against uniform QP, texture, and fixed shuffled ordering. Two development captures use eight sources and forty targets. Nonuniform rules assign offsets $(-2,-2,0,0,0,0,2,2)$ around QP 33; texture uses mean horizontal and vertical luma gradients. Actual sizes lie within three percent of uniform. Each stream generates its own support and source-anchored 30k+1k scene. Three predetermined runs vary generation, initialization, and optimization seeds together.

Predictability improves all four identity-equal means over uniform in both regions, with fewer bytes per capture. BBox LPIPS and DISTS improve for both identities; PSNR and SSIM increase for *0023_06* and decrease for *0013_01*. Predictability gives the lowest mean BBox DISTS; texture and shuffled ordering favor other metrics. Comparisons use actual sizes within the tolerance, without rate interpolation.

The eighty reference images are shared across runs; companion records retain seed dispersion, paired changes, and identity-equal BBox means.

B.2 SOURCE-COUNT AND IDENTITY TRANSFER

An earlier four/six-source study uses *0166_04* and *0173_02*, excluded from source-count calibration but used for a separate router. Forty common targets and all six actual points on *0166_04* are retained. Its 5k+25k+1k recipe differs from the main control: four source copies compete with $48-n$ generated views, giving source probabilities $16/60$ and $24/66$. Source multipliers 3/2 and unit support weight give expected coefficient shares 0.522/0.533; SSIM weight is 0.25.

Against linearly rate-aligned MV-HEVC, directed PSNR/SSIM/LPIPS/DISTS changes are $(-0.05, -0.0008, -0.0005, +0.0011)$ for four sources and $(-0.16, +0.0003, -0.0002, +0.0009)$ for six (PSNR in dB). DISTS improves at both counts; the other metrics retain mixed changes.

Table 11: **Allocation with complete-pipeline repetition.** Upper blocks: BBox metrics by identity; lower block: identity-equal canvas means and sizes. Entries average forty targets and three runs. KiB includes RGB and conditions. All rules use source-anchored Gaussian-only reconstruction with 31k updates.

Identity	Allocation	KiB	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
0013_01	Uniform QP	85.39	23.66	0.8645	0.0873	0.1198
	Predictability	85.00	23.66	0.8637	0.0860	0.1189
	Texture	84.80	23.64	0.8636	0.0865	0.1215
	Shuffled	83.59	23.64	0.8640	0.0861	0.1215
0023_06	Uniform QP	78.05	25.05	0.9196	0.0651	0.1095
	Predictability	77.04	25.20	0.9208	0.0633	0.1063
	Texture	78.61	25.24	0.9211	0.0621	0.1079
	Shuffled	77.94	25.17	0.9210	0.0633	0.1092
Identity-equal full-canvas means						
Both	Uniform QP	81.72	28.69	0.9635	0.0327	0.0940
Both	Predictability	81.02	28.76	0.9637	0.0319	0.0932
Both	Texture	81.71	28.77	0.9637	0.0327	0.0933
Both	Shuffled	80.77	28.74	0.9638	0.0326	0.0935

Two further identities, *0235_11* and *0326_07*, use eight decoded-source hulls and the same frozen residual router in both branches. Each has two SAGE rates and four actual MV-HEVC brackets, trained for 31k updates. All four mean metrics favor SAGE at both rates. High-rate within-capture PSNR intervals cross zero for both identities; LPIPS and DISTS intervals also cross zero for *0326_07*. These router outputs differ from the main Full mode.

B.3 Cross-Dataset Human Reconstruction

To examine whether the reconstruction rule transfers beyond the original capture setup, we render THuman2.0 scans (Yu et al., 2021) *0000*, *0100*, *0200*, and *0300* from forty-eight equally spaced cameras at 1024×1024. Every sixth camera supplies a coded source; forty others provide targets. Meshes supply observations and references, while reconstruction initializes from transmitted source silhouettes. The encoder estimates pose by full-capture triangulation.

The frozen-rule extension in Table 5 adds scans *0400*, *0450*, and *0500* under the same shared prior, with no scan-specific retuning.

All Gaussian branches transfer the fixed 30k+1k source-anchored rule, sharing per-scan initialization and condition arrays but generating support from their own decoded RGB. VVC uses QP 37; MV-HEVC provides adjacent actual-byte brackets. Scan 0000 also compares GLIC with 512-pixel sources, ECSIC with half-resolution rectified pairs, and TMIV with source-silhouette depth. Exact pose coding preserves conditioning skeletons. Gaussian branches output direct renders of these static scenes.

Equal-scan averaging after log-byte alignment gives 0.85 dB higher BBox PSNR, 0.0244 higher SSIM, and 0.0294/0.0196 lower LPIPS/DISTS than MV-HEVC. On scan 0000, SAGE's 40.97-KiB output improves all four BBox means over GLIC (41.40 KiB) and ECSIC (51.56 KiB). TMIV at 65.29 KiB attains lower DISTS (0.1411), while SAGE improves PSNR, SSIM, and LPIPS. Figure 3 illustrates the frontal appearance differences.

A fixed-bitstream control shares 41,955 bytes, support, 7,318 initial Gaussians, auxiliary losses, and 31k updates on scan 0000. Uniform supervision uses each source once with unit weight; anchoring repeats it four times with weight 1.8. BBox PSNR/SSIM rise from 25.34/0.8863 to 29.10/0.9168, while LPIPS/DISTS fall from 0.1337/0.1740 to 0.0962/0.1606. Complete canvas and BBox vectors accompany the paper.

