PRESERVING IDENTITY: A NEAR-FREE LUNCH IN ULTRA-LOW-RATE FACE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

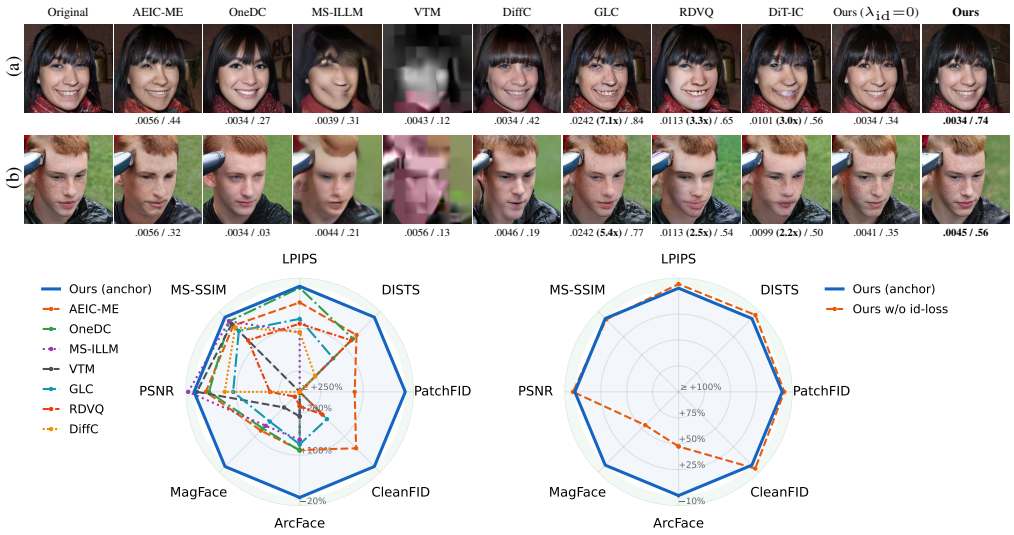
Generative codecs enable image compression at ultra-low bitrates, but on faces they exhibit *identity drift*: the reconstruction is perceptually plausible yet depicts a *different person*, a failure that standard fidelity/perceptual metrics do not capture. We build a face codec by repurposing a one-step diffusion face-restoration prior as the decoder of a compression pipeline. To this, we add an ArcFace-aware *identity loss* at training time. On a *rate–distortion* basis – with actual entropy-coded bitstreams – the loss yields a large identity gain (measured by *held-out* recognition networks) at an almost negligible cost on every other axis: a *near-free lunch*. On FFHQ, the same codec without the identity loss needs 50.9% more bitrate to match our identity, while the BD-rate on all distortion, perceptual, and distributional metrics stays within $\pm 5\%$; recent generative codecs need 116–225% more bitrate for our identity. Controlled experiments show the gain is *identity-specific structural alignment* (drift suppression) rather than generic feature regularization or high-frequency hallucination. The identity gain transfers across codecs and datasets, and the codec extends to a single resolution-agnostic model that keeps its identity lead on in-the-wild unaligned faces, and to face-video compression — a near-free route to identity-faithful face compression. Our code is included in the supplementary material.

1 INTRODUCTION

Neural image codecs with generative decoders now operate at *extreme-low* bitrates – well below 0.01 bits per pixel – by letting a strong generative prior synthesize plausible detail that a handful of transmitted bits can only loosely constrain (Mentzer et al., 2020; Muckley et al., 2023; Careil et al., 2024). On faces, however, this synthesis has a failure mode that ordinary compression metrics do not surface: the decoder confidently reconstructs a *different person*. The output is sharp, realistic, and in-distribution, yet its *identity* has drifted from the source. We call this **identity drift**, and at the bitrates where generative face codecs are most attractive it is severe – at ~ 0.004 bpp a reconstruction can look excellent while a face-recognition network no longer matches it to the original (Fig. 1). This goes unnoticed because the metrics used to certify low-rate reconstructions are *identity-blind*. Distortion (PSNR) rewards pixel-closeness to the reference, while perceptual (LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022)) and distributional (FID (Heusel et al., 2017)) scores reward a realistic, well-aligned face – yet none of them constrains *who* it depicts; a plausible but wrong face is, by these measures, a good reconstruction. Yet for the applications that motivate ultra-low-rate face coding – video calls, avatars, archival – identity is exactly the content that must survive: a distinct quality axis that must be optimized and measured explicitly.

We do so with a simple change to training – a face **identity loss** that penalizes the cosine distance between the reconstruction and the source in the embedding space of a frozen face-recognition network – and call the resulting codec **IdentiCodec**. Identity losses are established in face restoration and generation (Wang et al., 2021b; Zhou et al., 2022), and have been added to face codecs before (Chen & He, 2019; Xiao et al., 2022; Wang et al., 2025b). What is new in our study, to our knowledge, is threefold. **(Contribution 1)** We are the first to show that *at ultra-low bitrates* an identity loss becomes a *near-free lunch* – large identity gains at negligible cost to rate, distortion, and perceptual quality – whereas prior work at higher rates reports it as a trade-off (Chen & He, 2019; Xiao et al., 2022). **(Contribution 2)** We are the first to analyze *why* the ultra-low-rate regime reverses the usual tradeoff, with controlled experiments (wrong-identity/random-network, spatial, and spectral) that attribute the gain to identity-specific structural alignment. **(Contribution 3)** We are the first to bring an identity loss into a diffusion-based, ultra-low-rate *generative* codec: we repurpose a one-step

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.



(c) BD-rate vs. competitors (d) effect of the identity loss

Figure 1: **Identity drift at extreme-low bitrate, and a rate–distortion near-free lunch.** *(a,b)* Two subjects at comparable extreme-low bitrates (*bpp/MagFace*): competitors degrade (VTM, MS-ILLM, DiffC) or reconstruct the *wrong identity* (OneDC, AEIC-ME); ours without the loss ($\lambda_{\text{id}}=0$) drifts the same way, and the loss corrects it. GLC, RDVQ, and DiT-IC cannot reach this bitrate, so they are shown at their lowest rate with its exact multiple of ours marked (e.g. GLC needs 5–7×). *(c)* BD-rate radar, our model as the anchor: a competitor *inside* the ring needs more bitrate to match our quality (Table 1). *(d)* The same for our codec *without* the identity loss (zoomed to $+100\%$).

diffusion face-restoration prior (OSDFace (Wang et al., 2025a)) as its decoder – a base that already outperforms every published competitor on faces even *without* the loss – and adding the loss more than doubles the compression efficiency of the strongest competitors at matched identity (Table 1).

As Fig. 1 shows, at a comparable extreme-low bitrate our model preserves identity (measured with recognition networks *held out* from training) where every competing codec drifts or degrades. The trade-off one would expect – spend substantial rate or fidelity to gain identity – essentially never materializes: measured by all-metric BD-rate on actual bitstreams, identity is restored almost for free, at negligible cost on every other axis (quantified in Sec. 4.2). *Why does preserving identity come so cheap?* A controlled analysis attributes the gain to *identity-specific structural alignment*: at extreme-low rate the bitstream underdetermines facial geometry, and because identity *is* that geometry, pulling it back to the correct person corrects pixel- and structure-level errors as a by-product. Finally, both the objective and the codec generalize: the identity loss transplants onto other codecs, and our codec extends to a multi-resolution model and to a generative face-video compression pipeline.

2 Related Work

Neural and generative image compression. Learned image compression optimizes a rate–distortion objective end to end (Ballé et al., 2017; 2018; Minnen et al., 2018; Lee et al., 2019), and at extreme-low rate pairs it with generative (GAN or diffusion) decoders that restore realism: adversarial codecs (Mentzer et al., 2020; Muckley et al., 2023), diffusion codecs (Yang & Mandt, 2023; Careil et al., 2024), one-step or latent variants (Xue et al., 2025; Jia et al., 2024; Jiang et al., 2026), and single-step diffusion decoders specialized for compression (Chen et al., 2025; Jia et al., 2025; Li et al., 2025b; Vonderfecht & Liu, 2025); our base codec builds on a latent-diffusion backbone (Rombach et al., 2022). Exploiting the rate–distortion–perception trade-off (Blau & Michaeli, 2018; 2019), these look realistic at low rate but do not target face *identity*: at extreme-low rate they can render a convincing *different person* – the distinct axis we address.

Face-specific image and video compression. For *images*, face codecs bring in deep and generative priors – GAN-inversion onto a face manifold (Yang et al., 2025), StyleGAN (Karras et al., 2019) priors (Mao et al., 2024), token codebooks (Wang et al., 2025b), and Stable-Diffusion priors (Zhou et al., 2025). Closest to us, some already add a face-recognition-embedding loss (Chen & He, 2019;

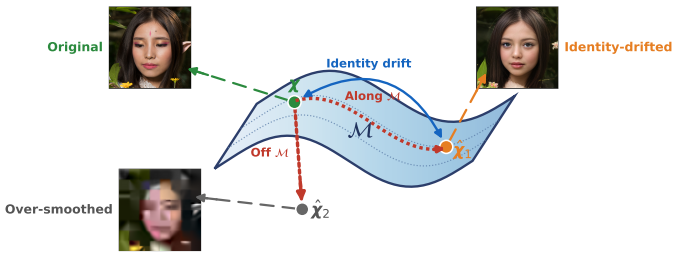


Figure 2: **Two failure modes on the face manifold $\mathcal{M}$. Identity drift** moves *along* $\mathcal{M}$ to a sharp but wrong-identity $\hat{\boldsymbol{x}}_1$ (OneDC, held-out MagFace 0.13); **over-smoothing** leaves $\mathcal{M}$ toward a blurry $\hat{\boldsymbol{x}}_2$ (VTM, MagFace 0.08). The drift (blue; Eq. 1) is our identity loss directly minimizes.

Xiao et al., 2022), but with a different aim and finding: they optimize downstream *recognition accuracy* at higher rates, where the identity term is a genuine cost (*lower* PSNR, *higher* rate), whereas we find it near-free at ultra-low rate and analyze its cause; Yang et al. (2025); Mao et al. (2024) instead target *scalable* human–machine coding. None releases code or weights at our operating point, so – as Zhou et al. (2025) does – we benchmark against the reproducible generative codecs of Sec. 4.1; our identity objective is orthogonal and transfers onto any of them. For *video*, generative face video coding (GFVC) transmits one intra key frame plus compact motion/keypoint parameters and animates it at the decoder (Siarohin et al., 2019; Wang et al., 2021a; Chen et al., 2022; Konuko et al., 2021; Chen et al., 2023; 2024). The key frame – the only image-compressed component – upper-bounds clip identity yet is usually coded by a conventional intra coder such as VVC (Bross et al., 2021); we replace only that key-frame codec with our identity-aware one (Sec. 6), leaving the animation backbone untouched.

Face recognition and identity-preserving generation. Angular-margin losses (Deng et al., 2019; Huang et al., 2020; Meng et al., 2021) learn discriminative face embeddings whose cosine similarity is the standard identity measure, and identity-embedding losses built on them are routine in blind face restoration and synthesis (Wang et al., 2021b; Zhou et al., 2022), where they steer a generator that sees the full (degraded) input. Compression differs: quality is bought with *transmitted bits*, so the added objective competes with the rate term and reshapes rate allocation – whether identity can improve while *barely* paying rate or fidelity (our near-free-lunch question) has no restoration equivalent; our contribution is that question's rate–distortion answer and mechanism, not the loss itself.

3 Method

Our IdentiCodec is a face codec obtained by adding a single identity term to the training objective of a strong extreme-low-rate image codec. We first formalize the problem we target and the identity loss that suppresses it, then argue it needs no side channel, and finally describe the base codec.

Identity drift and the training objective. In extreme-low-bitrate codecs, the compressed bitstream underdetermines the image, and the generative decoder must *synthesize* much of the missing detail. On faces, this produces a failure that distortion and perceptual metrics do not penalize, which we call **identity drift**: the decoder yields a sharp, realistic, in-distribution face whose *identity* differs from the source. We quantify it with a face-recognition network F_{eval},

$$\text{drift}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = 1 - \cos\big(F_{\text{eval}}(\hat{\boldsymbol{x}}),\, F_{\text{eval}}(\boldsymbol{x})\big), \tag{1}$$

the cosine distance between the reconstructed and original identities (larger is worse), where $\boldsymbol{x}$, $\hat{\boldsymbol{x}}$, F_{eval} are an input face, its reconstruction, and a recognition network held out from training (Sec. 4.1), respectively. Drift differs from *over-smoothing* – the classical distortion of low-rate coding – and the two are geometric *complements* on the face manifold. Drift is precisely the failure a generative decoder introduces in *escaping* over-smoothing: it resamples a sharp, on-manifold face that these same metrics reward even when it depicts the *wrong person* (Fig. 2). The two are thus orthogonal ways to lose identity – off-manifold blur versus on-manifold identity error.

To suppress this drift, we turn it into an explicit training target: we minimize the drift of Eq. 1, now on the frozen *training* network F_{train} (CurricularFace (Huang et al., 2020) IR-101). This training-network drift is our identity loss $\mathcal{L}_{\text{id}} = \text{drift}(\boldsymbol{x}, \hat{\boldsymbol{x}})$, added to the base codec's rate–

distortion objective $\mathcal{L}_{\text{base}}$ (**Base codec architecture** below; full form in Appendix B) with weight λ_{id}: $\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}}$. As a training loss, it adds no parameters and leaves inference unchanged.

A loss, not a side channel. One might go further and, *on top of* the loss, also *transmit* the identity: compress $F_{\text{train}}(\boldsymbol{x})$, send it as side information, and inject it into the decoder. We find this addition a poor trade. Even with a strong learned injector (a cross-attention identity adapter on every U-Net block), transmitting the identity barely moves quality beyond the loss alone: the BD-rate gain of this side-channel model over the loss-only model is under 1% on each of ArcFace, PSNR, and LPIPS (Appendix D). For that marginal gain, the channel adds an identity compressor/decompressor, a cross-attention block at every U-Net layer, and a face-recognition network on the encoder side — together $+7.9\%$ latency and $+7.6\%$ FLOPs over the base codec. The benefit is therefore better obtained from the training *objective* than from transmitted bits (full analysis in Appendix D).

Base codec architecture. Our codec builds on the AEIC (Zhang et al., 2025) framework, itself a one-step codec, but replaces its general-purpose one-step decoder with a one-step diffusion *face-restoration* prior: OSDFace (Wang et al., 2025a), a LoRA-adapted Stable-Diffusion-2.1 UNet (Rombach et al., 2022). We specialize this decoder for compression by training a *further* low-rank adapter (LoRA (Hu et al., 2022)) on the UNet with its input/output convolutions, keeping the OSDFace reference weights and the Stable-Diffusion backbone frozen, while the encoder and entropy model train end to end (full architecture and objective in Appendix A). This repurposing of a one-step face-restoration prior as the decoder of a compression pipeline is itself part of this work: even without the identity loss ($\lambda_{\text{id}}{=}0$) it already leads all published competitors (Sec. 4.2). This identity-free configuration is our ablation anchor, written **Ours** ($\lambda_{\text{id}}{=}0$) throughout.

4 Experiments

4.1 Setup

Datasets. We train and test on FFHQ (Karras et al., 2019) at 512^2 ($66{,}500$ training / $3{,}500$ held-out test images) and evaluate cross-dataset on a fixed $2{,}824$-image CelebA-HQ (Karras et al., 2018) test set at 512^2 that the model never sees during training (split details in Appendix B). **Metrics.** We report four families: distortion (PSNR, MS-SSIM (Wang et al., 2003)), perceptual (LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022)), distributional – *CleanFID* (Parmar et al., 2022) on full images (set-level realism, resizing-artifact-free) and *PatchFID* over 256^2 half-overlapping patches (Mentzer et al., 2020) (more sensitive to local texture) – and, as the axis of interest, *identity*: the cosine similarity between reconstruction and source in a recognition network's embedding space. To rule out circularity, the identity loss uses only CurricularFace (Huang et al., 2020) (F_{train}) while every reported identity number uses the *held-out* networks ArcFace (Deng et al., 2019) and MagFace (Meng et al., 2021), the two most standard held-out recognizers in face-identity evaluation (their mean, unless noted). Full curves are summarized by BD-rate (Bjøntegaard, 2001): the average extra bitrate a method needs to match a reference at equal quality (negative is better); unless noted, the reference is our model. **Competitors.** MS-ILLM (Muckley et al., 2023), RDVQ (Jiang et al., 2026), OneDC (Xue et al., 2025), GLC (Jia et al., 2024), DiT-IC (Shi et al., 2026), VTM (Bross et al., 2021) (all-intra), the released AEIC-ME (Zhang et al., 2025), and our own codec at $\lambda_{\text{id}}{=}0$ (Sec. 3), at matched extreme-low rates. We could *not* include any *face-specific* codec: to our knowledge none releases inference code or trained weights runnable at our extreme-low-rate operating point (Sec. 2), so — as in prior face-codec evaluations — the comparison is necessarily limited to these reproducible general and generative codecs. As the closest available stand-in for a purpose-built face codec, we *finetune every learned competitor that we can* on the same FFHQ face data — which also removes any training-domain confound — and report it alongside the released model (MS-ILLM and RDVQ); GLC, OneDC, and VTM cannot be retrained – no training code, missing auxiliary networks, and no trainable parameters, respectively. Exact models, bitrate accounting, and reproduction commands are in Appendix E. **Training configuration.** We first pretrain an *identity-agnostic* base codec that all operating points share; each is then finetuned from the shared model for 30k steps over the grid $\lambda_{\text{rate}} \in \{2, 4, 8, 16, 32\}$ — the rate–distortion weight that sets the operating point, with larger λ_{rate} spending fewer bits (roughly 0.035 down to 0.0045 bpp) — $\times\ \lambda_{\text{id}} \in \{0, 0.05, 0.1, 0.2\}$, where $\lambda_{\text{id}}{=}0$ is the anchor, and identity enters only through this finetuning loss. More details on implementation/training are provided in Appendix B.

4.2 RESULTS

Against external codecs. Our codec dominates prior work on identity at matched rate (Fig. 3): competitors need 116%–225% more bitrate to match it, and several never reach our perceptual/distributional quality at any rate ("–" in Table 1). MS-ILLM's (Muckley et al., 2023) higher PSNR (-13.9%) is the lone exception – regression toward the conditional mean, the *over-smoothing* corner opposite to identity drift (Sec. 3). Finetuning competitors on our FFHQ data does not close the identity gap (MS-ILLM and RDVQ (Jiang et al., 2026) still $\gtrsim 184\%$); even AEIC-ME (Zhang et al., 2025), which shares our entropy backbone, needs $+16.6\%$ more bitrate than our identity-loss-free Ours ($\lambda_{\text{id}}{=}0$) – a controlled gap (same FFHQ data, no identity loss) that isolates our base construction (the one-step OSDFace (Wang et al., 2025a) prior). That prior, not the training data, also drives our distributional lead: FFHQ-finetuning never improves a competitor's FID (unchanged for MS-ILLM, worse for RDVQ and AEIC-ME; Table 1). Both sources hold independently of the training data – Ours ($\lambda_{\text{id}}{=}0$) already beats every competitor and the identity loss widens the lead nearly for free – with Appendix H separating the two. Beyond the identity axis emphasized here, the full nine-metric rate–distortion curves (App. F, Fig. 13), the per-operating-point and iso-bpp breakdowns (App. B, Tables 4 and 7), and encode/decode latency and model size (App. R) are reported in the appendix.