The source-only control in Table 3 also reveals a region-dependent trade-off. On the same three scans, source-only raises canvas PSNR by 0.18 dB and lowers DISTS by 0.0112, whereas anchoring raises SSIM by 0.0012 and lowers LPIPS by 0.0028. BBox LPIPS has the opposite ordering. The outputs are unchanged; full-canvas and cropped perceptual evaluation use different spatial domains and resizing contexts.

Figures 9 and 10 complement Figure 3 with four scans. Full-body views expose silhouettes, and shared upper-body crops reveal appearance differences. The displayed metrics evaluate the entire target BBox rather than the enlarged patch; aggregate comparisons use all forty target cameras.

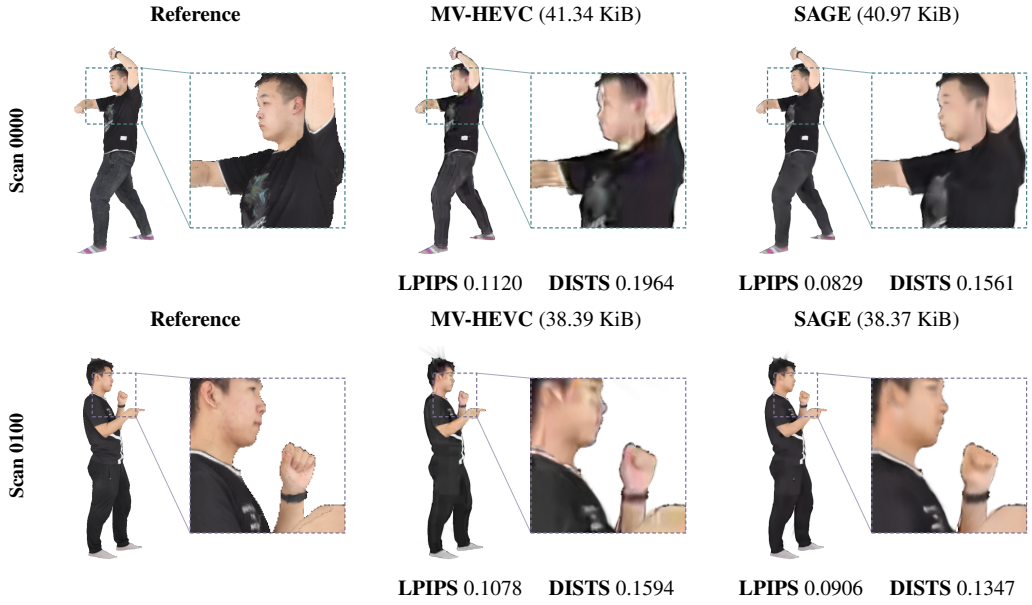


Figure 9: **THuman2.0 system transfer.** Scan 0000 uses camera 11 and scan 0100 uses camera 17. Each panel pairs a full-body rendering with a shared upper-body crop. Method labels give actual complete sizes; LPIPS and DISTS evaluate the entire target BBox. White backgrounds use each prediction's own foreground support.

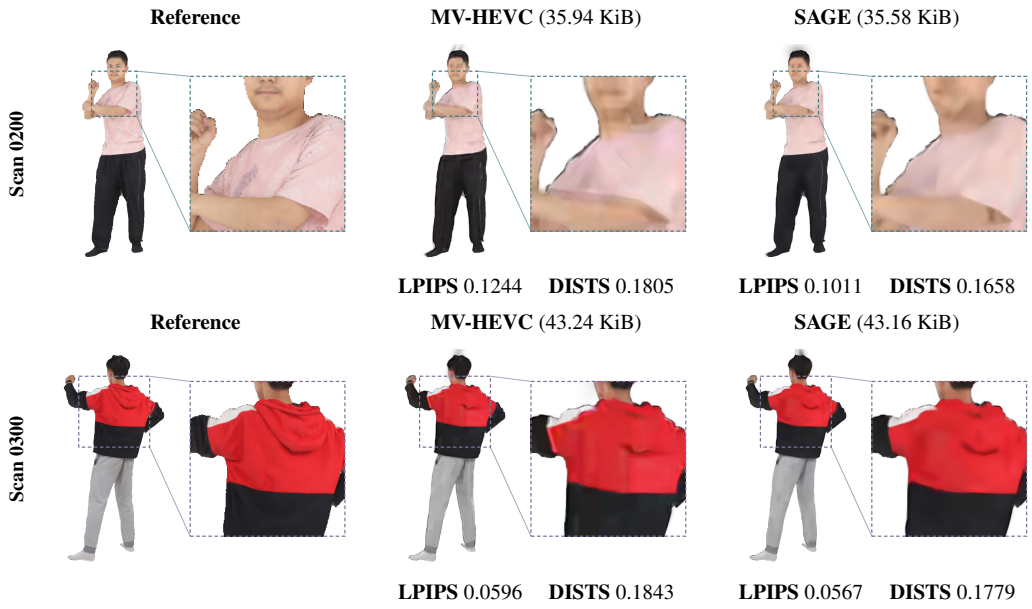


Figure 10: **Additional THuman2.0 scans.** Scans 0200 and 0300 at camera 17, using the same visualization and BBox metric protocol as Figure 9. Fixed camera selection covers different clothing, poses, and body orientations.