A rate–distortion near-free lunch. That advantage comes from the identity loss, and it is nearly free. Across the operating range the loss substantially lifts recognition-space fidelity: at the extreme-low rate ($\lambda_{\text{rate}}{=}32$; 0.004–0.005 bpp) ArcFace (Deng et al., 2019) similarity rises from 0.44 ($\lambda_{\text{id}}{=}0$) to 0.59 ($\lambda_{\text{id}}{=}0.05$) and MagFace (Meng et al., 2021) from 0.45 to 0.62; at ~ 0.008 bpp the two climb to 0.73 and 0.76 (from 0.59 and 0.61), and the gap persists at every rate (Appendix K, Fig. 20). Yet it costs almost nothing elsewhere: with our model as the BD-rate anchor (Table 1), the same codec trained *without* the loss ($\lambda_{\text{id}}{=}0$) needs 50.9% more bitrate to match our identity, while on every distortion, perceptual, and distributional metric the difference stays within $\pm 5\%$ (1.3% in our favor on MS-SSIM (Wang et al., 2003) to 4.8% against us on DISTS (Ding et al., 2022)) – an order of magnitude below the identity gap. Identity is thus restored nearly for free (Sec. 5 explains why).

Verification, not just mean cosine. A standard $1{:}1$ verification protocol on the held-out ArcFace confirms the gain is operational, not cosmetic. Sweeping the impostor-acceptance threshold (genuine = reconstruction vs. its own source, impostor = all cross pairs), the loss lifts the genuine-match rate most where it matters — at extreme-low rate and strict operating points: at $\lambda_{\text{rate}}{=}32$ the true-positive rate (TPR; genuine acceptance) at a false-positive rate (FPR; impostor acceptance) of 10^{-5} rises from 0.60 ($\lambda_{\text{id}}{=}0$) to 0.86. The generative competitors collapse there (OneDC (Xue et al., 2025) and MS-ILLM (Muckley et al., 2023) fall to 0.07–0.18 at matched bitrate), rendering convincing but wrong identities, so that mean cosine alone understates how much the loss recovers at the operating points a recognizer actually uses (App. G).

Generality. The identity gain is not a property of one codec, dataset, or population. **(i) Cross-dataset:** the effect is not a property of the training distribution. On CelebA-HQ, which the model never sees, the identity gain reproduces: at ~ 0.004 bpp ArcFace (Deng et al., 2019) improves from 0.37 to 0.53 and MagFace (Meng et al., 2021) from 0.38 to 0.55 ($\lambda_{\text{id}}{=}0$ vs. ours), with the same ordering at higher rates. Our codec also still dominates every external competitor on this held-out set, leading on identity by a wide margin while staying competitive-to-best on the other metrics; the full rate–distortion comparison is in Appendix I, with qualitative reconstructions in Appendix L. **(ii) Across codecs:** the identity gain also transfers beyond our codec: the *same* objective helps other architectures. We transplant the *same* loss ($\lambda_{\text{id}}{=}0.1$) into two structurally different learned codecs and finetune each on FFHQ with the loss on versus off: MS-ILLM (Muckley et al., 2023), a GAN codec, and RDVQ (Jiang et al., 2026), a vector-quantized codec – neither of which shares our diffusion decoder. In each case we change nothing but the objective: no architecture, entropy-model, or training-hyperparameter edits, only the identity term added to the codec's existing loss. Each codec trained with our loss is scored by all-metric BD-rate against its own id-*off* anchor. Adding the loss yields a large identity improvement – the id-loss model reaches a given identity level with 17.4% (MS-ILLM) to 39.8% (RDVQ) *fewer* bits – while every other metric moves only modestly, spanning 10.5% better to 5.5% worse in BD-rate, and on several (e.g. PatchFID for both codecs) the id-loss *also* helps. Crucially, the effect does *not* require a generative decoder: we obtain the same behavior on HPCM (Li et al., 2025a), a purely MSE + rate-optimized (non-generative) LIC codec whose objective contains no perceptual or adversarial term. At matched identity gain

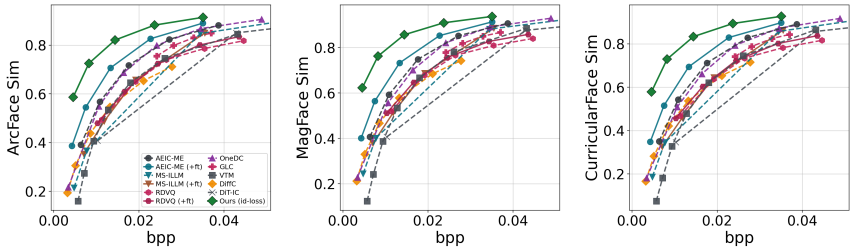


Figure 3: **Identity rate–distortion vs. competitors on FFHQ.** Cosine similarity vs. rate on the two *held-out* networks (ArcFace, MagFace) and, for reference, the training network (CurricularFace, excluded from reported numbers); ours (λ_{id}=0.05) vs. competitors (finetuned solid, public dashed). Full nine-metric RD and the identity-weight sweep are in App. F, K; GLC's bitrate is an estimate.

Table 1: **All-metric BD-rate (%) with our model as the anchor**, over the full FFHQ test set on entropy-coded bitstreams. "–" = never reaches our quality at any tested rate; Identity is the mean over the two *held-out* networks (ArcFace, MagFace); **best**/second per column marked. "+FFHQ-ft" rows are AEIC-ME, MS-ILLM, and RDVQ retrained on our FFHQ data.

Method	Identity	PSNR	MS-SSIM	LPIPS	DISTS	PatchFID	CleanFID
Ours (anchor)	**0.0**	0.0	**0.0**	0.0	0.0	0.0	0.0
Ours (λ_{id}=0)	50.9	−1.8	1.3	**−3.9**	**−4.8**	**−2.0**	**−4.6**
AEIC-ME	116.3	27.8	27.9	38.5	59.6	119.9	61.1
+FFHQ-ft	72.6	2.2	−0.2	34.3	106.1	–	523.3
OneDC	119.0	35.3	13.9	3.7	72.1	–	285.1
MS-ILLM	136.4	**−13.9**	13.5	109.3	338.3	–	370.4
+FFHQ-ft	184.0	−8.4	21.8	90.2	372.8	–	369.4
VTM	195.1	6.7	22.7	–	–	–	–
GLC	137.7	92.4	46.3	77.2	137.6	–	158.7
RDVQ	225.0	179.7	77.5	88.6	77.4	500.5	174.0
+FFHQ-ft	207.5	138.4	58.0	60.1	56.7	811.5	222.3
DiffC	259.6	72.5	34.1	108.4	198.3	–	–

Lower is better. **DiT-IC** (not tabulated): discrete quality levels give a single point in our rate range, so no reliable BD-rate; on the RD curves (Fig. 3) its identity is 0.41 vs. 0.78 at ≈0.01 bpp (≈ 2–3× the bitrate to match).

(λ_{id}=0.003, so the identity improvement is comparable in size to ours rather than over-pushed), the loss stays on the id-*off* rate–distortion frontier (−29.4% identity BD-rate at +1.5% PSNR, +0.4% MS-SSIM, +1.3% LPIPS) *and* improves every distributional/perceptual metric (DISTS −7.6%, PatchFID −24.2%, CleanFID −8.4%), the gains being largest at the lowest rates. So the near-free behavior is not specific to a generative decoder, and on a distortion-optimized codec the identity term additionally acts as a mild realism regularizer. The full BD-rate table (Table 13, three codecs) and per-metric RD curves are in Appendix J. **(iii) Across gender, age, and skin-tone:** the gain is also uniform across demographic subgroups. Stratifying the full FFHQ test set by gender, age, and skin tone (ITA), the held-out identity gain stays within 0.005 of the overall +0.093 (MagFace) for every subgroup, and the loss does not amplify pre-existing between-subgroup identity gaps (App. Q).

Identity-weight ablation. The identity weight λ_{id} is a dial on this trade, swept against Ours (λ_{id}=0) as the anchor. At our default λ_{id}=0.05 the identity BD-rate gain is −33.7% (equivalently, λ_{id}=0 needs +50.9% more rate, the reverse-anchor figure reported in Table 1) while every other axis stays within −1.3% (MS-SSIM (Wang et al., 2003), a gain) to +5.0% (DISTS (Ding et al., 2022)) – the near-free operating point. Increasing the weight gains more identity (−44.7% at 0.1, −53.8% at 0.2) at a progressively larger cost on the perceptual and distributional axes (up to +16% at 0.2), so the lunch is only near-free while the weight stays small (See Appendix K, Table 14). Even at these higher weights, though, the trade stays favorable when identity is the priority: the identity BD-rate improves by up to 53.8% against a perceptual/distributional penalty of at most ≈ 16%, so the identity gain outweighs the cost roughly threefold and the loss remains worthwhile well beyond the near-free regime. We default to λ_{id}=0.05 simply because it is where the penalty is negligible.

Complexity and identity efficiency. The identity edge does not come at a complexity cost. As a *one-step* decoder, our codec sidesteps the sampling overhead of multi-step diffusion codecs: at 512^2 its decode is a single U-Net forward (168 ms), whereas DiffEIC (Li et al., 2025b) (50-step) and DiffC (Vonderfecht & Liu, 2025) (27–50 U-Net forward passes) take 7.1 s and ≈ 1 s – 6–43× slower. Among *one-step* diffusion codecs it also sits on the best identity-versus-complexity fron-

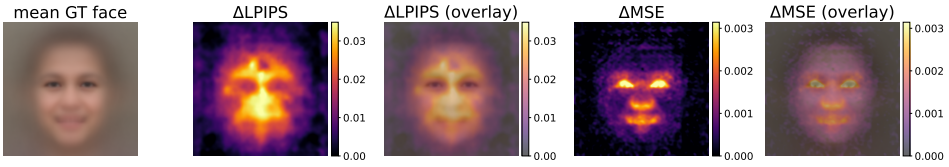


Figure 4: **Spatial attribution** (λ_{rate}=32, λ_{id}=0.1; FFHQ-3500, iso-λ_{rate}). Per-pixel improvement (anchor−id), averaged over the aligned test set; brighter means the identity loss reduces error more.

Table 2: **Quality improvement by super-region** (λ_{rate}=32, λ_{id}=0.1; FFHQ). Change from adding the identity loss, per super-region grouped by identity discriminativeness.

Super-region	Pixel frac.	ΔPSNR (dB)↑	ΔLPIPS↓
Inner identity (eyes/brows/nose/mouth)	6.1%	**+1.64**	**−0.029**
Face skin	23.1%	+1.08	−0.021
Periphery (ears/hair/hat)	28.3%	+0.14	−0.008
Background	28.4%	+0.04	−0.005

tier: at matched rate it attains far higher identity than OneDC (Xue et al., 2025) while decoding 3× faster and encoding ∼10× faster. Finally, the identity loss is training-only, so it adds *zero* inference cost: our model is the very same network as its identity-off anchor (Ours, λ_{id}=0) – identical parameters and latency – yet gains +0.11 ArcFace at a matched ultra-low bitrate. Full latency/size measurements, the identity-versus-complexity plot, and protocol are in Appendix R.

5 Why Is Identity Nearly Free?

The near-free gain of Sec. 4.2 poses a puzzle: raising recognition-space similarity would ordinarily trade against pixels, yet the identity loss lifts identity and fidelity *together*. We attribute this to **identity-specific structural alignment**: because identity *is* facial geometry, aligning a reconstruction to the correct person *is* an act of structural correction, so the fidelity that identity would otherwise cost is recovered by the same alignment. We run the analyses below at λ_{id}=0.1 — above our 0.05 default, where the improvement in the other metrics is easiest to see — in the λ_{rate}=32 extreme-low-rate cell where drift is largest. Two controls guard against reading the weight into the result. The effect is not a λ_{id}=0.1 artifact: Appendix C reproduces all three tests at the 0.05 default with the same trends. Nor is it the small extra rate the loss spends: matching the λ_{id}=0 anchor to the identity-loss model's *own* bitrate still leaves a +0.13 (λ_{id}=0.05)–+0.19 (0.1) held-out ArcFace (Deng et al., 2019) gain, so the extra 10–22% rate accounts for only 13–18% of the improvement and the mechanism carries the rest (Appendix C). The mechanism makes three predictions, each of which could have failed: the gain should (i) *concentrate on identity-bearing facial regions*; (ii) be *low/mid-frequency structural correction*, not high-frequency synthesis; and (iii) be *bound to identity correctness*, vanishing under a wrong or meaningless target. All three hold:

Test 1 — spatial attribution (where the fidelity gain lands). If the fidelity gain is structural alignment to identity, it should concentrate on the geometry that *carries* identity rather than spread uniformly. We parse each ground-truth face into four super-regions ordered by identity discriminativeness — *inner identity* (eyes/nose/mouth), *face skin*, *periphery* (ears/hair/neck/clothing), *background* — and measure the change from adding the loss (id − anchor, matched λ_{rate}) in LPIPS (Zhang et al., 2018) and PSNR per region over the 3,500-image test set (Fig. 4; per-region breakdown in Appendix C, Fig. 8). The improvement forms a clean *monotone gradient ordered by identity content* (Table 2): largest on inner-identity regions and falling off through face skin and periphery to background, with the nose alone gaining +2.2 dB — inner-identity regions improve ∼6× more than background in LPIPS and ∼40× in PSNR. This confirms prediction (i): the fidelity gain is reallocated to identity-defining regions, not spread uniformly.

Test 2 — frequency analysis (structural, not synthesized). Prediction (ii) concerns the *kind* of correction: structural alignment should reduce error at low/mid spatial frequencies (where facial geometry lives), whereas a mechanism working by *synthesizing* high-frequency detail would instead cut high-frequency error. Over the test set the radially-averaged error-energy reduction is overwhelmingly low/mid-frequency, −36.2% (low) and −15.7% (mid) versus +0.4% (high), out of −23.6% total (Fig. 9, Appendix C), confirming prediction (ii).

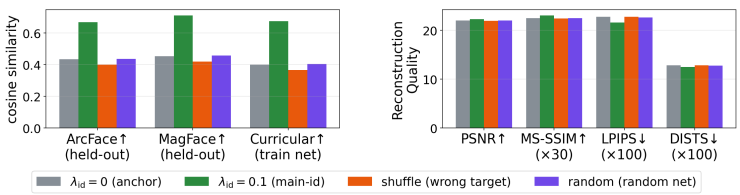


Figure 5: **Identity-specificity control** (λ_{rate}=32, λ_{id}=0.1; FFHQ, matched rate). Held-out identity (left) and reconstruction quality (right) for the correct-identity loss (*main-id*) against the wrong-target (*shuffle*) and random-network (*random*) controls and Ours (λ_{id}=0).

Test 3 — the specificity control (is the gain bound to identity?). The analyses above characterize the gain; this test pins down its cause by intervention. Prediction (iii) is that the whole effect — the identity gain *and* the fidelity moving with it — is driven by identity *content*: break its meaning and both should vanish. We retrain the λ_{rate}=32, λ_{id}=0.1 model with two controls that break it while holding all else fixed: *shuffle* (the same CurricularFace (Huang et al., 2020) network, *wrong* per-image target) and *random* (a randomly-initialized, meaningless feature space, correct pairing). The result is unambiguous (Fig. 5): with the correct target the loss lifts identity and reconstruction quality *together* (main-id beats the anchor in both panels), while both controls reproduce *neither* — identity falls back to the anchor (shuffle $\Delta{-}0.035$, random $\Delta{+}0.003$ over ArcFace (Deng et al., 2019) and MagFace (Meng et al., 2021)) and the quality edge disappears with it. Identity and fidelity thus improve as *one* identity-driven effect, appearing only for a correct target — exactly as expected if aligning identity *is* correcting structure.

Corollary: distributional quality follows. The loss also improves the distributional metrics at each operating point (PatchFID $2.89 \to 2.68$, CleanFID $9.07 \to 8.78$ at λ_{rate}=32; Appendix B, Table 4), but only faintly relative to the identity gain — an *emergent* consequence of per-image structural correction, not a set-level pull toward the manifold (quantified in Appendix C).

6 Extensions

Resolution-agnostic operation. A single multi-resolution finetune that injects the input resolution into a zero-initialised per-block *resolution FiLM* (App. O) spans the whole 256–1024 range at *no* transmitted cost, replacing a family of per-resolution codecs. On the full 3,500-image test set at practical bitrate it *matches* dedicated single-resolution codecs at the 512 anchor (PSNR within $\sim$0.3 dB) and *beats* them on every metric at 256 and 1024, with the identity lead over external codecs holding at all resolutions. The identity loss stays *near-free* here too: versus the same model without it (λ_{id}=0), held-out identity rises (up to $\sim$0.15 ArcFace, largest at low rate) while PSNR stays within 0.2 dB and the other metrics track the id-off model. Resolution conditioning adds only 5.2M parameters (0.47%) and a few milliseconds of latency (App. R).

In-the-wild large faces. The same model with *no* in-the-wild finetuning transfers to unconstrained web photographs. We evaluate on 3,619 single-face Flickr crops (Flickr-2W, causally independent of FFHQ, *zero* shared photo-ids) at their *native* 256–1024 resolution: a genuinely large-face, unaligned regime, unlike the small aligned crops of CPLFW/CFP. On orig-vs-reconstruction ArcFace identity our codec leads every baseline at matched bitrate: with our model as the BD-rate anchor, OneDC needs 32.3% more bits to match our identity (norm_crop; 19.7% under a fixed-landmark crop), AEIC-ME 40–54%, and VTM, MS-ILLM, GLC further behind. The identity loss secures this lead — the same model without it (λ_{id}=0) needs 22.7% more bits, yet still beats every competitor — and the advantage is largest exactly where faces are large and unaligned, with side-view cases holding identity while OneDC drifts to a different person. Full curves and qualitative examples are in Appendix P.

Face video via GFVC (intra replacement). Replacing only the key-frame codec — leaving the animation backbone (FOMM (Siarohin et al., 2019), an official JVET GFVC CTC anchor) untouched — carries our identity-aware intra coding into generative face video. The key frame is the sole image-compressed component of a GFVC clip, so its identity and texture propagate to every keypoint-rendered frame (Fig. 6): the VVC-intra key frame collapses into blocking and a wrong identity, whereas ours preserves both. Aggregated over the 33 CTC sequences (pooled RD, VVC-intra QP $\{22, 32, 42, 52\}$ anchor), the swap saves about 72% (LPIPS (Zhang et al., 2018)), 79% (DISTS (Ding et al., 2022)), and 78% (PatchFID (Mentzer et al., 2020)) of the *total* video bitrate at

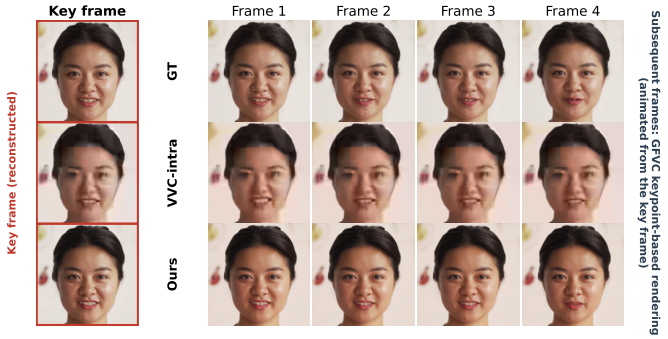


Figure 6: **GFVC key-frame replacement** (ClassC_002, official JVET CTC). Only the key frame is compressed – by VVC-intra (middle), by our IdentiCodec (bottom); the remaining frames are GFVC keypoint-based renderings animated from it. VVC-intra vs. our key frame scores ArcFace 0.36 / PatchFID 239@1.63 kbps vs. ArcFace 0.74 / PatchFID 37@1.80 kbps. Shown at 256×256.

matched perceptual quality (BD-rate). PSNR is essentially unchanged (≈ 18 dB) — as in Sec. 5, the gain is perceptual and identity, which PSNR does not register. Full curves over all 33 sequences and two further qualitative sequences are in Appendix M.

7 Limitations

The identity loss reduces but does not *eliminate* drift: at the extreme-low rate the hardest cases — unusual expression, lighting, or heavy accessories — still reconstruct as a perceptually plausible yet subtly *different* person (Appendix N, Fig. 27); even there the loss usually improves over the loss-free model, but a residual identity gap remains. Our identity objective supervises the *face*; everything else — background, hair accessories, clothing — is still synthesized by the generative decoder at a few thousandths of a bit per pixel and can be re-synthesized with altered content (Fig. 25(a), Appendix N: the cap logo becomes a different emblem while the face identity is preserved). This is generic to generative codecs at extreme-low rate, not an artifact of our loss: over the pure-background region our rate–distortion deviation is *lower* than AEIC-ME (Zhang et al., 2025)/RDVQ (Jiang et al., 2026), on par with OneDC (Xue et al., 2025), and essentially identical with and without the identity loss (App. N, Fig. 26) — so the objective adds no non-face hallucination of its own. The codec is thus identity-faithful but not pixel-faithful outside the face; region-aware supervision or rate allocation is a natural next step. In the face-video extension (Sec. 6) our contribution is the intra *key frame* only: the remaining frames come from the animation backbone's keypoint warping, which can distort the face under large pose or expression change (App. N, Fig. 25(b)) — a limitation of the GFVC backbone, not our codec. An *identity-conditioned backbone* that conditions inter-frame synthesis on the key-frame identity is a promising fix.

8 Conclusion

We presented a diffusion-based face codec for extreme-low-bitrate compression that leads all published competitors on identity, and showed that its key ingredient — a single ArcFace-aware identity loss, added to the training objective and not transmitted as a side channel — restores identity nearly for free: it brings a large identity advantage while every other quality axis moves by at most a few percent of rate, and improves every metric outright at each operating point. The loss targets *identity drift* — an on-manifold resampling to a sharp, realistic, but *wrong* person — the failure mode that distortion and perceptual metrics do not penalize at extreme-low-bitrate face compression. The gain transfers across codecs (MS-ILLM (Muckley et al., 2023), RDVQ (Jiang et al., 2026), HPCM (Li et al., 2025a), OneDC (Xue et al., 2025)), to an unseen dataset (CelebA-HQ), and to face video, where replacing only the GFVC key-frame codec propagates the identity gain through the entire clip. A controlled mechanism analysis attributes the gain to identity-specific *structural* alignment rather than generic deep-feature regularization: it collapses under wrong-identity and random-network controls, localizes to inner-face regions, and lives in low/mid frequencies rather than high-frequency synthesis. Because the objective is a drop-in loss that needs no architectural change, side information, or extra rate, it is a simple and broadly applicable route to making face codecs identity-faithful.

AI USE STATEMENT

We used generative AI tools (large language model assistants) in this work. The first table lists every task for which the ICLR 2027 AI policy requires disclosure; the second covers tasks with recommended disclosure.

Task with required disclosure	Used	Scope
Generate synthetic data sets	No	—
Help develop theoretical models or conceptual frameworks	No	the method (Sec. 3) and its mechanism analysis (Sec. 5) are the authors' own
Formulate mathematical claims	No	—
Provide critical ingredients for proving mathematical claims	No	—
Assist in the writing of proofs	No	—
Propose or refine hypotheses	No	the near-free hypothesis and the structural-alignment mechanism were designed by the authors
Design or provide feedback on research methodology or experiments	Yes (limited)	feedback on the evaluation and analysis protocols of Appendices E, C, O, P
Implement methods	No	the codec (diffusion-prior decoder, identity loss, training recipe) was designed and implemented by the authors; AI assisted only with code under their direction (second table)
Assist with translation	No	—
Clean and reformat dataset	Yes (limited)	data-preparation scripts (constructing the CelebA-HQ and in-the-wild test sets, format conversion) under the authors' direction
Support qualitative and thematic data analysis	No (N/A)	—
Interpret results	Yes (limited)	assisted the authors in reading intermediate results; all conclusions were drawn and verified by the authors

Task with recommended disclosure	Used	Scope
Formulate questions for surveys or interviews	No	—
Create or modify scientific figures or images	Yes	plotting scripts for result figures, under the authors' specification
Suggest experimental parameters	No	the rate/identity-weight grid was set by the authors
Create or edit software code	Yes	evaluation, analysis, and training-orchestration scripts; edits to training code under the authors' direction
Creation of artifacts	No	—
Draft parts of a research paper	Yes	drafts of text passages, later revised by the authors
Transcribe recordings of research material	No (N/A)	—
Summarize or analyse existing literature	Yes (limited)	summaries of the closest related methods for the availability audit (App. E)
Discover research topics or identify gaps	No	—
Brainstorming	Yes	discussion of analysis options with the authors
Sourcing/searching for information	Yes	verifying availability, step counts, and settings of related methods
Edit a research paper to improve readability	Yes	condensing and copy-editing
Identify relevant literature	No	—
Format references	No	—
Suggest a structure for a research paper	Yes (limited)	placement of individual paragraphs and figures
Propose a title or keywords for a research paper	No	—

We have reviewed all AI-assisted work: all code was reviewed and tested by the authors, every reported number was produced by the described pipelines on the stated data, and all AI-drafted text was checked against the experiments and the cited sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Every result in the paper is reproducible from public data and the released code. The training recipe is specified in Sec. 4.1: we train the encoder and entropy model end to end and, on the decoder side, finetune only a low-rank adapter on the UNet with its input/output convolutions, keeping the Stable-Diffusion backbone, the OSDFace (Wang et al., 2025a) reference weights, and the VAE frozen, for 30k steps at each operating point over the two-dimensional grid $\lambda_{\text{rate}} \in \{2, 4, 8, 16, 32\} \times \lambda_{\text{id}} \in \{0, 0.05, 0.1, 0.2\}$; the full hyperparameters (optimizer, learning-rate schedule, batch size, LoRA (Hu et al., 2022) rank, initialization) are in Appendix B (Table 3), whose Table 4 also lists the complete per-model configuration and metrics for all 20 models. The identity objective uses the frozen training network F_{train} (CurricularFace (Huang et al., 2020) IR-101), fed a fixed 0.75 center crop resized to 112×112 with $F_{\text{train}}(\boldsymbol{x})$ precomputed and cached; identity is *evaluated* only on the disjoint held-out networks ArcFace (Deng et al., 2019) (InsightFace `w600k`) and MagFace (Meng

et al., 2021), so no reported identity number sees its supervising network. Datasets and splits are as in Sec. 4.1: FFHQ-aligned 512^2 with a seeded $95/5$ shuffle into $66{,}500$ training and $3{,}500$ held-out test images (the same split for pretraining and finetuning), and a deterministic $2{,}824$-image CelebA-HQ cross-dataset test set (all $2{,}000$ validation images plus the first 824 training images, matching the official test-partition size). All methods are scored by one shared evaluation pipeline over the same nine metrics, and BD-rate is the Bjøntegaard integral on the $\log$-bitrate axis; the exact competitor models, bitrate sources, and reproduction commands are documented in Appendix E (Table 10). Our model and training code, the pretraining and identity-finetuning configurations, and the five pretrained checkpoints (each with a SHA-256 manifest) are provided for review as anonymized supplementary material and at an anonymous repository;[1] the complete evaluation and plotting scripts and the per-grid-cell configurations will accompany the de-anonymized public code release.

Ethics Statement

Face-identity reconstruction is inherently dual-use. The ability to preserve recognizable identity at extreme-low bitrate benefits bandwidth-constrained telepresence, archival, and accessibility, but the same fidelity could aid surveillance or, if misapplied to a generative decoder, the synthesis of misleading imagery. Our method restores the identity of the *source* subject being compressed rather than impersonating others, and it adds no identity side channel; still, we note that any high-fidelity face codec carries misuse risk and should be deployed with consent and provenance safeguards. All experiments use the publicly released FFHQ and CelebA-HQ datasets under their respective licenses, and we report identity on recognition networks disjoint from the one used in training to avoid overstating the effect.

References

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations (ICLR)*, 2017.

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 (ICLR)*, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, ITU-T Video Coding Experts Group (VCEG), Austin, TX, USA, 2001.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6228–6237, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 675–685, 2019.

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.

Marlène Careil, Matthew J. Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Chardon, I. Cretois, and C. Hourseau. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13(4):191–208, 1991.

Bolin Chen, Zhao Wang, Binzhe Li, Rongqun Lin, Shiqi Wang, and Yan Ye. Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression. In *2022 Data Compression Conference (DCC)*, pp. 13–22, 2022.

[1] `https://osf.io/9bz6p/?view_only=250e064dc0704032bb999991db57485b`

Bolin Chen, Jie Chen, Shiqi Wang, and Yan Ye. Generative face video coding techniques and standardization efforts: A review. *arXiv preprint arXiv:2311.02649*, 2023.

Bolin Chen, Yan Ye, Jie Chen, Ru-Ling Liao, Shanzhi Yin, Shiqi Wang, Kaifa Yang, Yue Li, Yiling Xu, Ye-Kui Wang, Shiv Gehlot, Guan-Ming Su, Peng Yin, Sean McCarthy, and Gary J. Sullivan. Standardizing generative face video compression using supplemental enhancement information. *arXiv preprint arXiv:2410.15105*, 2024.

Zheng Chen, Mingde Zhou, Jinpei Guo, Jiale Yuan, Yifei Ji, and Yulun Zhang. Steering one-step diffusion model with fidelity-rich decoder for fast image compression. *arXiv preprint arXiv:2508.04979*, 2025.

Zhibo Chen and Tianyu He. Learning based facial image compression with semantic fidelity metric. *Neurocomputing*, 2019. doi: 10.1016/j.neucom.2019.01.086.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

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.

Jarek Duda. Asymmetric numeral systems: Entropy coding combining speed of huffman coding with compression rate of arithmetic coding. *arXiv preprint arXiv:1311.2540*, 2013.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14771–14780, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26088–26098, 2024.

Zhaoyang Jia, Zihan Zheng, Naifu Xue, Jiahao Li, Bin Li, Zongyu Guo, Xiaoyi Zhang, Houqiang Li, and Yan Lu. Cod: A diffusion foundation model for image compression. *arXiv preprint arXiv:2511.18706*, 2025.

Shiyin Jiang, Wei Long, Minghao Han, Zhenghao Chen, Ce Zhu, and Shuhang Gu. Differentiable vector quantization for rate-distortion optimization of generative image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14440–14450, 2026.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Goluck Konuko, Giuseppe Valenzise, and Stéphane Lathuilière. Ultra-low bitrate video conferencing using deep image animation. In *ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4210–4214, 2021.

Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuqi Li, Haotian Zhang, Li Li, and Dong Liu. Learned image compression with hierarchical progressive context modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Towards extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):888–899, 2025b.

Qi Mao, Chongyu Wang, Meng Wang, Shiqi Wang, Ruijie Chen, Libiao Jin, and Siwei Ma. Scalable face image coding via StyleGAN prior: Toward compression for human-machine collaborative vision. *IEEE Transactions on Image Processing*, 33:408–422, 2024. doi: 10.1109/TIP.2023.3343912.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Matthew J. Muckley, Alaaeldin El-Nouby, Karen Ullrich, Herve Jegou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, 2023.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. Dit-ic: Aligned diffusion transformer for efficient image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. OSDFace: One-step diffusion model for face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021a.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Yongbo Wang, Haonan Wang, Guodong Mu, Ruixin Zhang, Jiaqi Chen, Jingyun Zhang, Jun Wang, Yuan Xie, Zhizhong Zhang, and Shouhong Ding. Switchable token-specific codebook quantization for face image compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp. 1398–1402, 2003.

Jiuhong Xiao, Lavisha Aggarwal, Prithviraj Banerjee, Manoj Aggarwal, and Gerard Medioni. Identity preserving loss for learned image compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wenhan Yang, Haofeng Huang, Jiaying Liu, and Alex C. Kot. Facial image compression via neural image manifold compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3812–3824, 2025. doi: 10.1109/TCSVT.2023.3274739.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. *arXiv preprint arXiv:2512.12229*, 2025.

Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yimin Zhou, Yichong Xia, Bin Chen, Mingyao Hong, Jiawei Li, Zhi Wang, and Yaowei Wang. FaSDiff: Balancing perception and semantics in face compression via stable diffusion priors. *arXiv preprint arXiv:2505.05870*, 2025.

A CODEC ARCHITECTURE

For self-containedness we detail our base codec. The encoder and entropy path follow the AEIC (Zhang et al., 2025) framework; the integration of a one-step diffusion face-restoration prior (Wang et al., 2025a) as the codec's decoder – trained end to end with the encoder and entropy model – is part of this work (Sec. 3). On top of this base, the paper's focus is the training-time identity loss $\mathcal{L}_{\mathrm{id}}$, which touches neither the bitstream nor the inference path. Fig. 7 diagrams the full pipeline.

Encoder. A lightweight analysis transform g_a maps the 512^2 input face $\boldsymbol{x}$ to a compact latent $\boldsymbol{y}$, downsampling by $\times 32$ to a 320-channel feature map. There is no diffusion or heavy computation on the encoder side; the asymmetry (cheap encoder, strong decoder) is what makes the codec practical at extreme-low rate.

Entropy model (AEIC-ME (Zhang et al., 2025)). The latent is quantized (rounding, with a straight-through estimator (Bengio et al., 2013) for gradients) and coded under a learned prior: a hyper-prior ($h_a \to z\,(160\text{ ch}) \to h_s$) predicts per-element mean and scale, and a four-part checkerboard context model (He et al., 2021) adds spatial autoregression, so each symbol is coded conditioned on already-decoded neighbors. An rANS (Duda, 2013) range coder with 16-bit CDFs turns the symbols into the bitstream whose length is the rate R.

Decoder (one-step diffusion). The bitstream is entropy-decoded back to $\hat{\boldsymbol{y}}$, and a synthesis transform g_s produces the decoder latent $\boldsymbol{l}_T$ together with a residual. The latent is refined by a Stable-Diffusion-2.1 UNet (Rombach et al., 2022) whose time-conditioning and cross-attention are removed, so a *single* forward pass – not an iterative sampling loop – replaces the usual diffusion process; it uses the OSDFace (Wang et al., 2025a) one-step face-restoration reference weights. The refined latent plus residual is mapped to the reconstruction $\hat{\boldsymbol{x}}$ by a pruned Stable-Diffusion VAE decoder ("halfDecoder"). At inference the model is fully deterministic and identity-free: no recognition network is involved.

What is trained. The Stable-Diffusion backbone, the OSDFace (Wang et al., 2025a) reference weights, and the VAE are *frozen*; we train only a low-rank adapter on the UNet together with its input/output convolutions, plus the encoder and entropy model. The training objective is given in Appendix B (Eq. 2).

Identity supervision (training only). The identity term compares the reconstruction and the source in the embedding space of the frozen training network F_{train} (CurricularFace (Huang et al., 2020) IR-101), which is fed a 112×112 aligned crop. To avoid a second recognition pass per step we precompute and cache the ground-truth embedding $F_{\mathrm{train}}(\boldsymbol{x})$ offline, so the only added cost at training time is one embedding pass on $\hat{\boldsymbol{x}}$. This branch exists solely during training and is absent at inference.

B TRAINING DETAILS AND FULL CONFIGURATION GRID

This appendix collects the setup details behind Sec. 4.1.

Data. The FFHQ split is a seeded $95/5$ shuffle of its 70k aligned faces into $66{,}500$ training and $3{,}500$ held-out test images (the same split for pretraining and finetuning). The CelebA-HQ cross-dataset test set is deterministic: all $2{,}000$ official validation images plus the first 824 training images, matching the official test-partition size ($2{,}824$); none of it is seen in training.

Identity-network input. The identity network F_{train} takes RGB in $[-1, 1]$ and outputs a 512-dimensional unit-norm embedding. Since faces are FFHQ-aligned we feed it a fixed center crop (0.75) resized to 112×112 rather than running a detector, and precompute and cache $F_{\mathrm{train}}(\boldsymbol{x})$ for the whole dataset, so each training step costs only one extra embedding pass on $\hat{\boldsymbol{x}}$.

Objective. Finetuning optimizes a standard rate–distortion objective with our identity term as the only addition:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{id}}\,\mathcal{L}_{\mathrm{id}}, \\ \mathcal{L}_{\mathrm{base}} &= \lambda_{\mathrm{rate}}\,R + 2.0\,\ell_2 + 1.0\,\mathrm{DISTS} + 0.1\,\mathrm{CLIP} + 0.1\,\mathrm{GAN}_{\mathrm{DINOv3}}. \end{aligned} \tag{2}$$

$\mathcal{L}_{\mathrm{base}}$ is a weighted sum of the entropy-model rate R, a pixel ℓ_2 term, two perceptual terms (DISTS (Ding et al., 2022) and a CLIP (Radford et al., 2021) feature distance), and a vision-aided

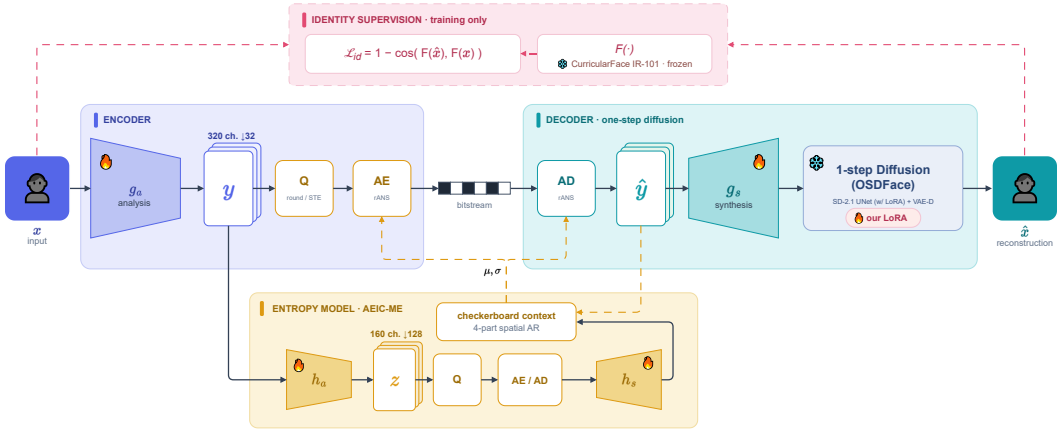


Figure 7: **Base codec architecture.** Lightweight encoder g_a and AEIC-ME entropy model (hyperprior + checkerboard context + rANS) produce the bitstream; a one-step diffusion decoder (SD-2.1 UNet with OSDFace weights, frozen with a trained LoRA; pruned SD-VAE "halfDecoder") reconstructs $\hat{\boldsymbol{x}}$. The snowflake marks the frozen diffusion backbone (only its LoRA is trained). The identity supervision (dashed, red) is training-only: the cached embedding $F_{\text{train}}(\boldsymbol{x})$ and the loss $\mathcal{L}_{\text{id}}$ touch neither the bitstream nor inference.

Table 3: **Training hyperparameters**, shared by all 20 grid models; only the loss weights λ_{rate} and λ_{id} differ between models.