B.4 Optional Appearance Refinement

The Gaussian-only result is the primary reconstruction output. The optional support-conditioned RGB module does not participate in scene reconstruction or alter the optimized Gaussian representation. Full output mixes fidelity-oriented and perception-oriented predictions, $\hat{I}_j^F$ and $\hat{I}_j^P$, at query camera j:

$$\hat{I}_j^A = \mathrm{clip}\Big(\hat{I}_j^F + \hat{A}_j g_\phi(\mathbf{f}_j)(\hat{I}_j^P - \hat{I}_j^F), 0, 1\Big). \tag{15}$$

The fixed spatial mixer g_ϕ outputs a gate in $[0, 1]$. Its sixteen-channel input $\mathbf{f}_j$ concatenates Gaussian RGB, generated RGB, fidelity RGB, perception RGB, absolute expert disagreement, and alpha, in that order. Products are elementwise and clipping restricts RGB to $[0, 1]$. Alpha modulates the fidelity-to-perception increment; Gaussian parameters remain unchanged. Teacher-refined Gaussian output instead updates the scene through 31k additional teacher-guided steps, for a total of 62k.

The matched Full comparison in Table 12 uses five captures from two identities and sixteen rate–capture groups. Both branches share reconstruction rules and frozen appearance networks, generating support from their own decoded RGB. Scene construction uses 5k initialization, 25k base, and 1k terminal updates. MV-HEVC values are interpolated between measured complete sizes. We average views within rates, rates within captures, and captures within identities. SAGE improves all four means for both identities under this common output function.

Table 12: **Matched Full outputs.** Identity-equal means over two identities and five captures. Gains denote PSNR/SSIM increases and LPIPS/DISTS reductions.

Statistic	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
SAGE Full	**28.79**	**0.9626**	**0.0299**	**0.0929**
MV-HEVC Full	28.69	0.9601	0.0319	0.1015
Gain	+0.10	+0.0025	0.0020	0.0086

Before Full rendering, nonconstant spherical-harmonic coefficients are scaled by 0.8. Each 172,708-parameter expert receives Gaussian RGB, generated-minus-Gaussian RGB, alpha, a five-pixel alpha boundary, and mean RGB disagreement. A 48-channel stem and four residual blocks (dilations 1, 2, 4, 1) predict a $0.12\tanh(\cdot)$ residual and sigmoid gate. Alpha modulates the residual, whose deployed increment is scaled by 0.20. The spatial mixer adds two 32-channel convolutions and a sigmoid head; the full module has 359,721 parameters. Networks use eight-group normalization and sigmoid linear unit (SiLU) activation. These fixed operations define $\hat{I}^F, \hat{I}^P$; the core Gaussian-only output directly renders its checkpoint.

Training and checkpoint selection use three development identities, excluding the five-unit identities *0013_01* and *0023_06*. The fidelity expert is selected at step 4,000 of 12,000, the perception expert at 2,000 of 3,000, and the spatial mixer trains for 3,000 steps on 96 views with 24 validation views. Fixed experts supply its inputs; target masks are used only for evaluation. Sixteen SAGE states and thirty-two MV-HEVC brackets yield 640 paired rate–view measurements over 200 unique references.

The Full-output comparison in Table 12 uses the same frozen appearance networks for both coded frontends. The main evaluation uses one spatial-mixer checkpoint selected on validation views; query costs are reported separately in Appendix B.6.1.

The experts combine image, foreground, structural, edge, and perceptual losses with foreground-harm penalties. Spatial mixing additionally regularizes gate magnitude and variation. Validation selection uses $\Delta P + 50\Delta S + 20(-\Delta L) + 10(-\Delta D)$ for PSNR, SSIM, LPIPS, and DISTS changes. The accompanying method notes give all loss coefficients, architectures, and checkpoint rules.

B.5 Compression Baselines

We use eight source cameras and forty targets per task. VVC uses the VVC Test Model (VTM), version 24, on an ordered random-access YUV 4:2:0 sequence. MV-HEVC uses the HEVC test model with multiview extensions (HTM), version 16.3, with groups $(01, 07, 13)$, $(19, 25, 31)$, and $(37, 43)$. Learned stereo frontends code four adjacent source pairs. All complete packages include reconstruction conditions. Table 13 isolates the source-codec effect with the same scene decoder.

Table 13: **Coding frontends under a common receiver.** VVC minus MV-HEVC uses per-view log-byte alignment over three rates per identity and all 360 paired views. Positive changes indicate improved quality from VVC-coded evidence.

Identity	Rates	ΔPSNR	ΔSSIM	−ΔLPIPS	−ΔDISTS
0012_09	3	-0.06	+0.0014	+0.0006	+0.0013
0345_08	3	+0.12	+0.0026	+0.0012	+0.0068
0811_06	3	+0.20	+0.0030	+0.0019	+0.0120
Equal-ID mean	–	+0.09	+0.0024	+0.0012	+0.0067

After rate averaging, SSIM and both perceptual metrics improve on each identity; the first identity has a negative mean PSNR effect. Linear-byte alignment gives the same rounded aggregate PSNR change.