Initialization	base codec pretrained for 300k steps on the same FFHQ training split
Optimizer	AdamW, initial learning rate 5×10^{-5}
LR schedule	staged decay to 2×10^{-5} / 1×10^{-5} / 5×10^{-6} / 1×10^{-6} at 10k / 25k / 28k / 29k steps
Discriminator	vision-aided GAN (DINOv3 backbone), AdamW, learning rate 2×10^{-5}
Steps / batch	30k steps; effective batch 32 (4 per GPU $\times$ 2 GPUs $\times$ 4-step gradient accumulation)
Input	FFHQ-aligned 512^2 crops
Trained parameters	encoder + entropy model (dense); UNet LoRA rank 32 + UNet input/output convolutions
Gradient clipping	max-norm 0.1
Rate weight	λ_{rate} ramped to its target over the first 1,000 steps
Hardware	2$\times$ A100 (80 GB); $\approx$1.4 days per operating point

GAN term with a DINOv3 (Siméoni et al., 2025) discriminator that keeps reconstructions sharp and on the natural-face manifold; λ_{rate} selects the operating point.

Two-stage training. Every model is produced in two stages. *(i) Pretraining* trains the base codec — the encoder, the entropy model, and a UNet LoRA (rank 32) with its input/output convolutions, while the Stable-Diffusion backbone, the OSDFace reference weights, and the VAE stay frozen — for 300k steps on the FFHQ training split under a rate–distortion objective $\lambda_{\text{rate}} R + 2\,\ell_2 + \text{LPIPS} + 0.1\,\text{CLIP}$ (fixed λ_{rate}=0.8, effective batch 8, learning rate 1×10^{-4}). It uses *no* adversarial or identity term, so it is entirely identity-agnostic and is the single checkpoint shared by all operating points. *(ii) Finetuning* then continues this one checkpoint for 30k steps at each of the 20 grid cells under the deployment objective (Eq. 2): it swaps LPIPS for DISTS, adds the vision-aided (DINOv3) adversarial term, ramps λ_{rate} to its target ($\{2, 4, 8, 16, 32\}$), and — for identity models — adds $\lambda_{\text{id}}\mathcal{L}_{\text{id}}$, at effective batch 32 with the staged learning-rate decay of Table 3. The identity loss therefore enters *only* in finetuning, and λ_{id}=0 reproduces the base codec and is the anchor.

Hyperparameters. Table 3 lists the training hyperparameters shared by every model in the grid (and by the identity-specificity controls of Test 3, Sec. 5); only λ_{rate} and λ_{id} vary between models.

Table 4: **Full configuration grid: per-model metrics on the FFHQ test split** ($3{,}500$ **images).** Rows are grouped by identity weight λ_{id} (top block = the anchor, λ_{id}=0); within each block, rate weight λ_{rate} runs $32 \to 2$ (low to high bitrate). ArcFace and MagFace are the held-out identity networks; CurricularFace (F_{train}) is the training network, listed for reference only. pFID and cFID denote patch-FID and clean-FID, respectively. ↑/↓ mark whether higher or lower is better.

		Identity (cosine sim.)			Distortion		Perceptual		Distribution	
λ_{rate}	bpp	ArcFace↑	MagFace↑	Curr.↑	PSNR↑	MS-SSIM↑	LPIPS↓	DISTS↓	pFID↓	cFID↓
$\lambda_{\mathrm{id}} = 0$ *(the anchor, no identity loss)*										
32	0.0041	0.435	0.454	0.400	22.03	0.750	0.228	0.129	2.89	9.07
16	0.0074	0.589	0.614	0.568	23.50	0.802	0.173	0.109	2.16	6.99
8	0.0132	0.736	0.764	0.729	25.00	0.849	0.132	0.093	1.75	5.62
4	0.0225	0.843	0.870	0.848	26.30	0.887	0.101	0.079	1.48	4.45
2	0.0340	0.898	0.922	0.908	27.19	0.910	0.082	0.070	1.34	3.75
$\lambda_{\mathrm{id}} = 0.05$ *(default)*										
32	0.0045	0.586	0.623	0.578	22.21	0.762	0.221	0.126	2.68	8.78
16	0.0082	0.725	0.763	0.730	23.69	0.812	0.169	0.108	2.09	6.88
8	0.0144	0.821	0.857	0.833	25.17	0.856	0.129	0.092	1.71	5.51
4	0.0237	0.882	0.909	0.894	26.42	0.891	0.099	0.079	1.49	4.43
2	0.0351	0.914	0.938	0.927	27.26	0.912	0.081	0.069	1.33	3.71
$\lambda_{\mathrm{id}} = 0.1$										
32	0.0050	0.668	0.711	0.674	22.32	0.769	0.216	0.125	2.59	8.79
16	0.0088	0.777	0.818	0.790	23.79	0.816	0.167	0.107	2.05	6.82
8	0.0151	0.850	0.885	0.866	25.27	0.860	0.127	0.091	1.70	5.47
4	0.0245	0.895	0.923	0.911	26.48	0.893	0.098	0.078	1.46	4.39
2	0.0359	0.921	0.945	0.936	27.31	0.914	0.081	0.069	1.33	3.70
$\lambda_{\mathrm{id}} = 0.2$										
32	0.0056	0.741	0.788	0.760	22.41	0.775	0.214	0.124	2.61	8.67
16	0.0097	0.823	0.865	0.844	23.89	0.822	0.164	0.106	2.04	6.85
8	0.0160	0.876	0.909	0.895	25.33	0.863	0.126	0.091	1.70	5.40
4	0.0256	0.909	0.937	0.928	26.55	0.896	0.097	0.078	1.48	4.33
2	0.0371	0.928	0.952	0.945	27.36	0.915	0.080	0.069	1.33	3.64

Table 4 lists every model behind the body results: the full $\lambda_{\mathrm{rate}} \times \lambda_{\mathrm{id}}$ grid of Sec. 4.1, with the operating point (bpp) and all nine quality metrics per model. Reading down each rate block, raising λ_{id} lifts all three identity scores and slightly improves the distortion, perceptual, and distributional metrics, at a bpp premium that grows with the weight – the per-cell view of the near-free lunch summarized by Table 1 and the identity sweep of Table 14.

C MECHANISM DETAILS

This appendix reports the full numbers behind the three mechanism tests of Sec. 5: the complete per-part spatial breakdown, the band-by-band frequency numbers, a robustness check across the three recognition networks, and the identity–distribution decoupling behind the corollary of the Synthesis (Sec. 5). All measurements use the same λ_{rate}=32, λ_{id}=0.1 cell against its iso-λ_{rate} anchor counterpart, on the $3{,}500$-image FFHQ test split, with the change reported as id − anchor.

Full per-part spatial breakdown. The four-super-region summary is in the body (Table 2); Fig. 8 resolves it into all 16 semantic parts individually. Within the inner-identity group the nose gains $+2.17$ dB, the upper lip $+1.93$ dB, and the brows $\sim +1.5$ dB, while the periphery (ears/hair/hat) and background barely move.

Frequency analysis: spectra and full numbers. Fig. 9 plots the radially-averaged signal and error power spectra behind Test 2 (frequency analysis), and Table 5 gives the band-by-band figures. The signal columns show that the identity loss *does* synthesize slightly more high-frequency signal energy ($71.9\% \to 74.3\%$ of the ground truth's), but this added high frequency is not source-aligned: it leaves the high-frequency error essentially unchanged ($+0.4\%$; only 7.7% of the total error any-

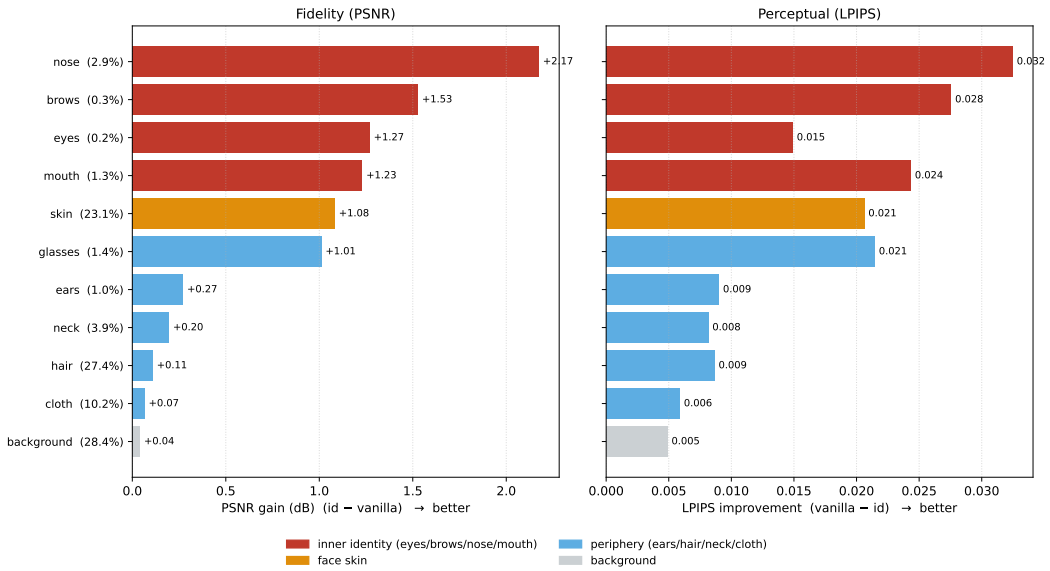


Figure 8: **Per-region attribution** (λ_{rate}=32, λ_{id}=0.1; FFHQ-3500, iso-λ_{rate}; detail for Test 1, spatial attribution). Per-region PSNR gain (left) and LPIPS improvement (right), ordered by PSNR gain.

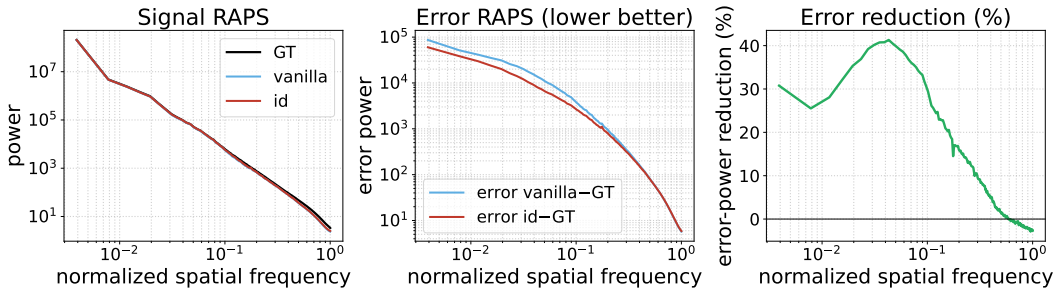


Figure 9: **Frequency analysis** (λ_{rate}=32, λ_{id}=0.1; FFHQ-3500, 512^2, iso-λ_{rate}). Signal and error radially-averaged power spectra (RAPS), as labeled in the panels.

way). High-frequency synthesis is thus present but not the driver of the gain — consistent with the low/mid-frequency structural correction reported in the main text.

Robustness across recognition networks. The reported identity number is the mean over the two held-out networks (ArcFace (Deng et al., 2019), MagFace (Meng et al., 2021)), so no single net drives it. Fig. 10 plots the per-image similarity distributions for all three nets. The mean gain is consistent across them (ArcFace $0.435 \rightarrow 0.668$, MagFace $0.454 \rightarrow 0.711$, CurricularFace (Huang et al., 2020) $0.400 \rightarrow 0.674$). ArcFace carries a heavier low tail – 10.5% of the anchor's reconstructions fall below 0.3 versus 8.7% for MagFace, a residue of its stricter detector alignment on cropped faces – yet the identity loss shrinks that tail sharply (to 2.4%), and the two held-out nets agree on the direction and magnitude of the improvement throughout. Averaging the two therefore reflects a genuine, net-independent recovery rather than an artifact of one embedding's outlier behavior. As a further check with a different margin family, a held-out recognizer trained with the additive *cosine* margin, CosFace (Wang et al., 2018) (the official InsightFace `glint360k_cosface_r100`), shows the same rate-dependent trend (Table 6): the gain is largest at the lowest rate ($+0.165$) and tapers to $+0.018$ at the highest (mean $+0.096$), tracking ArcFace and MagFace rate by rate; we report it here as corroboration and do *not* fold it into the reported identity mean. One caveat remains: all of these held-out networks are, like F_{train}, margin-trained CNN recognizers on similar face corpora, so they

Table 5: **Frequency-band analysis** (λ_{rate}=32, λ_{id}=0.1; FFHQ-3500, iso-λ_{rate}). Left: recovered signal power as a percentage of the ground-truth band power. Right: the id-vs-anchor change in reconstruction-error band power (negative = error reduced).

	Signal (% of GT band)		Error change
Band (norm. freq.)	λ_{id}=0	Id	(id vs. anchor)
Low (0–0.1)	98.7%	99.0%	−36.2%
Mid (0.1–0.5)	85.6%	89.5%	−15.7%
High (0.5–1.0)	71.9%	74.3%	+0.4%
Total	–	–	−23.6%

Table 6: **A further held-out recognizer, CosFace, shows the same near-free trend** (FFHQ, official `glint360k_cosface_r100`; the five operating points of Table 4, λ_{rate}=32 → 2, ≈ 0.004–0.035 bpp). Δ is id-on minus id-off. CosFace tracks the two reported held-out networks rate by rate; it is *not* folded into the reported identity mean.

λ_{rate}	id-off	id-on	Δ
32	0.412	0.577	+0.165
16	0.571	0.725	+0.154
8	0.729	0.829	+0.101
4	0.847	0.891	+0.044
2	0.908	0.926	+0.018
mean Δ			+0.096

are held out from *training* but not architecturally unrelated; confirmation on a structurally different judge (e.g. a human verification study) is a natural complement we leave to future work.

Identity–distribution decoupling (detail for the Synthesis corollary, Sec. 5). Consistent with a structure-driven account, the distributional response is roughly an *order of magnitude* smaller than the identity gain (50.9% identity BD-rate vs. within-5% distributional BD-rate) and, in the decomposition (Fig. 11), is entirely covariance and not concentrated where identity improved most – the signature of an emergent by-product, not a targeted anchoring. A drifted anchor output is, after all, already a "confident wrong person," so most of the identity fix *repositions* it *within* an already-realistic manifold: large in identity, faint in set-level distribution. The same disproportion is why identity needs a dedicated objective: distributional metrics greatly under-register it.

Reproduction at the λ_{id}=0.05 default, and a rate-matched control. The body isolates the mechanism at λ_{id}=0.1 for visibility; here we confirm that none of the three tests depends on that choice and that the gain is not an artifact of the extra rate the loss spends. *(i) The λ_{id}=0.05 default reproduces every test.* Test 1: the per-super-region improvement keeps the same monotone gradient ordered by identity content — inner identity +1.01 dB / −0.019 LPIPS (Zhang et al., 2018), face skin +0.66 / −0.013, periphery +0.07 / −0.0045, background +0.05 / −0.0035 — smaller in magnitude than at λ_{id}=0.1 (the lighter weight), but with the identical ordering. Test 2: the error reduction is again low/mid-frequency (−24.5% low, −9.5% mid) versus a slight high-frequency increase (+0.7%). Test 3: retraining the two controls at λ_{id}=0.05 reproduces the collapse — against the anchor's held-out identity (0.435/0.454/0.400 for ArcFace (Deng et al., 2019)/MagFace (Meng et al., 2021)/CurricularFace (Huang et al., 2020)), the correct-target loss lifts identity to 0.586/0.623/0.578 while *shuffle* (Δ−0.014 over ArcFace/MagFace) and *random* (Δ+0.001) both fall back to the anchor.

(ii) The gain survives a rate-matched comparison. At a fixed λ_{rate} the identity loss lands at a slightly higher bitrate than its λ_{id}=0 anchor (0.0041 → 0.0045 bpp at λ_{id}=0.05; → 0.0050 at λ_{id}=0.1). To remove that confound we interpolate the anchor's own rate–identity curve (its λ_{rate}=32 and 16 points) to the identity-loss model's exact bitrate and compare there (Table 7). Because the anchor curve is nearly flat at this extreme-low-rate end, the +10–22% extra rate buys it only a small identity rise, so a large rate-matched gain remains at both weights (+0.13 at λ_{id}=0.05, +0.19 at λ_{id}=0.1 on

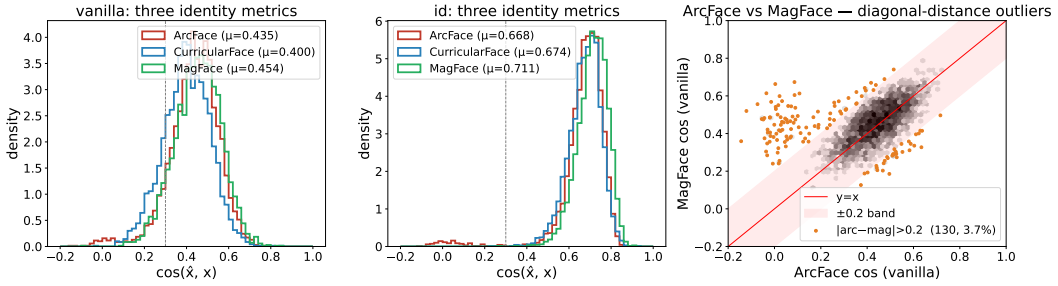


Figure 10: **Per-image identity-metric distributions across the three recognition networks** (λ_{rate}=32 cell, FFHQ test): the training-side CurricularFace and the two held-out nets, ArcFace/MagFace. Robustness check for the distribution analysis (ArcFace detector-alignment outliers).

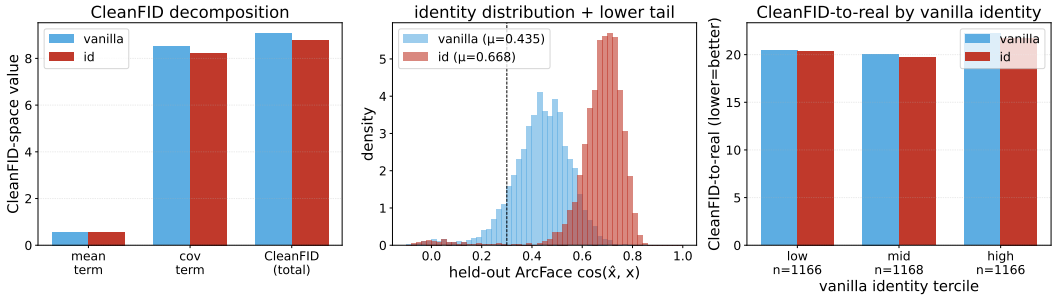


Figure 11: **Distribution decomposition** (λ_{rate}=32, λ_{id}=0.1; FFHQ-3500, iso-λ_{rate}; detail for the Synthesis corollary, Sec. 5). CleanFID mean/covariance split, the identity-similarity distribution, and CleanFID-to-real by identity tercile. Identity improves dramatically yet CleanFID moves far less, and its small change is entirely covariance and not concentrated in the low-identity tercile: CleanFID greatly under-registers identity, so distribution anchoring is not the mechanism.

ArcFace); the extra rate accounts for just 13–18% of the raw improvement. The near-free identity gain is therefore a structural-alignment effect, not a bitrate effect.

D TRANSMITTING IDENTITY: EVEN A STRONG SIDE CHANNEL IS REDUNDANT

Sec. 3 argues that transmitting the identity as side information is redundant: it is the training *objective*, not extra transmitted bits, that preserves identity. Our first attempt used a *weak* injector – a global channel-wise FiLM (Perez et al., 2018) – and there the learned identity rate R_{id} collapses *perfectly to* 0.0: under $\mathcal{L}_{\text{id}}$-only training the transmitted code degenerates to an input-independent constant ($\cos(\hat{\mathbf{i}}_i, \hat{\mathbf{i}}_j)$=1.0 across 256 distinct inputs) and the rate–distortion objective spends no bits on the channel at all. One might dismiss this as the injector being too weak to exploit a code that a stronger, spatially resolved conditioning path would use. This appendix answers that objection directly: we replace the injector with the strongest one we would reasonably deploy, show the conclusion is unchanged, and – crucially – quantify the tiny BD-rate movement against the compute the channel actually costs.

A strong identity adapter. We transmit the identity through an entropy-coded transform coder – an MLP analysis transform maps the embedding $F_{\text{train}}(\boldsymbol{x})$ to a 128-dimensional latent, which is scalar-quantized and entropy-coded under a factorized-Gaussian prior at rate R_{id}, and an MLP synthesis transform reconstructs the code $\hat{\mathbf{i}}$ (renormalized to the unit sphere). We then inject it with an image-prompt-adapter-style *decoupled cross-attention* module: a small tokenizer maps $\hat{\mathbf{i}}$ to K=4 conditioning tokens, and a *zero-initialized* cross-attention block (query = U-Net features, key/value = id tokens) is attached to *every one of the* 16 U-Net transformer blocks. Because the SD back-

Table 7: **Rate-matched (iso-bpp) identity control** (λ_{rate}=32; FFHQ-3500). The λ_{id}=0 anchor is interpolated along its own rate–identity curve to the identity-loss model's bitrate ("Anchor @ bpp"), isolating the gain attributable to the identity loss rather than to the extra rate. "Raw gain" is against the iso-λ_{rate} anchor (the λ_{rate}=32 λ_{id}=0 point, 0.435/0.454/0.400, as used in the body); "Iso-bpp gain" is against the rate-matched "Anchor @ bpp". Held-out ArcFace/MagFace and training-side CurricularFace shown.

Weight (bpp)	Net	Anchor @ bpp	Id-on	Raw gain	Iso-bpp gain
λ_{id}=0.05	ArcFace	0.456	0.586	+0.151	**+0.129**
(0.0045)	MagFace	0.476	0.623	+0.169	+0.146
	CurricularFace	0.423	0.578	+0.178	+0.155
λ_{id}=0.1	ArcFace	0.478	0.668	+0.233	**+0.190**
(0.0050)	MagFace	0.498	0.711	+0.257	+0.213
	CurricularFace	0.447	0.674	+0.274	+0.228

Table 8: **BD-rate of the ID-adapter (inject ON) over the loss-only model** (PCHIP; five operating points; full 3,500-image held-out FFHQ). Negative = fewer bits at matched quality. All three metrics change by well under 1%.

	ArcFace	PSNR	LPIPS
BD-rate vs. loss-only	−0.57%	−0.91%	−0.61%

bone's text cross-attention is removed when it is repurposed as a one-step restoration decoder, this is a freshly constructed, dedicated identity path – not a revived text branch. Zero-initializing each adapter's output projection makes the block start as the identity map, so it can only *add* conditioning on top of an already-working decoder and never degrades the base path: the strongest, most favorable version of the side channel.

Two-phase protocol. So that the redundancy test runs against a channel that genuinely carries identity rather than a strawman, we train in two phases. Phase 1 (15k steps) trains the full channel with two auxiliary losses that make the transmitted code faithful and used, $\mathcal{L}_{\text{rec}} = 1 - \cos(\hat{\mathbf{i}}, \mathbf{i})$ and $\mathcal{L}_{\text{con}} = 1 - \cos(F_{\text{train}}(\hat{\boldsymbol{x}}), \hat{\mathbf{i}})$ (with $\mathbf{i} = F_{\text{train}}(\boldsymbol{x})$ the cached true embedding), building a channel with $\cos(\hat{\mathbf{i}}, \mathbf{i}) \approx 0.44$. Phase 2 (20k steps) then continues under $\mathcal{L}_{\text{id}}$ alone – exactly the regime our shipped model uses – letting the rate–distortion objective decide how many bits the channel deserves. We run this at five operating points $\lambda_{\text{rate}} \in \{2, 4, 8, 16, 32\}$ and evaluate inject-ON reconstructions on the full 3,500-image held-out FFHQ set, with practical (header-excluded, rANS) bpp measured identically to the main curves so the adapter points land on the same axes as the loss-only +id-loss frontier.

Result 1: injection yields only a marginal BD-rate gain. Even with this strong adapter the objective still spends almost nothing on the channel: under $\mathcal{L}_{\text{id}}$-only training the learned rate R_{id} is only 2.8–22 bits per 512×512 image, on average ≈ 0.5% of the bitstream (at most 1.8% at the lowest rate). Placed on the same axes as the loss-only +id-loss curve (λ_{id}=0.05; Fig. 12), the inject-ON points sit essentially *on* that frontier. We quantify the difference with BD-rate, the standard rate–distortion summary and a far more reliable measure than any single-point comparison. Over the loss-only model the ID-adapter is consistently but slightly better on all three metrics (Table 8): ArcFace (Deng et al., 2019) −0.57%, PSNR −0.91%, LPIPS −0.61% (negative = fewer bits at matched quality). The gain is real but small – well under 1% – so transmitting the identity through even a strong adapter shifts the frontier only marginally.

Result 2: the channel is not free. That marginal change carries a real cost (Table 9; measured on an RTX 8000 at 512×512). On the decoder, the adapter adds 48.0M parameters (+5.8% over the 833M U-Net), +3.9% decode FLOPs, and +3.5% decode latency. Worse, *transmitting* the identity forces the *encoder* to run a 65.1M-parameter face-recognition network (24.2 GFLOPs, +9.2 ms/image) to produce $F_{\text{train}}(\boldsymbol{x})$ on every image – a dependency the loss-only model incurs only *once, at training time* and never ships. End to end the side channel adds +7.9% wall-clock and

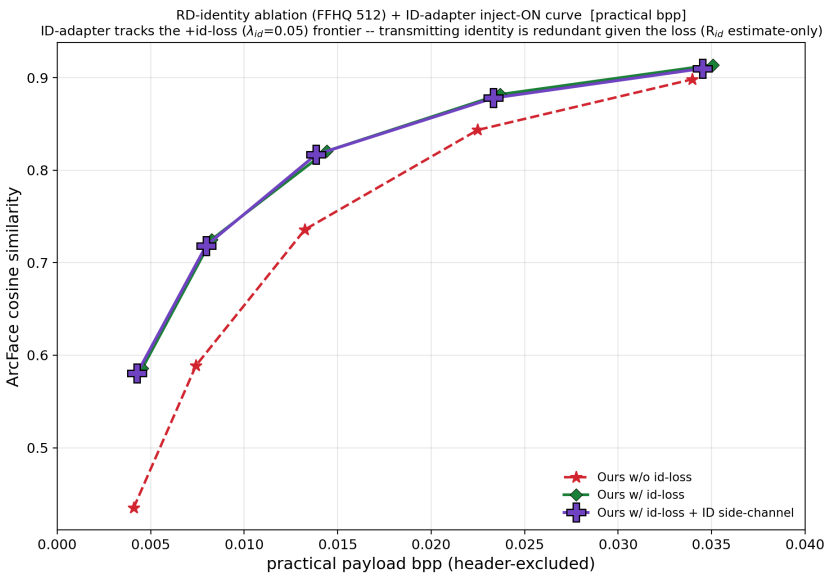


Figure 12: **Transmitting the identity tracks the loss-only frontier.** ArcFace similarity vs. practical bpp on the full 3,500-image held-out FFHQ set. Red (*Ours w/o id-loss*): our anchor, λ_{id}=0. Green (*Ours w/ id-loss*): the loss-only model, λ_{id}=0.05. Purple (*Ours w/ id-loss + ID side-channel*): the strong ID-adapter with the identity additionally transmitted and injected. The id-loss lifts the anchor to the green frontier (large, red→green); transmitting the identity on top (green→purple) shifts it only marginally (sub-1% BD-rate).

Table 9: **Complexity overhead of transmitting identity** (ID-adapter; RTX 8000, 512×512). Decoder-side cost is the cross-attention adapter; encoder-side cost is the face-recognition network the transmitter must run to produce the embedding. Percentages are relative to the loss-only base codec.

	Added by the side channel	Overhead vs. base
Params	decoder: id-codec + tokenizer + 16× cross-attn	+48.0M (+5.8% of U-Net)
	encoder: face-recognition net (F_{train})	+65.1M
	total	+113.1M
Latency	decoder: cross-attn injection	+6.1 ms (+3.5%)
	encoder: identity extraction $F_{\text{train}}(\boldsymbol{x})$	+9.2 ms
	full pipeline (base 193.6 ms)	+15.3 ms (+7.9%)
FLOPs	decoder: cross-attn injection	+25.9 GFLOPs (+3.9%)
	encoder: identity extraction (F_{train}, IR-101)	+24.2 GFLOPs
	total (base U-Net 662.7 GFLOPs)	+50.1 GFLOPs (+7.6%)

+7.6% FLOPs to the pipeline, plus an entropy-coded identity codec and a per-image transmitted code, all for a sub-1% BD-rate gain.

Why, and the weak-injector limit. The reason is the loss itself: once $\mathcal{L}_{\text{id}}$ shapes the base path to preserve identity, the reconstruction already carries it through the base bitstream, so the rate–distortion-optimal side-info rate is near zero and even a strong injector has nothing extra to place. The weak global-FiLM injector we first tried makes this starker still: beyond collapsing its rate to zero (above), ablating its injection at fixed rate moves every metric by $|\Delta| \leq 0.001$ even on the checkpoint whose code is most faithful ($\cos \approx 0.46$). The decisive evidence is our shipped IdentiCodec, which carries *no* side channel at all yet reaches held-out ArcFace (Deng et al., 2019) up to 0.85 at λ_{rate}=8 and 0.91 at λ_{rate}=2 purely from the loss (Table 4). Weighed against the sub-1% BD-rate it gains, the transmitted channel costs +7.9% latency, +7.6% FLOPs, a 65M-parameter encoder-side recognition network, and an entropy-coded identity codec – a poor trade at any rate we

Table 10: **Competitor evaluation recipes.** "FFHQ-ft" marks the learned competitors we could finetune on our FFHQ data; the rest are evaluated as released (Sec. 4.1 explains why GLC, OneDC, and VTM cannot be finetuned). "Rate control" is the knob swept to trace each curve; "Points" is the number of operating points.

Codec	Model source	FFHQ-ft	Points	Rate control
AEIC-ME	released model	yes	5	rate weight
OneDC	released model	no	6	Lagrangian λ
MS-ILLM	official zoo (quality/vlo)	yes	4	quality index
GLC	released (natural-image)	no	4	quality index
VTM	VTM-24.0, all-intra	n/a	7	QP 44–63
RDVQ	released model	yes	5	rate weight
DiffC	released (SD-1.5)	n/a	6	recon. timestep
DiT-IC	released (SANA-600M)	n/a	3	quality level

test. The gain is real but too small to justify the added complexity: identity is better obtained from the objective than from bits.

E COMPETITOR RECIPES AND REPRODUCTION

This appendix documents, for every competitor in Table 1, the exact model we evaluated, how its bitrate was obtained, whether we could finetune it on faces, and the command that reproduces it. Table 10 summarizes; per-codec notes follow. All reconstructions – ours and every competitor – are scored by one shared pipeline on the same $3{,}500$-image FFHQ (and $2{,}824$-image CelebA-HQ) test sets, over the identical nine metrics of Sec. 4.1, and BD-rate is integrated on the $\log$-bitrate axis with our model as the anchor. Bitrate accounting: for our model and for MS-ILLM (Muckley et al., 2023), OneDC (Xue et al., 2025), RDVQ (Jiang et al., 2026), and AEIC-ME (Zhang et al., 2025), bpp is the size of the *actual* entropy-coded bitstream, counted as payload stream bytes (the standard LIC accounting; fixed container framing excluded) – our latents are coded with the rANS (Duda, 2013) coder of Sec. 3, and RDVQ with its released real coder (causal top-k tensor-rANS). VTM (Bross et al., 2021) bpp is its full standard-format bitstream. The public GLC (Jia et al., 2024) release exposes no real entropy coder, so GLC alone uses its entropy-model rate estimate; this omits real-coding overhead and therefore, if anything, favors GLC. Concretely, inflating GLC's estimated rate by 20% – well beyond the 1–3$\%$ estimate-to-real gap typical of hyperprior entropy coders – shifts its operating points rightward on the log-rate axis and can only *increase* its reported $+137.7\%$ identity BD-rate (Table 1); the estimate can thus at most understate GLC's deficit, and no plausible correction overturns its $>100\%$ identity gap.

BD-rate protocol. Each entry follows the standard Bjøntegaard construction: a method's operating points are mapped to (quality, $\log_{10}$ bitrate), interpolated with a shape-preserving monotone piecewise-cubic Hermite (PCHIP) spline — not a single global cubic, which overshoots (Runge) on the competitors' sparse rate ladders (see Appendix O) — and the log-rate gap to the anchor's curve is integrated over the quality range where the two *overlap*; the result is the average bitrate change at matched quality. When a method never reaches the anchor's quality range on a metric there is no overlap and the entry is "–". The identity column is a single BD-rate computed on the per-rate *mean* of the two held-out similarities (ArcFace (Deng et al., 2019), MagFace (Meng et al., 2021)).

AEIC-ME (Zhang et al., 2025). The released AEIC-ME model, evaluated at five rate points; we additionally finetune it on our FFHQ training split (the "+FFHQ-ft" row) with AEIC-ME's own rate–distortion–perception objective (*no* identity loss, same loss weights and 30k-step schedule), each rate initialized from the corresponding released checkpoint.

OneDC (Xue et al., 2025). Evaluated as released over six Lagrangian-λ points (plus one explicit low-bpp target). Finetuning is infeasible: its training pipeline depends on auxiliary networks that are absent from the public release, so face retraining cannot be reproduced. Reconstructions are scored through the shared external-codec path, `scripts/eval_external_codec.py --family OneDC --model_name <pt> --bpp <b>`.

MS-ILLM (Muckley et al., 2023). We use the official model zoo (the standard quality levels plus the two very-low-rate "vlo" variants) via `scripts/run_msillm.py --quality <q>`,

which writes reconstructions and the real per-image bpp. We additionally *finetune* it on FFHQ (the "+FFHQ-ft" row): only the generator is trained, with the encoder and entropy model frozen so the bitstream statistics – and thus the operating points – are unchanged; the finetuned generator is loaded over the matching pretrained architecture by `scripts/run_msillm_finetuned.py --ckpt <last.ckpt> --base_quality <q>`.

GLC (Jia et al., 2024). The public GLC checkpoint is the natural-image model with no face finetune (the release ships inference only, no training code), so it is evaluated as released and labeled accordingly, via `scripts/run_glc.py --model_path <GLC_image.pth.tar>` over four quality indices.

VTM (Bross et al., 2021). The JVET VVC reference software VTM-24.0 in all-intra mode (`encoder_intra_vtm.cfg`, perceptual QP adaptation off), encoding each image as a single YUV420 10-bit intra frame at constant QP over 44–63, via `scripts/run_vvc_codec.py --qp <QP>`; bpp is the resulting bitstream size.

RDVQ (Jiang et al., 2026). Evaluated at five rate-weight points, both as released and FFHQ-finetuned (the "+FFHQ-ft" row). Reproducing the finetune on Turing GPUs required resuming with the discriminator reset and a `bf16`→`fp32` cast in the autoregressive predictor; see the RDVQ finetune script in our release.

DiffC (Vonderfecht & Liu, 2025). A training-free codec that transmits a pretrained diffusion model's reverse trajectory by reverse-channel coding; we run the released implementation on Stable Diffusion 1.5 over six reconstruction-timestep points (`recon_ts` $\in \{591, \ldots, 150\}$), which trade rate against fidelity, via `gen_rd.py --size <res> --recon_ts <t>`. The payload is the reverse-channel bitstream counted in bytes (header-excluded, matching our accounting), scored through the same external-codec path. Being training-free, it admits no face finetune; we evaluate it at 512 (full test set) and, for the multi-resolution study, at 256 and 1024.

DiT-IC (Shi et al., 2026). A recent diffusion-transformer codec that anchors a one-step SANA-600M latent-diffusion prior in an entropy-coded pipeline. We run the released merged checkpoints at the three lowest of its four discrete quality levels (`q1`–`q3`, measured at 0.010, 0.043, 0.080 bpp on faces; `q4` lies far above our regime), via the authors' inference code, and score the reconstructions through the same external-codec path with the real rANS bitstream. It is a released model with only discrete quality levels and no face finetune, and only its lowest point (`q1`, 0.010 bpp) falls inside our extreme-low rate range, so a Bjontegaard BD-rate is not well-defined (it swings with the sparse ladder); we therefore show DiT-IC on the rate–distortion curves (Fig. 3) rather than as a BD-rate table row. Despite excellent per-point realism (e.g. FID 11 at `q1`), its identity at $\approx$0.01 bpp is only 0.41 (vs. 0.78 for ours), and matching our identity costs it $\approx$2–3$\times$ the bitrate — the same realism-without-identity failure at extreme-low rate, on the most recent generative codec.

F FULL FFHQ RATE–DISTORTION

Fig. 13 gives the complete nine-metric rate–distortion comparison on FFHQ behind the identity strip of Fig. 3 and the BD-rate of Table 1.

G IDENTITY VERIFICATION: ROC AND TPR@FPR

Mean cosine similarity summarizes identity with a single number; here we report the full *verification* behavior that a face-recognition system actually operates on. We use the held-out ArcFace (Deng et al., 2019) embedding (not the CurricularFace (Huang et al., 2020) network the loss trains against, so this is not circular). For every image we form a *genuine* score — the cosine between a reconstruction and *its own* source — and *impostor* scores against every other source in the 3,500-image test set (the full $3{,}500 \times 3{,}499$ off-diagonal). Sweeping a decision threshold on this score traces a receiver operating characteristic (ROC) curve (Fig. 14); we report the true-positive rate (TPR; the fraction of genuine pairs accepted) at fixed false-positive rates (FPR; the fraction of impostor pairs wrongly accepted) in Table 11. Each codec is taken at its operating point *nearest* in bitrate to our reference, omitting a codec at a rate point whose bitrate range does not reach it (at the extreme-low rate this excludes VVC, RDVQ, and GLC, whose lowest points exceed 2$\times$ our bitrate); where an included competitor still sits at a *higher* bitrate than ours (e.g. AEIC-ME at 0.0063 vs. our 0.0045 bpp) the comparison, if anything, favors the competitor.

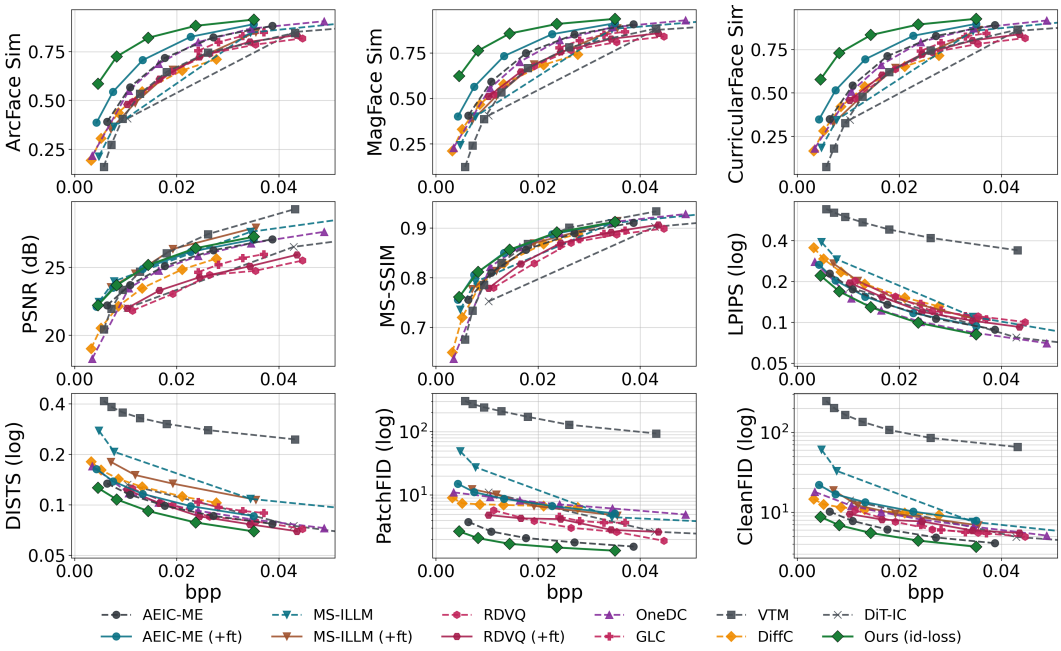


Figure 13: **Rate–distortion vs. competitors on FFHQ (all metrics).** Identity (held-out recognition) and perceptual/distortion metrics against bitrate; our default identity-aware codec (λ_{id}=0.05) vs. MS-ILLM, RDVQ, OneDC, GLC, VVC, and AEIC-ME (finetuned solid, public dashed). GLC's bitrate is an entropy-model estimate (App. E).