Appearance across camera queries. To complement the single-view comparisons in the main paper, Figure 11 renders the same low-budget scene from three calibrated cameras. The views expose different surface regions while retaining each method's fixed scene state. This comparison makes body coverage and silhouette behavior visible alongside the local appearance examined in Figure 8.

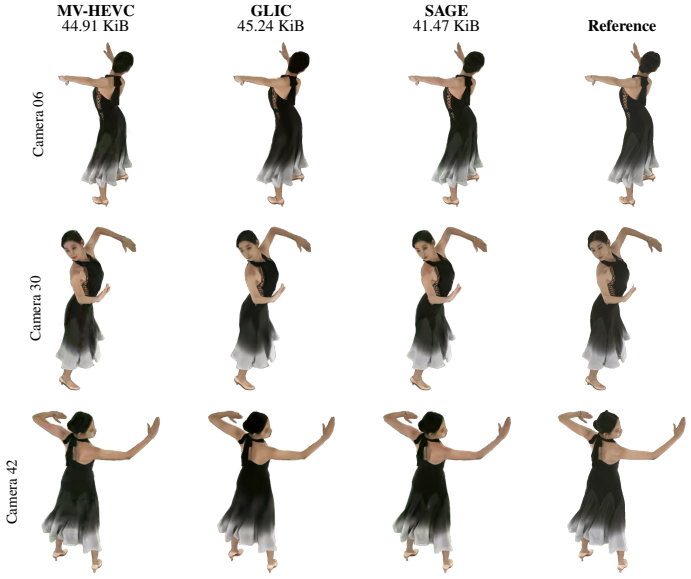


Figure 11: **Low-budget multiview appearance.** Rows show cameras 06, 30, and 42 of *0811_06/30*. All methods use the same 31k Gaussian states as Table 1; headers give actual complete sizes. Each row shares one crop, and predictions use their own alpha for white-background display.

B.5.1 Immersive texture and depth

TMIV is the Test Model for Immersive Video. Our TMIV 24 configuration uses VVenC 1.12 and VVdeC 2.3 at QPs 22, 28, 34, 40, 46, and 51. Source-silhouette carving supplies depth. Raw synthesis, occupancy, and hole filling isolate foreground processing; the third identity's pre-coding PSNR rises from 20.46 to 24.51 dB with the latter two steps. The companion retains the full three-identity, four-metric deployment atlas and all six measured rates per variant.

B.5.2 Multi-rate Learned-Codec Comparison

The multi-rate comparison covers three ECSIC settings and three SAGE source settings for each of three development identities. All eighteen reconstructions use a 31k-step Gaussian receiver and forty target cameras. Complete sizes include the archived conditions and container; each frontend generates support from its own decoded RGB. Figure 4 includes the *0811_06* subset, while the companion retains all 720 view-level metric vectors from the eighteen states. The 62k teacher study uses separate states.

At each identity, every measured SAGE package is smaller than ECSIC's lowest-rate package and improves all four canvas metrics. Relative to that measured reference, package reductions range from 7.09% to 38.12%, and PSNR differences range from 0.06 to 0.88 dB; the same four-metric ordering holds within the foreground bounding box. ECSIC's higher-rate points recover additional detail, with mixed perceptual changes at the highest setting. These percentages describe pointwise comparisons with the actual lowest-rate ECSIC package. The companion retains all eighteen operating points.

B.5.3 Other learned frontends

Stereo Image Compression with Latent Shifts and Stereo Attention (SASIC) (Wödlinger et al., 2022) uses its released Cityscapes model with four rectified pairs. Its entropy-coding adapter preserves the model's quantized reconstruction and scale lower bound. ECSIC uses official InStereo2k weights and a separate identified arithmetic-coding adapter. Both transmit inverse rectification. GLIC uses frozen official image models and actual entropy strings. Decoded inputs generate their own support and receive 31k Gaussian updates.

SASIC improves PSNR by 0.24–0.51 dB over the corresponding SAGE scenes at 2.05–2.41 times the complete size. It supplies one released rate model, hence one measured setting per identity. The companion measurements retain these points and all extended-computation configurations separately. Figure 4 presents the common-budget multirate comparison.

GLIC's six same-model source resolutions isolate spatial sampling from model changes. Lower-resolution outputs are restored to the calibrated canvas before support generation and reconstruction. Their LPIPS improvement slopes decrease from 1.54×10^{-3} to 2.55×10^{-5} per KiB, while DISTS retains an intermediate acceleration and PSNR reverses slightly at the highest resolution. Higher-budget native inputs favor GLIC in some comparisons; complete measurements remain in the companion.

Foreground processing. The common connectivity rule thresholds predicted alpha at 0.05, retains its largest eight-connected component, dilates it by eight pixels, and sets exterior color to black. Empty foreground leaves the image unchanged. This prediction-only step is separate from pure Gaussian rendering cost. An independent source-geometric control filters centers by agreement with four of eight dilated source silhouettes; it reduces large perceptual spikes but leaves several other metric reversals.

Independent ECSIC support repetitions use three prior seeds at four native rates, with fixed Gaussian seeds. Their mean LPIPS gains diminish, whereas highest-rate PSNR, SSIM, and DISTS reverse. These results distinguish execution variability and model transitions from a smooth idealized rate response.