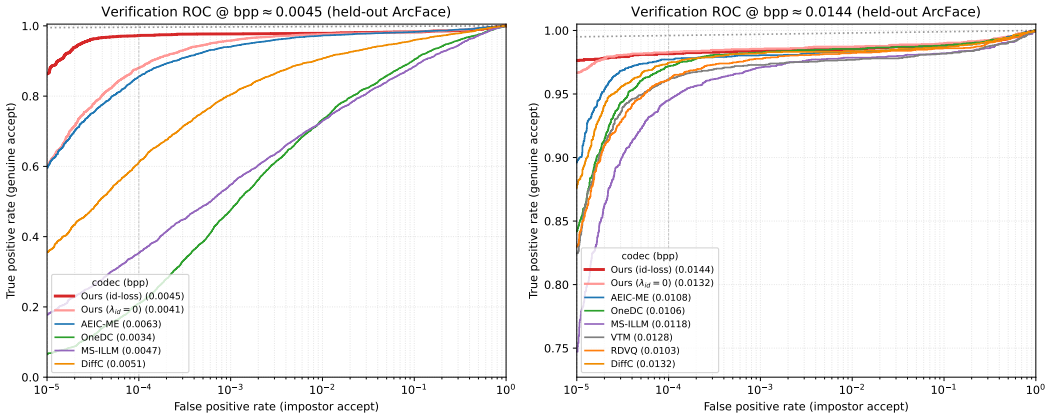


Figure 14: **Identity-verification ROC (held-out ArcFace).** Genuine = reconstruction vs. its own source; impostor = all cross pairs. Left: extreme-low rate (bpp≈0.0045); right: bpp≈0.0144. Legend lists each codec's actual bitrate. Dashed verticals mark FPR=$10^{-4}, 10^{-5}$. At the ultra-low rate our identity-aware codec (red) dominates; the generative competitors drop away at strict FPR.

Two effects stand out. First, at lenient thresholds (FPR $\geq 10^{-3}$) all reasonable codecs verify well — impostors (distinct identities) are trivially separable — so the identity signal is a *margin*, not a separability, effect; it surfaces only at strict operating points. Second, at those strict points and at extreme-low rate the differences are large: the generative competitors OneDC (Xue et al., 2025) and MS-ILLM (Muckley et al., 2023) *collapse* (TPR at FPR=10^{-5} of 0.07 and 0.18), consistent with their rendering a convincing but wrong person, while our codec rises from 0.60 without the loss to 0.86 with it — best by a wide margin. The pattern is the paper's thesis made operational: at the strict operating points a recognizer actually uses, a generative decoder sacrifices identity for realism, and the identity loss is what restores it.

Table 11: **Verification TPR at fixed FPR** (held-out ArcFace, FFHQ 3,500-image test set), at two bitrates. "Genuine" is the mean genuine cosine. Each codec at its nearest operating point; bitrate listed. **Best** per column in bold. At the extreme-low rate, VVC/RDVQ/GLC are omitted (their lowest available bitrate exceeds 2× our reference, so they are not rate-matched there). Competitors: AEIC-ME, OneDC, MS-ILLM, VVC, RDVQ, DiffC.

Codec	bpp	Genuine	TPR@10^{-2}	TPR@10^{-3}	TPR@10^{-4}	TPR@10^{-5}
Extreme-low rate (bpp≈0.0045)						
Ours (λ_{id}=0.05)	0.0045	**0.586**	**0.979**	**0.977**	**0.973**	**0.863**
Ours (λ_{id}=0)	0.0041	0.435	0.976	0.958	0.880	0.597
AEIC-ME	0.0063	0.391	0.972	0.940	0.856	0.596
OneDC	0.0034	0.217	0.733	0.476	0.210	0.065
MS-ILLM	0.0047	0.214	0.729	0.548	0.353	0.177
DiffC	0.0051	0.306	0.907	0.805	0.611	0.355
Low rate (bpp≈0.0144)						
Ours (λ_{id}=0.05)	0.0144	**0.821**	0.985	0.984	0.982	**0.976**
Ours (λ_{id}=0)	0.0132	0.736	**0.987**	**0.985**	**0.983**	0.967
AEIC-ME	0.0108	0.568	0.983	0.981	0.977	0.896
VVC	0.0128	0.534	0.977	0.973	0.962	0.825
OneDC	0.0106	0.550	0.985	0.983	0.972	0.842
MS-ILLM	0.0118	0.487	0.979	0.971	0.945	0.747
RDVQ	0.0103	0.480	0.982	0.978	0.961	0.830
DiffC	0.0132	0.546	0.984	0.982	0.975	0.876

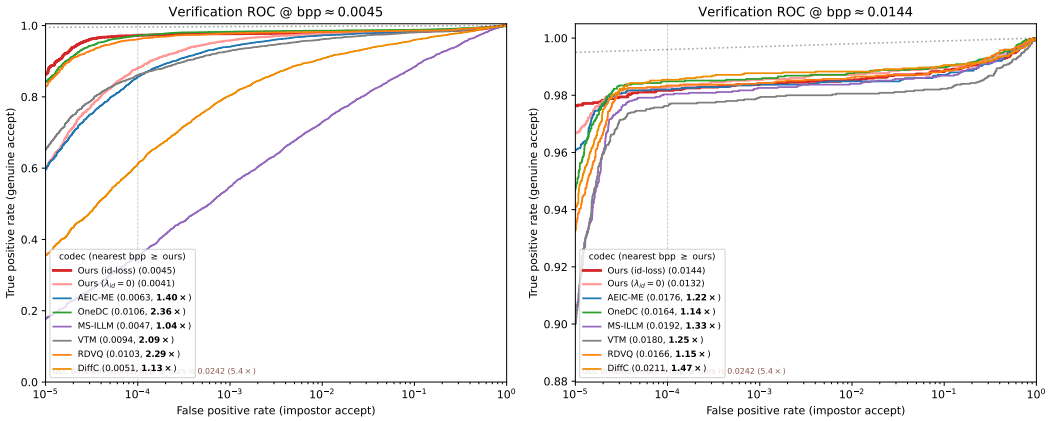


Figure 15: **Verification ROC with each competitor at its nearest bitrate ≥ ours** (supplementary to Fig. 14). Each competitor is read at its nearest operating point with bitrate at least ours; its multiple of our bitrate is bold in the legend. Left: bpp≈0.0045 (full range, so the generative collapse is visible); right: bpp≈0.0144 (y-axis zoomed). Even given up to 2.36× our rate, competitors sit below our identity-aware codec at strict FPR. GLC omitted (nearest point ≥ ours is 5.4× our rate).

Robustness to the bitrate match. The comparison above (Fig. 14, Table 11) is our central analysis; as a supplementary check that its conclusion does not hinge on the exact bitrate match, Fig. 15 instead gives each competitor its nearest operating point at a bitrate ≥ ours — at least as many bits as we spend. Even so handed *more* rate than us — up to 2.36× at the extreme-low rate — our identity-aware codec still leads at the strict operating point: at bpp≈0.0045 its TPR at FPR=10^{-5} (0.863) exceeds OneDC at 2.36× our rate (0.842), RDVQ at 2.29× (0.830), VVC at 2.09× (0.651), and DiffC (Vonderfecht & Liu, 2025) — even at only 1.13× our rate — (0.355). GLC is omitted here: its lowest available point is 0.0242 bpp (5.4× our rate), too far above to be a rate comparison.

H ISOLATING THE BASE CODEC FROM THE IDENTITY LOSS

Our competitor-facing advantage (Table 1) combines two ingredients: the *base codec* — the OSDFace-prior decoder in our compression pipeline, with *no* identity loss (λ_{id}=0) — and the iden-

Table 12: **Identity BD-rate decomposition** (FFHQ; mean of ArcFace and MagFace; + means the competitor needs more bitrate to match). Left: vs. the base codec (λ_{id}=0). Middle: vs. the full model (λ_{id}=0.05, as in Table 1). Right: the increment the identity loss adds. Indented +FFHQ-ft rows are the FFHQ-finetuned variants (only AEIC-ME, which shares our backbone, narrows the vs.-base gap). Competitors: AEIC-ME, OneDC, MS-ILLM, GLC, VTM, RDVQ, DiffC.

Competitor	vs. base (λ_{id}=0)	vs. full (λ_{id}=0.05)	id-loss increment
AEIC-ME	+49.8%	+116.3%	+66.5
+FFHQ-ft	+16.6%	+72.6%	+56.1
OneDC	+55.3%	+119.1%	+63.8
MS-ILLM	+73.4%	+136.4%	+63.0
+FFHQ-ft	+99.2%	+184.0%	+84.8
GLC	+63.4%	+138.1%	+74.7
VTM	+103.1%	+194.8%	+91.7
RDVQ	+107.9%	+225.1%	+117.2
+FFHQ-ft	+99.2%	+207.5%	+108.3
DiffC	+111.6%	+259.6%	+148.0

tity loss on top of it. To separate them, we recompute the identity BD-rate using the base codec as the anchor (Fig. 16, Table 12). Even with no identity loss, the base codec alone already forces every competitor to spend +49.8% (AEIC-ME) to +107.9% (RDVQ) more bitrate to match its identity, and it leads on the perceptual and distributional axes as well (Fig. 16) — so a large part of the win is the base architecture, matching the λ_{id}=0 anchor of Sec. 4.2. The identity loss then *adds* a comparable second increment (+63 to +117 percentage points), lifting the gap to the +116%–+225% reported against the full model in Table 1. Both contributions are substantial, but the vs.-base column is inflated by the competitors not being face-specialized: when the strongest of them, AEIC-ME, is itself finetuned on our FFHQ split, its identity gap to our base codec narrows from +49.8% to +16.6% (Table 12, +FFHQ-ft rows). The same finetuning does *not* shrink the gap for the architecturally different MS-ILLM (+73.4% → +99.2%) or RDVQ (+107.9% → +99.2%), confirming this residual is a property of our base *architecture*, not merely of face-domain data. The dataset-independent split is therefore uneven — the identity loss is the larger contributor, the base prior the smaller — though both are non-trivial and free at inference.

I Cross-dataset rate–distortion (CelebA-HQ)

Fig. 17 reports the full rate–distortion comparison on the 2,824-image CelebA-HQ test set, which our model never sees during training. The out-of-distribution behavior mirrors FFHQ: our codec dominates all competitors on the identity axes by a wide margin at extreme-low rate and remains competitive-to-best on the distortion, perceptual, and distributional metrics; only MS-ILLM (Muckley et al., 2023)'s PSNR exceeds ours – its over-smoothing signature (Sec. 4.2).

J Generality: per-metric RD curves

Fig. 18 shows the per-metric rate–distortion curves behind the cross-codec BD-rate of Table 13: for MS-ILLM (Muckley et al., 2023) (GAN codec), RDVQ (Jiang et al., 2026) (VQ codec), and HPCM (Li et al., 2025a) (a purely non-generative MSE + rate-optimized LIC), the id-*off* finetuned arm (gray) versus the same codec finetuned with our identity loss (red), at matched bitrate. We use λ_{id}=0.1 for the two generative codecs and λ_{id}=0.003 for HPCM, where the smaller weight matches the identity gain to ours rather than over-pushing it (a large weight on a distortion-optimized codec buys a big identity jump but leaves the frontier). The red gap annotations give the per-point change; identity (top-left of each block) improves throughout, while the perceptual and distortion panels stay close to the anchor – and for HPCM the distributional panels (DISTS, PatchFID, CleanFID) move *below* the anchor, largest at the lowest rates, so on the non-generative codec the identity loss is not merely near-free but a mild realism regularizer. Fig. 19 shows qualitative HPCM reconstructions at the lowest rate, where id-off drifts to a different person and the identity loss restores the correct face at essentially the same bitrate.

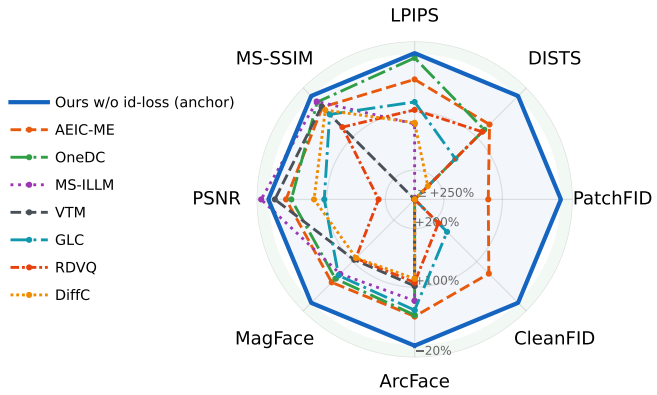


Figure 16: **Base codec** (λ_{id}=0) **vs. competitors** (FFHQ, full test set, all metrics; anchor ring = base codec). Each competitor's BD-rate to match the base codec; inward means it needs more bitrate. Even *without* the identity loss, every competitor sits inside the ring on the held-out identity axes (ArcFace, MagFace).

Table 13: **Cross-codec BD-rate of the identity loss (FFHQ).** All-metric BD-rate (%) of each codec trained *with* our identity loss, relative to the same codec's id-*off* finetuned anchor. Negative = the id-loss model needs that many *fewer* bits to match the anchor at equal quality (better). We use λ_{id}=0.1 for the two generative codecs (MS-ILLM, RDVQ) and λ_{id}=0.003 for the non-generative HPCM, matching its identity gain to ours. Identity is the held-out mean (ArcFace, MagFace).

Codec (+id-loss vs. id-off anchor)	Identity	PSNR	MS-SSIM	LPIPS	DISTS	PatchFID	CleanFID
MS-ILLM (GAN codec)	−17.4	2.4	−2.9	0.3	1.3	−6.9	5.5
RDVQ (VQ codec)	−39.8	−2.2	−5.9	2.1	3.6	−10.5	−5.6
HPCM (non-gen. LIC codec)	−29.4	1.5	0.4	1.3	−7.6	−24.2	−8.4

Lower is better

K Identity-weight sweep: per-metric RD curves

Fig. 20 shows the per-metric rate–distortion curves behind the identity-weight ablation of Table 14: our codec at $\lambda_{\text{id}} \in \{0.05, 0.1, 0.2\}$ against Ours (λ_{id}=0). Identity (bottom row) rises monotonically with λ_{id} and every setting is far above the anchor, while distortion and perceptual quality (top two rows) stay close across settings and only degrade mildly at λ_{id}=0.2.

L Additional qualitative results

Figs. 21 and 22 show extreme-low-bitrate reconstructions on eight test images of each dataset (columns: Original / AEIC-ME (Zhang et al., 2025) / OneDC (Xue et al., 2025) / MS-ILLM (Muckley et al., 2023) / VTM (Bross et al., 2021) / DiffC (Vonderfecht & Liu, 2025) / Ours (λ_{id}=0) /

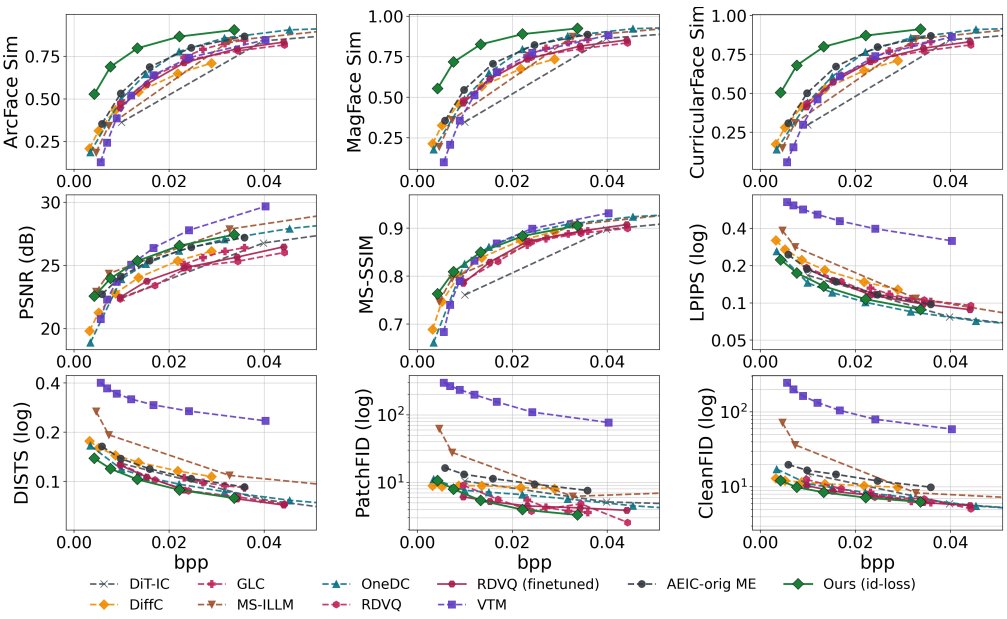


Figure 17: **Rate–distortion vs. competitors on CelebA-HQ (cross-dataset).** Same metrics and protocol as Fig. 13, on the held-out CelebA-HQ test set. Our default identity-aware codec (λ_{id}=0.05) vs. MS-ILLM, RDVQ, OneDC, GLC, VVC, and AEIC-ME.

Table 14: **Identity-weight (λ_{id}) ablation on FFHQ.** All-metric BD-rate (%) of our model at each λ_{id} against Ours (λ_{id}=0) as the anchor. Negative = our model needs that many fewer bits to match the anchor at equal quality (better). Identity is the held-out mean (ArcFace, MagFace).

Setting	Identity	PSNR	MS-SSIM	LPIPS	DISTS	PatchFID	CleanFID
Ours (λ_{id}=0, anchor)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ours (λ_{id}=0.05)	−33.7	1.8	−1.3	4.1	5.0	2.0	4.8
Ours (λ_{id}=0.1)	−44.7	3.8	−0.8	7.5	8.8	3.3	9.3
Ours (λ_{id}=0.2)	−53.8	8.0	1.3	13.2	14.9	12.4	16.3

Lower is better

Ours; each cell annotated *bpp / held-out MagFace* identity similarity). All reconstructions are native 512^2; our model preserves identity at a comparable extreme-low rate while competitors drift or degrade. Where OneDC appears at a lower bitrate than ours, note that its rate ladder is sparse: its next operating point (0.0106 bpp, 2.3× ours) still attains a lower test-set held-out identity than ours at 0.0045 bpp (0.550 vs. 0.586 ArcFace (Deng et al., 2019)).

M Additional GFVC results

Fig. 23 gives the rate–distortion curves behind the BD-rate figures in Sec. 6: pooled over the 33 CTC sequences, replacing the VVC-intra key frame with our codec (both followed by the same FOMM (Siarohin et al., 2019) animation backbone) shifts the whole clip's RD curve, with large gains in identity (ArcFace (Deng et al., 2019)) and perceptual/distributional quality (DISTS (Ding et al., 2022), PatchFID (Mentzer et al., 2020), LPIPS (Zhang et al., 2018)) at a fraction of the rate, while PSNR is comparable (≈ 18 dB). Fig. 24 shows two further sequences: as in the main text, only the key frame is image-compressed – by VVC-intra (middle row) or our identity-aware codec (bottom row); the remaining frames are GFVC keypoint renderings animated from it.

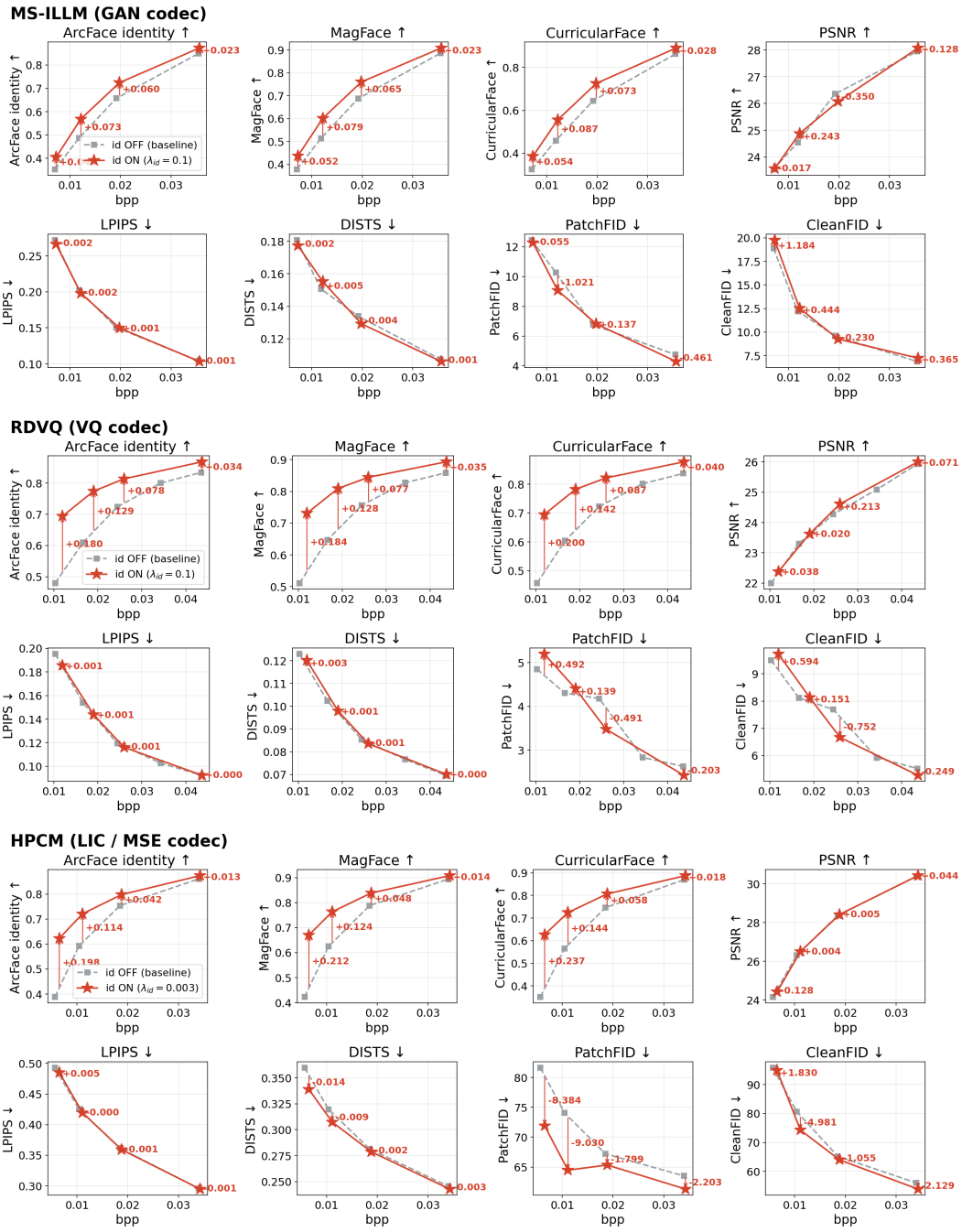


Figure 18: **Per-metric RD curves across codecs, with vs. without the identity loss.** For each codec, id-off arm (gray, dashed) vs. the same codec trained with our identity objective (red) across eight metrics; annotations mark the id-on − id-off gap at each rate point. Top two blocks are generative codecs (MS-ILLM, RDVQ) at λ_{id}=0.1; the bottom block is HPCM, a non-generative MSE + rate LIC, at λ_{id}=0.003 (matched identity gain). On HPCM the identity panels rise while the distortion panels stay on the anchor and the distributional panels (DISTS, PatchFID, CleanFID) fall below it. Summarized as BD-rate in Table 13.

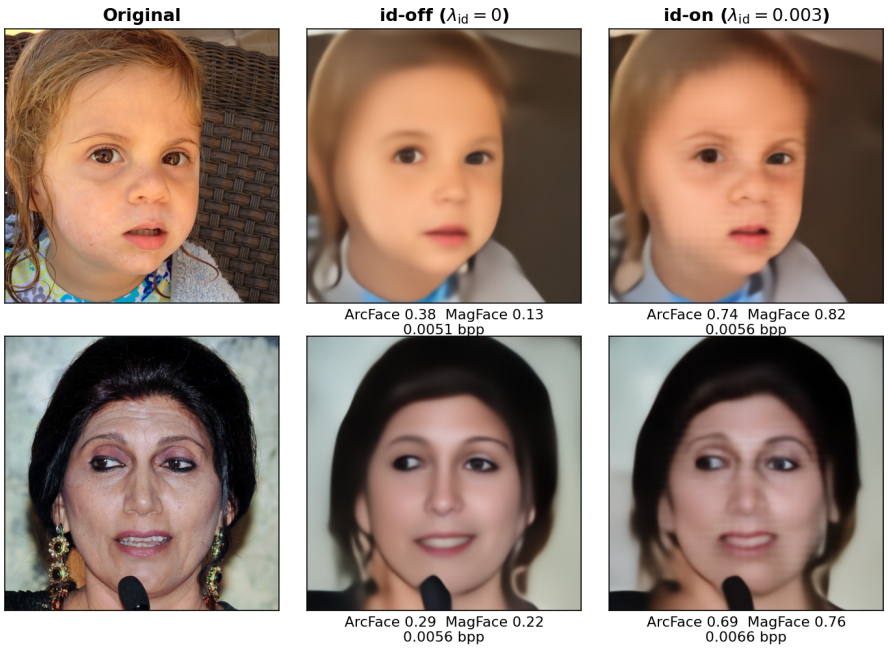


Figure 19: **Identity drift and its recovery on the non-generative HPCM codec (lowest rate, λ_{rate}=0.00003).** Original vs. id-off (λ_{id}=0) vs. id-on (λ_{id}=0.003) reconstructions at nearly equal bitrate. Without the identity loss the MSE-optimized codec produces a smooth, plausible, but *different* person (low ArcFace/MagFace to the original); adding the loss restores the correct facial structure at essentially the same rate. These are strong-recovery examples chosen for illustration.

N LIMITATIONS: QUALITATIVE EXAMPLES

Fig. 25 illustrates the two limitations discussed in Sec. 7: non-face hallucination by our intra codec (a) and inter-frame warping by the GFVC backbone (b).

Non-face hallucination is generic to generative codecs. To check that the qualitative effect of Fig. 25(a) is not specific to our codec, we quantify it over the *pure-background* region only – CelebAMask-HQ label *background*, with hair, hat, and clothing excluded so that regions our identity supervision indirectly helps do not confound the comparison. On a 500-image FFHQ subset we parse each ground-truth image, restrict all metrics to the background mask, and report background LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022) against bitrate (Fig. 26). The picture is unambiguous: our codec deviates from the true background *less* than AEIC-ME (Zhang et al., 2025) and RDVQ (Jiang et al., 2026) across the whole rate range and is on par with OneDC (Xue et al., 2025), so it is not a worse offender than other generative codecs. Most tellingly, the curves for our model *with* and *without* the identity loss are essentially identical – the identity objective neither causes nor worsens non-face hallucination; the effect is a property of extreme-low-rate generative decoding shared by all codecs in this class. (A background sharpness ratio near one for every codec confirms the non-face region is *synthesized*, i.e. hallucinated, rather than blurred away.)

Residual identity drift at the extreme-low rate. The identity loss reduces but does not *eliminate* drift. Fig. 27 shows the eight FFHQ test images with the *lowest* reconstructed identity under our model at λ_{rate}=32 (≈0.004 bpp). Even with the loss these reconstruct as high-quality but perceptually *different* people (identity ∼0.2–0.3 vs. a test-set mean of 0.62); the loss still helps on most of them over the loss-free model, though a residual gap remains and one case is marginally worse with the loss. Such cases are rare — only 21 of 3,500 fall below an identity of 0.45 — and cluster at unusual expression, lighting, or heavy accessories.

We rank and annotate identity here with *MagFace* (Meng et al., 2021) using a *fixed* 0.75 center-crop (*no* face detector), for two reasons. First, MagFace is held out from training — the loss supervises only CurricularFace (Huang et al., 2020) — so the ranking is not circular. Second, our default

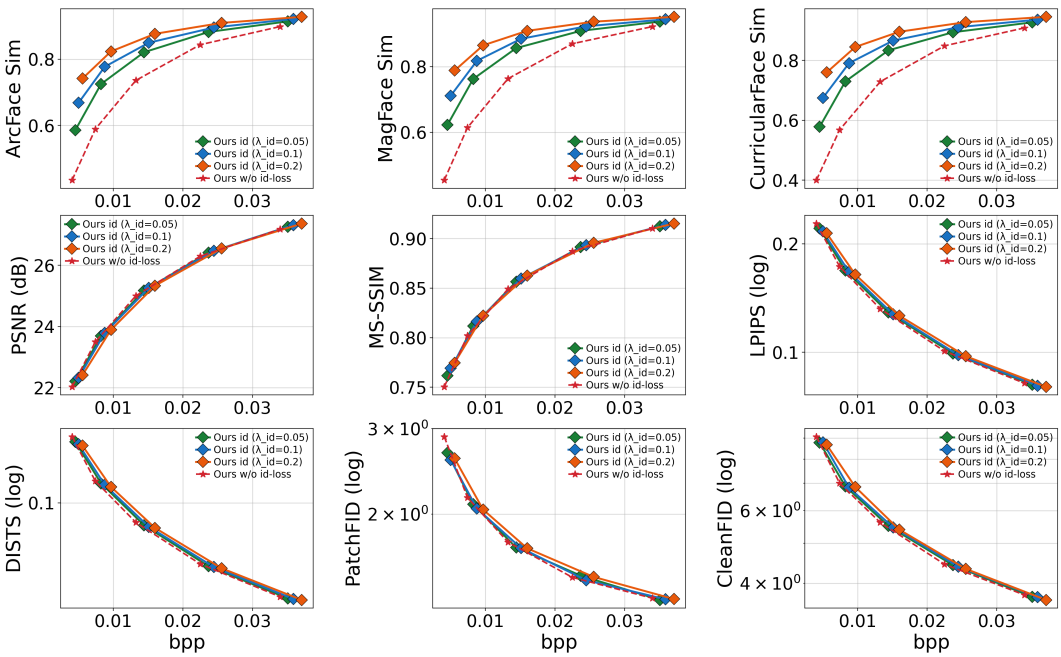


Figure 20: **Identity-weight (λ_{id}) ablation on FFHQ (per-metric RD curves).** Our codec at $\lambda_{\text{id}} \in \{0.05, 0.1, 0.2\}$ vs. Ours ($\lambda_{\text{id}}{=}0$) (red dashed). Summarized as BD-rate in Table 14.

identity eval extracts embeddings via insightface *detection*, which fails on tightly-cropped, occluded, or large-pose faces and then returns a near-zero *detection-failure* cosine that is a metric artifact rather than identity drift; a fixed center-crop — valid because these faces are FFHQ-aligned — always embeds the visible face consistently, so a low cosine reflects genuine drift. (Ranking by the detection-based metric instead surfaces occlusion artifacts — sunglasses, hands, poles — not true failures.)

O Resolution-agnostic (multi-resolution) rate–distortion

The codec is made *resolution-agnostic* by a multi-resolution finetune: at each step one training resolution is drawn from $\{256, 384, 512, 640, 768, 896, 1024\}$ (weighted toward the 512 anchor and the $256/1024$ ends) and the whole batch is downsampled to it on the fly, while a zero-initialised per-block *resolution FiLM* — conditioned on a continuous scale $s{=}\log_2\sqrt{HW}/512$ — adapts the codec across resolutions at no transmitted cost (the decoder recomputes s from the signalled (H, W)). One such model replaces a family of per-resolution codecs.

Single model vs. dedicated per-resolution codecs. Fig. 28 isolates the cost of resolution-agnosticism itself: the one multi-resolution model against five dedicated single-resolution codecs (each finetuned and evaluated at a single fixed scale with no resolution conditioning), on the *full* 3,500-image test set at practical bitrate at all three resolutions. At the 512 anchor the two coincide — identity and perceptual curves overlap and the single-resolution codec leads by only $\lesssim$0.3 dB PSNR. At 256 and 1024 the multi-resolution model *dominates* the single-resolution ones on every axis — identity (ArcFace, MagFace, CurricularFace), perceptual (LPIPS, DISTS, PatchFID, CleanFID), *and* PSNR — so one scale-invariant model is not merely as good as a codec family but strictly better away from the anchor, consistent with multi-scale training regularising the ultra-low-rate regime.

Figure 29 evaluates this single model against competitors — MS-ILLM (Muckley et al., 2023), OneDC (Xue et al., 2025), RDVQ (Jiang et al., 2026), GLC (Jia et al., 2024), and VTM (Bross et al., 2021) — at $256/512/1024$; Fig. 30 summarises the same as a per-resolution BD-rate radar with the multi-res model as the anchor. To keep each panel a matched comparison, 512 uses the full 3,500-image test set with practical (real-payload) bitrate, while 256 and 1024 use the 500-image ref-



Figure 21: **FFHQ reconstruction samples** (the first nine test images of the FFHQ test split, excluding `00053`, which appears in Fig. 1). Columns: Original / AEIC-ME / OneDC / MS-ILLM / VTM / DiffC / Ours (λ_{id}=0) / Ours; each cell is annotated with its *per-image bpp / held-out MagFace* (OneDC and DiffC are fixed-rate here, hence a constant bpp). Cells are downsampled to 256×256 for the figure.



Figure 22: **Cross-dataset (OOD) reconstruction samples on CelebA-HQ** (first eight test images, `celebahq_00000–00007`). Columns: Original / AEIC-ME / OneDC / MS-ILLM / VTM / DiffC / Ours (λ_{id}=0) / Ours; each cell is annotated with its *per-image bpp / held-out MagFace* (OneDC and DiffC are fixed-rate here, hence a constant bpp). Our model is trained only on FFHQ. Cells are downsampled to 256×256 for the figure.

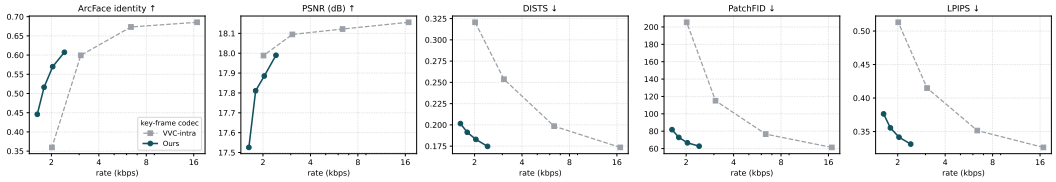


Figure 23: **GFVC key-frame-swap rate–distortion** (FOMM backbone, pooled over the 33 official CTC sequences, 512^2 Class C&D). Total video bitrate vs. each metric; VVC-intra QP $\{22, 32, 42, 52\}$ (anchor) vs. our λ sweep.

erence subset on which the competitors were evaluated; released/pretrained competitor variants only (finetuned versions excluded); bitrate accounting follows App. E (GLC alone uses its entropy-model rate estimate). Across all three resolutions the resolution-agnostic model leads on the held-out identity metrics (ArcFace, MagFace) and the perceptual/distortion metrics (LPIPS, DISTS, PatchFID, CleanFID) at ultra-low bitrate, and is competitive on PSNR — in the radar every competitor lies inside the anchor ring (i.e. needs more bitrate to match our quality) on nearly every non-PSNR axis. Fig. 29 also carries the same model trained *without* the identity loss ($\lambda_{\mathrm{id}}{=}0$, lighter curve): the loss remains near-free across the whole 256–1024 range — it lifts all three identity metrics (ArcFace, MagFace, CurricularFace; largest at low rate, e.g. $+0.15$–0.17 ArcFace/MagFace at the lowest rate, tapering to ~ 0.02 at the highest) while PSNR stays within 0.2 dB and the perceptual/distributional metrics stay on the id-off curve at every resolution.

P In-the-wild large-face operation

The main-set and CelebA-HQ evaluations use aligned crops; here we stress the codec on *unconstrained, large, mixed-resolution* faces from the open web.

Data. We use **Flickr-2W**, a 2017 crawl of $20{,}745$ natural Flickr photographs. Individual file timestamps predate the FFHQ release by ~ 1.5 years, and matching the $20{,}727$ Flickr photo-ids against the official FFHQ metadata yields *zero* shared photos — so this set is causally independent of our FFHQ training data and serves as an unseen-identity generalization test. We keep only photographs containing exactly one detected face and crop a square at the FFHQ face ratio (face$=0.272$ of the frame, background included), *pose-preserved* (no similarity alignment, so the crop stays in-the-wild) and constant-padded (mid-grey) at the borders, resized to the FFHQ-equivalent resolution rounded to a codec-valid multiple of 64. Keeping only faces whose effective resolution lands in $[256, 1024]$ gives $3{,}619$ **single-face crops** spanning 256–1024 (mode ~ 320) — a genuinely large-face regime, complementary to the tiny aligned crops of CPLFW/CFP.

Protocol. There are no identity labels, so identity is measured *orig-vs-reconstruction*: the ArcFace embedding of the reconstruction against that of the original crop. We report two crop protocols that bracket registration sensitivity: (i) a *margin*-0.6 *fixed bbox* using the same 5-point landmarks for original and reconstruction (registration-sensitive, pixel-aligned), and (ii) *norm_crop*, which re-detects and similarity-aligns each reconstruction independently and penalises a non-detected reconstruction with $\cos = 0$ (registration-invariant; $3{,}601/3{,}619$ originals detect). Competitors follow the in-the-wild recipe of App. E: OneDC, AEIC-ME, VTM (VVC all-intra), MS-ILLM and GLC, each swept over its rate ladder at native resolution. RDVQ and DiffC are excluded because they cannot ingest mixed native resolutions in one pass (DiffC resizes every image to a fixed square, which would break the shared-landmark protocol), matching their absence from the CPLFW/CFP in-the-wild comparison.