B.5.4 General-Scene Receiver Transfer

We also transfer the released MVSGaussian (Liu et al., 2024) weights to 0811_06/30 without human-specific fitting. Each query uses the three nearest views selected from the same eight decoded sources and the same 70,089-byte package. Calibrated cameras and the source-derived hull set the depth range; inference uses 512-pixel inputs, 64/8 depth planes, and a 1024-pixel output. The method predicts a query-conditioned hybrid representation, whereas SAGE reuses an optimized Gaussian scene.

Over all forty targets, MVSGaussian obtains PSNR/SSIM/LPIPS/DISTS of 20.09/0.6447/0.2202/0.2971, compared with 30.61/0.9706/0.0192/0.0947 for the paired SAGE Gaussian state. Prepared-input output times are 21.05 and 1.90 ms, respectively, on the same H800. These timings exclude CPU input preparation and model loading; the transfer's full adapter timings are retained separately. The transferred model recovers the body but exhibits missing limb detail and foreground artifacts. This comparison characterizes one released general-scene receiver on the specified human capture, alongside the iteratively constructed representation.

B.6 Reconstruction and Query Efficiency

Gaussian-only reconstruction separates scene preparation from repeated queries. In common time units,

$$C_{\mathrm{Gaussian}}(N) = C_{\mathrm{setup}} + NC_{\mathrm{render}}, \tag{16}$$

where N counts queries, C_{render} is the time per render, and C_{setup} covers decoding, support generation, and scene optimization. Full additionally invokes experts and, when needed, query-specific support generation. The measured Full throughput below uses cached support on the prescribed camera grid.

B.6.1 Component profiling

Scene construction and rendering have different computational requirements. Table 14 reports representative components on one H800; these are separate measurements rather than one end-to-end trace. A complete cold-process Full measurement appears in Appendix B.6.3.

Table 14: **Construction and query costs.** Measurements use one H800; memory is in gibibytes (GiB). Queries share a five-unit, batch-one setup and framework-allocated memory; construction uses process-level memory. Full includes both experts and spatial mixing, with generated support cached.

Operation	Wall time	Peak GPU memory
Source decoding	0.58 s	Not measured
Support generation, 48 views	174.96 s	33.24 GiB
Gaussian construction, 31k	468.31 s	1.93 GiB
Gaussian query	1.63 ms/view	0.56 GiB
Full cache-hit query	119.01 ms/view	1.50 GiB

Across five captures, 2,000 synchronized queries per mode at batch one and float32 precision yield 614.85 views/s for Gaussian output and 8.40 views/s for Full, calculated as total queries divided by elapsed time. Separate VTM RGB/skeleton encoding takes 613.23 s serially or 420.53 s with two concurrent streams, excluding prediction, candidate search, and pose extraction.

The two-capture development search (*0013_01/100*, *0023_06/100*) takes 2639.28 s: 1135.12 s for sixteen source-cycle calls, 47.74 s for scoring, and 1455.92 s for ten encode/decode jobs. Timing starts with prepared images and conditions; subsequent transfers reuse the frozen template. Parallel execution details accompany the implementation notes.

B.6.2 Query Cost and Decoder Storage

On one fixed 1023×1023 capture (*0013_01*, time 149), synchronized H800 profiling measures 54.05 ms per expert and 9.03 ms for spatial mixing, using GPU-resident inputs, batch one, 32-bit floating-point precision (FP32), and three blocks of forty queries. Independently timed Full cache-hit queries take 118.89 ms, including rendering and support transfer. Split execution reproduces Full RGB exactly. Experts dominate query computation; Table 14 reports the separate five-capture aggregate.

Across five tasks, Gaussian attributes occupy 6.32–9.58 mebibytes (MiB). Forty support images require 119.77 MiB as unsigned 8-bit integers or 479.06 MiB in 32-bit floating point (FP32); shared query networks occupy 1.37 MiB, excluding Diffuman4D. These quantities measure decoder storage, separately from transmitted bytes.

B.6.3 End-to-end cold-process reconstruction

Three fresh-process executions decode the same 73,733-byte VVC RGB-and-condition payload for *0012_09*, generate a fresh 48-view Diffuman4D support grid, construct the source-only visual hull, perform 30k Gaussian and 1k terminal updates, and load the frozen experts and spatial mixer. Full output uses 0.8 spherical-harmonic (SH) stabilization and 0.2 expert scaling.

An outer monotonic clock measures decoding through the first synchronized Full image and through forty outputs, including weight loading and startup with the operating-system page cache retained. Queries use support-grid cameras, batch one, and native 1024×1024 resolution. Across the three executions, PSNR spans 29.69–29.73 dB, SSIM rounds to 0.9723, LPIPS spans 0.0163–0.0164, and DISTS spans 0.1041–0.1047. The companion binds each complete quality vector to its measured execution time.

The median first-output latency is 678.7 s (range 666.5–682.2 s). Across 360 timed loaded-scene queries, aggregate throughput is 8.47 views/s, computed from total queries divided by total syn-

chronized wall time. Resource sampling detects no other GPU process during these repetitions. The experiment measures known-grid Full reconstruction. It does not establish the latency or quality of a new-camera support-generation path.

B.7 Human-Region Analysis

At the lowest anchor, five archived Gaussian captures use 2.73% fewer bytes than the larger MV-HEVC bracket. Identity–time-equal interior PSNR is 22.52 versus 22.43 dB, eight-pixel-boundary PSNR 18.28 versus 18.26 dB, silhouette intersection over union 0.9710 versus 0.9708, and boundary F1 0.9531 versus 0.9513. A separate nine-view diagnostic trades 0.13 dB lower BBox PSNR for better SSIM and perceptual scores.

Boundary F1 thresholds alpha and reference masks at 0.5 and extracts contours by one-pixel erosion. Precision/recall use three-pixel square tolerance; either empty contour gives zero. Scores average forty views, then five captures equally.

B.8 Identity-Level Effects

The 640 rate–view pairs comprise 200 unique images and two identities. Views, rates, times, and identities are averaged hierarchically. Table 15 gives both identity vectors and their mean. Resampling two identities with replacement yields only three means: either vector or their midpoint, with probabilities $1/4$, $1/4$, and $1/2$. Identity effects are descriptive; execution repeats characterize pipeline variability.

Table 15: **Identity-equal Full-output effects.** Time and rate means are nested within each identity; LPIPS/DISTS columns report reductions. The two identity-specific vectors and their equal-weight mean describe the evaluated captures.

Identity	Capture instants	ΔPSNR $\uparrow$	ΔSSIM $\uparrow$	$-\Delta$LPIPS $\uparrow$	$-\Delta$DISTS $\uparrow$
0013_01	3	0.07	0.0026	0.0027	0.0095
0023_06	2	0.13	0.0023	0.0014	0.0077
Equal-ID mean	5	0.10	0.0025	0.0020	0.0086

Pooling gives 28.94/28.82 dB for SAGE/MV-HEVC because time and rate coverage differ across identities. Historical within-capture resampling uses identity–time units, rates, and views. The identity-equal PSNR gain is 0.1016 dB with linear-byte alignment and 0.1000 dB with log-byte alignment.

B.9 Gaussian Refinement Studies

Fixed-input development controls isolate supervision from source coding. To examine an extended optimization budget, Table 16 compares teacher-refined Gaussian outputs after 62k updates, averaged over three seeds and forty targets per identity. Both methods receive the extended budget; MV-HEVC uses its preselected larger measured package. At smaller sizes, SAGE improves SSIM and LPIPS on all three identities. PSNR improves on two and ties at displayed precision on the third, whose DISTS favors MV-HEVC.

Table 16: **Teacher-refined reconstruction.** Canvas means over three complete 62k runs per identity and method. Sizes are complete measured packages; bold marks best quality, including displayed ties.

Identity	Method	KiB	PSNR	SSIM	LPIPS	DISTS
0012_09	MV-HEVC	45.51	**30.58**	0.9753	0.0144	**0.0948**
	SAGE	42.42	**30.58**	**0.9766**	**0.0143**	0.0949
0345_08	MV-HEVC	39.78	31.21	0.9779	0.0200	0.0902
	SAGE	37.08	**31.37**	**0.9795**	**0.0163**	**0.0854**
0811_06	MV-HEVC	44.91	31.29	0.9722	0.0190	0.1070
	SAGE	41.47	**31.47**	**0.9747**	**0.0179**	**0.0966**

Teacher and regional controls. A frozen PointSplat teacher (Guo et al., 2026) consumes each branch's decoded sources and poses. Additional 31k updates alternate source/teacher supervision at fixed population with parent color/alpha preservation. Detached source-reprojection confidence does not outperform constant-mass weighting on canvas means.

Forty-five executions across three seeds and identities give canvas/head PSNR gains of 0.22–0.40/0.29–0.43 dB, but a lowest-rate face change of -0.25 ± 0.34 dB. All baseline seeds are retained. Head-prioritized patches reduce canvas PSNR by 0.10 dB. Normalized densification yields 25.2% fewer primitives and $+0.04$ dB, with DISTS worse by 0.0011. Upper-region teacher supervision improves perceptual means while reducing PSNR/SSIM. The companion retains all structural, regional and rate-specific controls, complete metric vectors, and 24,000 synchronized query measurements.

C Analysis and Outlook

Deployment. Twelve self-contained LightGaussian streams use 656.2–1,169.7 KiB, with PSNR 29.65–29.90 dB and DISTS 0.1226–0.1007, including reversals. SAGE/MV-HEVC instead transmit observations for shared-prior reconstruction.

Rate–distortion summaries. Figure 4 summarizes 26 native 31k states: 6 SAGE, five constant-support, six MV-HEVC, four GLIC, and five ECSIC settings. Metric-wise upper hulls impose positive, nonincreasing directed slopes; raw measurements retain reversals. Numerical comparisons use all measured vectors.

Integrated rate–distortion comparison. We extend the native-resolution, 31k Gaussian-only curve on *0811_06/30* with three independently encoded and reconstructed lower-rate SAGE points. Their complete packages are 44,239, 45,691, and 47,515 bytes at QP 41, 39, and 37, respectively. The corresponding canvas PSNRs are 29.802, 30.066, and 30.163 dB. Each point uses its own decoded sources, generated support, and Gaussian state; forty target views and all four original quality metrics are retained in the accompanying records.

Table 17 reports Bjøntegaard-delta rate (BD-rate) from piecewise-linear curves in log complete bytes (Bjøntegaard, 2001). For BD-rate, we invert only the longest common quality interval on which each *raw* curve has exactly one crossing; the rule and 0.10 dB minimum were fixed before evaluating the new points. Negative values indicate fewer SAGE bytes at matched PSNR within the stated interval. All nine SAGE, six MV-HEVC, four GLIC, and five ECSIC working points are retained, including non-monotone segments outside the integration domain. The piecewise-linear estimates characterize the locally invertible quality ranges of this capture.

Table 17: **Measured-curve rate–distortion integration.** BD-rate uses the common quality interval shown for canvas; BBox uses its corresponding region-specific interval. Positive ΔPSNR favors SAGE over a shared complete-rate interval; NA means none exists.

Baseline	Canvas PSNR interval	Canvas BD-rate	BBox BD-rate	Canvas ΔPSNR
MV-HEVC	29.80–30.15	-47.64%	-48.04%	$+0.801$
GLIC	30.04–30.32	-27.10%	-27.16%	$+0.386$
ECSIC	29.80–30.40	-61.97%	-61.47%	NA

The shared-rate intervals are 67.69–83.03 KiB against MV-HEVC and 57.71–76.58 KiB against GLIC. ECSIC has a shared quality interval but no shared measured rate interval, so its BD-rate is defined here while its integrated ΔPSNR is NA. BBox BD-rate uses 25.97–26.30, 26.19–26.47, and 25.97–26.56 dB for MV-HEVC, GLIC, and ECSIC, respectively.

C.1 Queries Outside the Construction Support Grid

To test whether source anchoring also benefits cameras absent from the generated support, we reserve eight of the forty target cameras on THuman2.0 scans *0000* and *0400*. Their indices are 3, 9, 15, 21, 27, 33, 39, and 45. The original eight sources remain fixed. We regenerate support on the other

Table 18: **Support-grid-held-out queries.** BBox quality from direct Gaussian rendering after 31k updates; eight reserved cameras per scan.

Scan	Rule	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
0000	Uniform	25.37	0.8808	0.1310	0.1723
0000	Anchored	27.98	0.9053	0.0984	0.1611
0400	Uniform	21.82	0.8783	0.1229	0.1799
0400	Anchored	23.79	0.8980	0.0874	0.1662
Mean	Uniform	23.59	0.8796	0.1269	0.1761
Mean	Anchored	25.88	0.9017	0.0929	0.1636

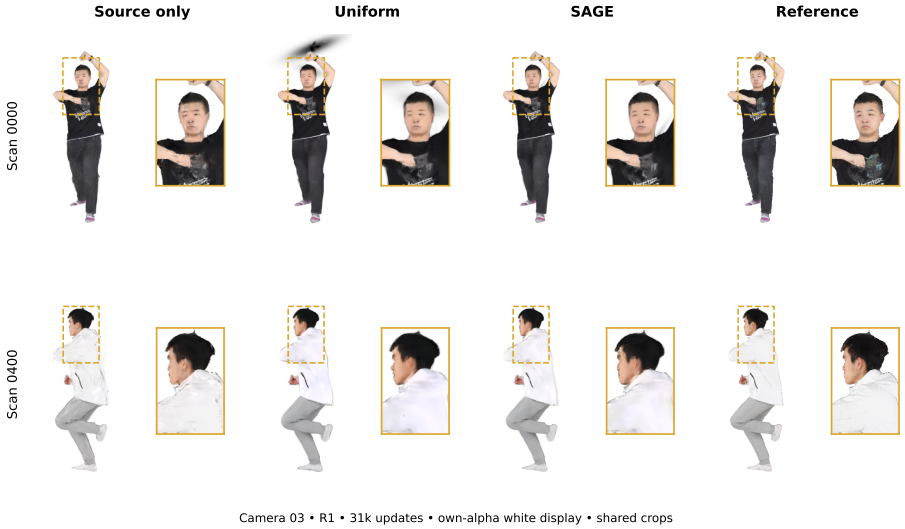


Figure 12: **Reconstruction beyond the support grid.** Both scans use reserved camera 3. Source-only, uniform, and SAGE show direct 31k Gaussian renders from the same coded observations; the reference and each prediction use their own silhouette/alpha for white-background display. Identical body crops and head insets expose appearance and silhouette differences.

thirty-two targets and exclude the reserved cameras from generated RGB and scene supervision. Camera calibration and the original full-capture pose conditions remain available.

Uniform and anchored reconstruction share the complete stream, source-hull initialization, generated support, auxiliary losses, seed, and 31k update budget. Uniform sampling uses eight source and thirty-two support entries; anchoring repeats each source four times and applies source photometric weight 1.8. Thus, the source sampling probabilities are $8/40$ and $32/64$, respectively. Both branches are evaluated after scene construction is complete. Table 18 reports all eight excluded cameras, averaged within each scan and then equally across scans.

On the eight excluded cameras per scan, source anchoring changes scan-equal BBox PSNR by +2.29 dB and SSIM by +0.0221; the directed LPIPS and DISTS reductions are +0.0340 and +0.0125, respectively. These queries directly render the optimized scene without generating support at the queried cameras. Figure 12 visualizes camera 3 on both scans, now including source-only reconstruction from the same coded observations. Its upper-body detail uses a shared crop across methods. The companion measurements retain all thirty-two supported-camera results. Because the two camera groups occupy different positions, we compare supervision rules within the same reserved group.

Generating support for a new query. Full output additionally consumes generated RGB at the requested camera. We measure this path on the anchored *0000* scene, requesting camera 3 after construction without its support. The scene and appearance networks are loaded before timing. Each request launches a fresh prior process that generates twelve fixed target views jointly, in-

cluding the requested view; timing ends after the first Full RGB output is synchronized. Prior weight loading, generation, support reading, rendering, both experts, and spatial mixing are included. Scene construction is measured separately.

Across three fresh prior processes, request-to-first Full output takes 54.76 s on average (range 52.31–58.07 s). After support is cached, ninety batch-one queries take 118.15 ms per query, computed from total elapsed time. Measurements use one H800, half-precision prior inference, and float32 query networks. Each repetition starts a fresh prior process while retaining the operating system's page cache; five query warm-ups precede thirty cache-hit measurements. The query uses an existing calibrated camera omitted from construction support. The measured generation group comprises cameras 3, 9, 15, 21, 27, 33, 39, 45, 1, 7, 13, and 19, conditioned on the original eight sources. Its joint generation time is charged to the first requested output rather than divided by the number of generated images.

C.2 Repeated Reconstruction on Held-Out Cameras

To test sensitivity to stochastic scene construction, we repeat the complete support-generation and reconstruction path on scans *0000* and *0400*. Their fixed packages contain 41,955 and 42,232 bytes, respectively. Each execution retains eight sources, thirty-two support cameras, and eight excluded query cameras. Within a repetition, uniform and anchored reconstruction share generated support and initial Gaussian parameters; each runs 30k base and 1k terminal updates.

Table 19 reports paired effects on the excluded cameras. For each scan, the prior/base/terminal seed tuple is (20260822, 20260812, 20260814) for R1. R2 and R3 use 20260923 and 20260924, respectively, for all three stages. Support is regenerated for each tuple before either reconstruction branch is scored.

Table 19: **Full-pipeline execution variability.** Anchored minus uniform BBox effects on eight held-out cameras per scan. Perceptual differences are reversed: positive values indicate improvement. Standard deviations describe three executions within each scan.

Scan	Repetition	ΔPSNR ↑	ΔSSIM ↑	$-\Delta$LPIPS ↑	$-\Delta$DISTS ↑
0000	R1	2.607	0.0245	0.0326	0.0113
	R2	2.647	0.0232	0.0340	0.0113
	R3	2.689	0.0243	0.0330	0.0106
	Mean	2.648	0.0240	0.0332	0.0111
	Sample SD	0.041	0.0007	0.0007	0.0004
0400	R1	1.975	0.0196	0.0355	0.0137
	R2	2.129	0.0167	0.0342	0.0181
	R3	1.992	0.0192	0.0345	0.0156
	Mean	2.032	0.0185	0.0347	0.0158
	Sample SD	0.084	0.0016	0.0007	0.0022
Equal-scan mean		2.340	0.0213	0.0340	0.0134

Averaging repetitions within each scan and then weighting the two scans equally gives a BBox PSNR change of +2.340 dB, an SSIM change of +0.0213, and LPIPS/DISTS reductions of +0.0340/+0.0134. All four directed metrics improve in all six paired executions. The within-scan sample standard deviations of the PSNR gain are 0.041 and 0.084 dB. Repeated generation and reconstruction characterize execution variability on these two scans; the reference set remains sixteen held-out images.

The source-only receiver keeps the source-hull initialization and 31k updates, but uses only decoded RGB and silhouettes for supervision. Figure 12 shows this control alongside uniform and anchored reconstruction. Across the two scans, anchoring improves PSNR and SSIM on average, while source-only has lower LPIPS and DISTS; the complete view-level measurements accompany the paper.

C.3 Input-Rate Interaction

To examine whether source anchoring acts specifically on codec distortion, we compare the fixed, non-annealed uniform and anchored 31k Gaussian recipes using two source conditions on *0811_06/30*. The first uses the masked images before VTM encoding; the second uses their QP 39 reconstructions in a 45,691-byte complete package. Each condition generates its own support and source-hull

initialization, while the source cameras, prior weights, target cameras, optimization budget, and scoring rules remain fixed. The pre-codec images serve as a high-fidelity input control. Table 20 reports paired effects; only the QP 39 row is a transmitted RD working point.

Table 20: **Anchoring by source-input condition.** Directed canvas changes from uniform to anchored reconstruction across forty targets; positive LPIPS/DISTS reductions indicate improvement.

Input	ΔPSNR	ΔSSIM	LPIPS reduction	DISTS reduction
Pre-codec	$+1.684$	$+0.00581$	$+0.00601$	$+0.01386$
QP 39 decoded	$+1.164$	$+0.00431$	-0.00960	$+0.00849$

Anchoring improves PSNR, SSIM, and DISTS under both input conditions. Its LPIPS direction changes with compression: the paired mean improves for pre-codec sources and worsens for QP 39 sources. Relative to the high-fidelity control, compression changes the canvas anchoring effect by -0.521 dB PSNR, -0.00150 SSIM, -0.01561 LPIPS reduction, and -0.00537 DISTS reduction. The BBox comparison has the same directional interaction. These results separate the general benefit of observation-weighted reconstruction from its rate-dependent perceptual response on this capture.