Result. Our resolution-agnostic model — with no in-the-wild finetuning — leads on identity at every matched bitrate (Fig. 31). With our model as the BD-rate anchor and integrating over the measured overlap (PCHIP, App. E), OneDC needs 32.3% (norm_crop) and 19.7% (margin-0.6) more bits to match our identity; AEIC-ME needs 40.4% and 54.2% more, and the margins over GLC (47.6%/73.9%), MS-ILLM (61.7%/97.8%) and VTM (95.3%/194.2%; norm_crop/margin-0.6) are larger still. The gap is largest precisely in this large-unaligned regime — wider than on the small aligned CPLFW/CFP crops — indicating that the identity advantage is not an artifact of aligned evaluation but survives, and grows, under in-the-wild pose and scale. Figure 33 shows side-view

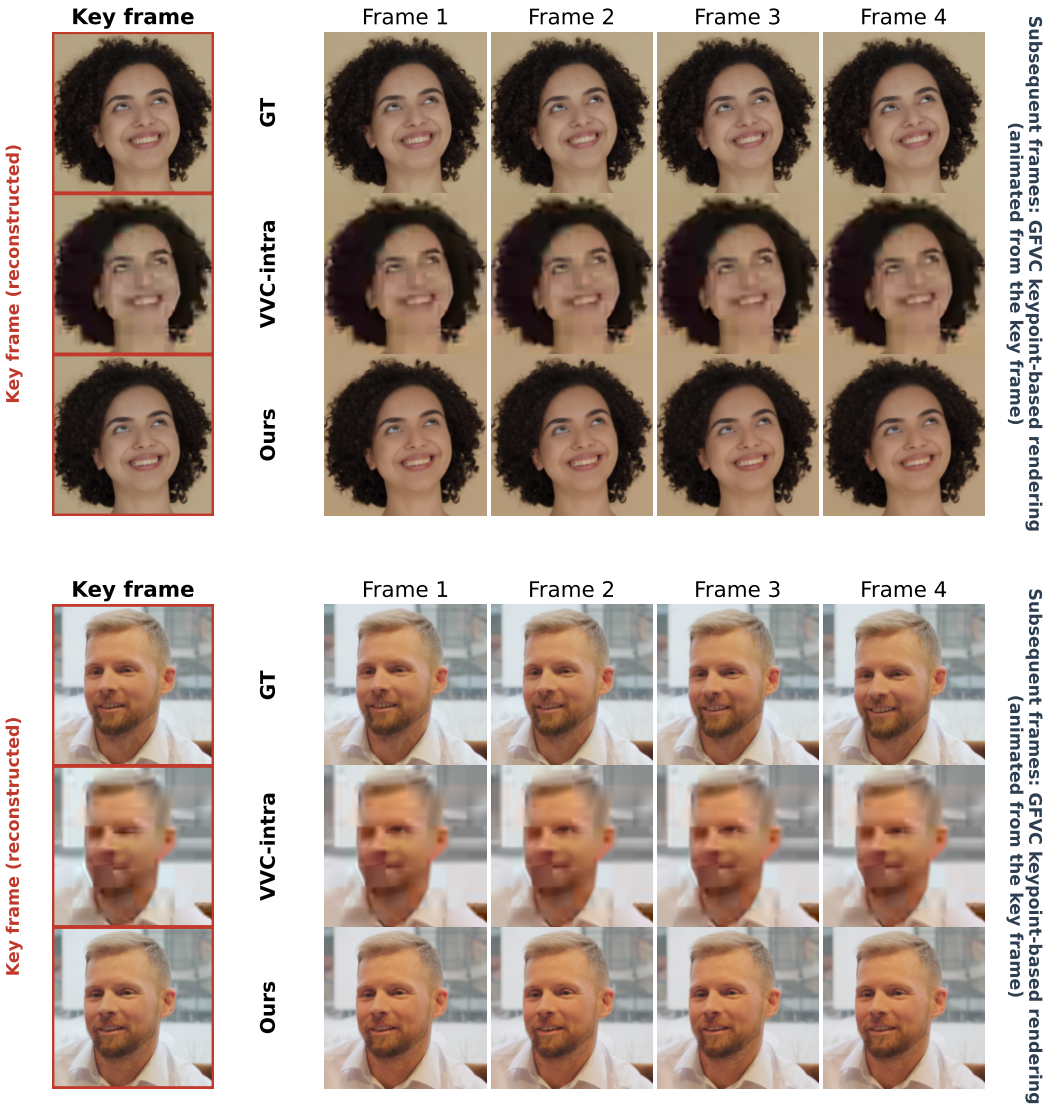


Figure 24: **GFVC key-frame replacement, two further sequences.** *Top:* ClassC_011 (comparable total rate 1.50 vs. 1.73 kbps; VVC-intra ArcFace 0.25 / PatchFID 165, ours ArcFace 0.62 / PatchFID 48). *Bottom:* ClassD_006 (1.99 vs. 2.17 kbps; VVC-intra ArcFace 0.44 / PatchFID 302, ours ArcFace 0.74 / PatchFID 85). Only the key frame is image-compressed (VVC-intra middle row / ours bottom row); the remaining frames are GFVC keypoint renderings animated from it. Frames shown at 256×256.



(a) Non-face hallucination (b) GFVC inter-frame warping (backbone, not our codec)

Figure 25: **Limitations. (a)** Non-face content is still synthesized: the cap logo and the region left of the face (red circles) are re-synthesized, while the face identity is preserved. **(b)** A GFVC mid-frame (#0125): the animation backbone's keypoint warping distorts the face under large motion – a backbone limitation, not our intra codec.

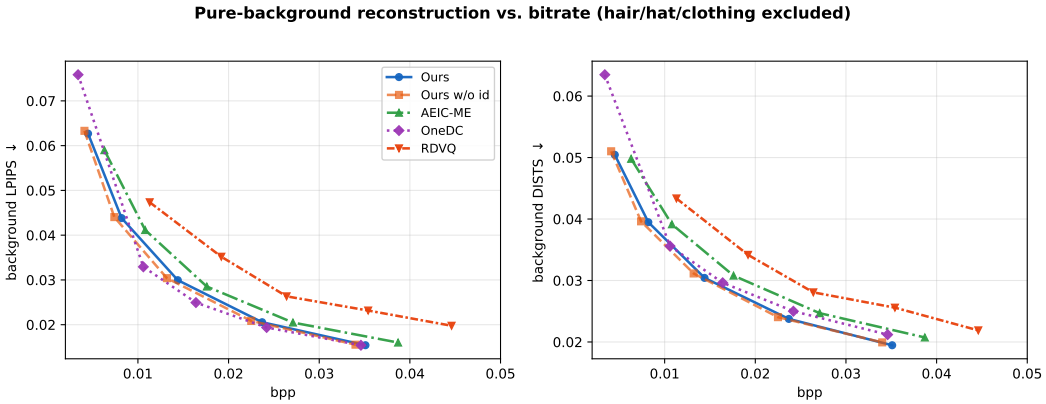


Figure 26: **Pure-background rate–distortion (hair/hat/clothing excluded),** 500**-image FFHQ subset.** Background-region LPIPS (left) and DISTS (right) against bitrate; lower is better. Our codec (with and without the identity loss, which overlap) deviates less than AEIC-ME and RDVQ and matches OneDC, indicating our model does not hallucinate non-face content more than other generative codecs, and the identity loss adds none.

cases where our reconstruction preserves the person's identity at ultra-low rate while OneDC and the other baselines drift to a plausible but visibly *different* person. The identity gain is *near-free* in this in-the-wild regime as well: the same model trained without the loss ($\lambda_{\text{id}}=0$) traces the *same* PSNR/MS-SSIM/LPIPS/DISTS rate–distortion curves (Fig. 32; $|\Delta\text{PSNR}|\leq 0.17$ dB, the perceptual axes indistinguishable), so identity is the only axis that moves — the loss buys it at no distortion or perceptual cost, exactly as on aligned FFHQ.

Q Demographic subgroup fairness

Recognition networks carry known demographic biases, so we check whether the near-free identity gain is *uniform* across subgroups or concentrated in some of them. We stratify the full 3,500-image FFHQ test set by three attributes derived from the ground-truth image only (the label is shared by every codec): **gender** and **age**, predicted by the InsightFace buffalo_l gender/age head, and **skin**



Figure 27: **Failure cases: the eight lowest-identity reconstructions** (λ_{rate}=32, $\approx$0.004 bpp), ranked by held-out MagFace with a fixed 0.75 center-crop (no detector, so occlusion/pose cannot create a detection-failure artifact). Columns: Original / AEIC-ME / OneDC / MS-ILLM / VTM / DiffC / Ours (λ_{id}=0) / Ours; each cell is annotated with its *bpp / MagFace* (in red). The generative codecs synthesize a plausible but *different* person while VTM/MS-ILLM over-smooth; the identity loss usually helps yet does not fully close the gap on these hardest cases. Cells are downsampled to 256×256 for the figure.

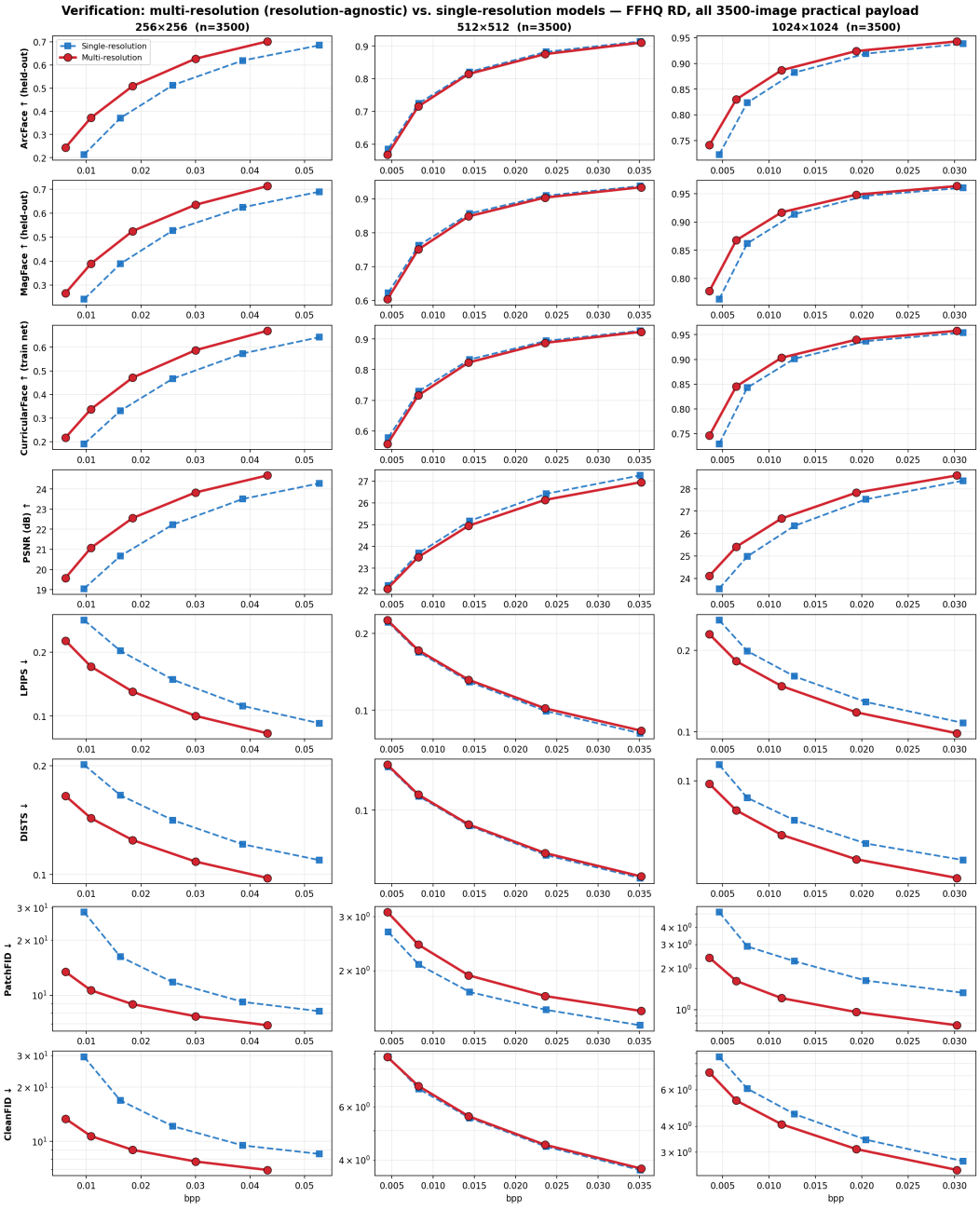


Figure 28: **One resolution-agnostic model vs. dedicated single-resolution codecs** (FFHQ, full 3,500-image test set, practical bitrate). Red = the single multi-resolution model; blue dashed = per-resolution codecs (one finetune per scale, no resolution conditioning). At 512 they coincide (the single-resolution codec is ahead by $\lesssim$0.3 dB PSNR only); at 256 and 1024 the multi-resolution model is better on *every* metric, PSNR included.

tone, summarized by the Individual Typology Angle (ITA) (Chardon et al., 1991), the model-free CIELab descriptor ITA $= \arctan\big((L^* - 50)/b^*\big)$ standard in skin-tone fairness audits. Skin pixels are isolated with a face-parsing mask and ITA is taken from the median L^*, b^* after removing specular, shadow, and low-chroma pixels; we deliberately avoid a race/ethnicity classifier. ITA is a *photometric* descriptor confounded by illumination, so its bins index skin brightness in the capture, not ethnicity. For each image we compute the identity gain $\Delta = \cos(\hat{\boldsymbol{x}}_{\text{on}}, \boldsymbol{x}) - \cos(\hat{\boldsymbol{x}}_{\text{off}}, \boldsymbol{x})$ averaged

Table 15: **The near-free identity gain is uniform across demographic subgroups** (FFHQ, held-out MagFace, mean over the five rates). Δ is the id-on minus id-off cosine; "id-off" and "id-on" are the absolute identities of the $\lambda_{\mathrm{id}}{=}0$ anchor and our model. The gain varies by at most 0.005 across subgroups within each attribute, and the between-subgroup identity spread does not grow with the loss (it *shrinks* for gender and age). CurricularFace gives the same picture (Δ in $[0.098, 0.105]$ across all subgroups).

Attribute	Subgroup	n	Δ MagFace	id-off	id-on
Gender	M	1611	$+0.089$	0.731	0.820
	F	1889	$+0.096$	0.719	0.816
Age	$<$30	991	$+0.091$	0.734	0.826
	30–45	1244	$+0.094$	0.722	0.816
	$>$45	1265	$+0.093$	0.720	0.814
Skin tone (ITA)	Light	1162	$+0.095$	0.725	0.820
	Intermediate	637	$+0.093$	0.726	0.819
	Tan–Dark	1689	$+0.091$	0.724	0.815
Overall	—	3500	$+0.093$	0.725	0.818

over the five rate points, using the two *held-out* recognizers (MagFace and CurricularFace), and report per-subgroup means with 95% bootstrap confidence intervals.

Table 15 and Fig. 34 show the gain is essentially constant: Δ lies in $[+0.089, +0.096]$ (MagFace) across all eight subgroups, within 0.005 of the overall $+0.093$, and every 95% interval excludes zero by a wide margin. The loss also does not amplify pre-existing disparities: the between-subgroup spread of absolute identity is small to begin with (≤ 0.014 cosine) and does not increase with the loss — it *shrinks* for gender ($0.012 \to 0.005$) and age ($0.014 \to 0.012$) and stays negligible for skin tone (≤ 0.005 either way). We therefore find no evidence of bias amplification: the near-free identity benefit is delivered uniformly across the gender, age, and skin-tone strata we can measure. The main caveat is inherited from the evaluation networks and the ITA proxy, both of which are photometric and demographically imperfect; a study with self-reported attributes and human raters is future work.

R Complexity: latency, model size, and identity efficiency

Protocol. We report per-image wall-clock encode/decode latency at 256^2, 512^2, and 1024^2, plus parameter counts, on a single Quadro RTX 8000 (the App. A hardware), fp32. Neural codecs are timed with `torch.cuda.synchronize()` around `compress`/`decompress`; the first four images are discarded as warm-up and we average the next twenty. VTM (VVC intra, QP 60) is timed on CPU via the reference binary's reported total time. Latency is essentially rate-independent for the neural codecs (fixed architecture); parameter count is the size metric (peak GPU memory is dominated by the fp32 caching allocator and is not comparable across environments). The reference row is the *multi-step* diffusion codec DiffEIC, included only to contrast sampling cost; DiffC is a *multi-step* diffusion competitor and is timed on the same footing. Cross-environment absolute milliseconds are indicative; the robust quantities are the ratios within one harness and the number of U-Net forward passes.

Findings. (i) *One step beats multi-step.* As a single-forward decoder, IdentiCodec decodes in 167.6 ms, versus 7.1 s for DiffEIC (50 forwards, $\sim 43\times$) and ~ 1 s for DiffC ($24{+}27$ forwards, $\sim 6\times$); cost tracks the number of U-Net forwards, not model size. (ii) *Best among one-step diffusion codecs.* At matched rate (Fig. 35) our decode is $3.1\times$ faster than OneDC and our encode $\sim 10\times$ faster, at markedly higher identity – OneDC's heavier cross-attention U-Net and full-size VAE cost more per forward. (iii) *The identity loss is free.* It is training-only, so IdentiCodec is the very same network as its identity-off anchor (Ours, $\lambda_{\mathrm{id}}{=}0$) – identical parameters and latency – while gaining $+0.11$ ArcFace at a matched ultra-low bitrate. Finally, the resolution-agnostic model of Sec. 6 adds only the per-block resolution-FiLM heads – 5.2M parameters (0.47% of the model) and a few milliseconds of latency – so one scale-invariant model replaces a per-resolution codec family at little cost.

Table 16: **Inference complexity** (RTX 8000, per-image, warm-up$=4$, mean of 20; latencies in ms). Enc./Dec. $=$ encode/decode; diff. $=$ diffusion; the parenthesised count is the number of U-Net forward passes per image (for DiffC, encode $+$ decode). Our one-step decoder is a single forward; the DiffC competitor and the DiffEIC reference row are multi-step. Competitors: AEIC-ME, OneDC, MS-ILLM, VTM, DiffC, DiT-IC. The identity loss is training-only, so "Ours (λ_{id}=0)" and IdentiCodec share parameters and latency exactly; the multi-resolution model adds 5.2M FiLM parameters (0.47%) at a few-ms latency overhead. At 1024^2 our models and AEIC-ME run the single-forward (non-tiled) decoder.

			256^2		512^2		1024^2	
Codec	Type (U-Net fwd.)	Params	Enc.	Dec.	Enc.	Dec.	Enc.	Dec.
IdentiCodec (ours)	1-step	$\sim$1.10B	15.2	69.5	18.7	167.6	65.6	793
Ours (λ_{id}=0)	1-step	$\sim$1.10B	15.2	69.5	18.7	167.6	65.6	793
Ours (multi-res.)	1-step	$\sim$1.10B	18.4	74.1	21.6	171.2	68.6	799
AEIC-ME	1-step	$\sim$1.08B	16.5	70.0	19.7	168.0	65.3	793.8
OneDC	1-step	$\sim$1.32B	81.9	250.0	182.5	520.3	672	1929
DiT-IC	1-step	$\sim$1.07B	56.4	106.1	145.5	243.8	508	870
MS-ILLM	N/A (GAN)	181.5M	29.2	35.7	53.3	78.9	154	242
VTM	N/A	—	$\sim$601	$\sim$46	$\sim$1012	$\sim$53	$\sim$2190	$\sim$75
DiffC	multi-step (24+27)	$\sim$1.07B	704	707	1074	1037	5092	4965
DiffEIC *(ref.)*	multi-step (50)	$\sim$1.38B	81.0	3134	233.8	7146	607	28881

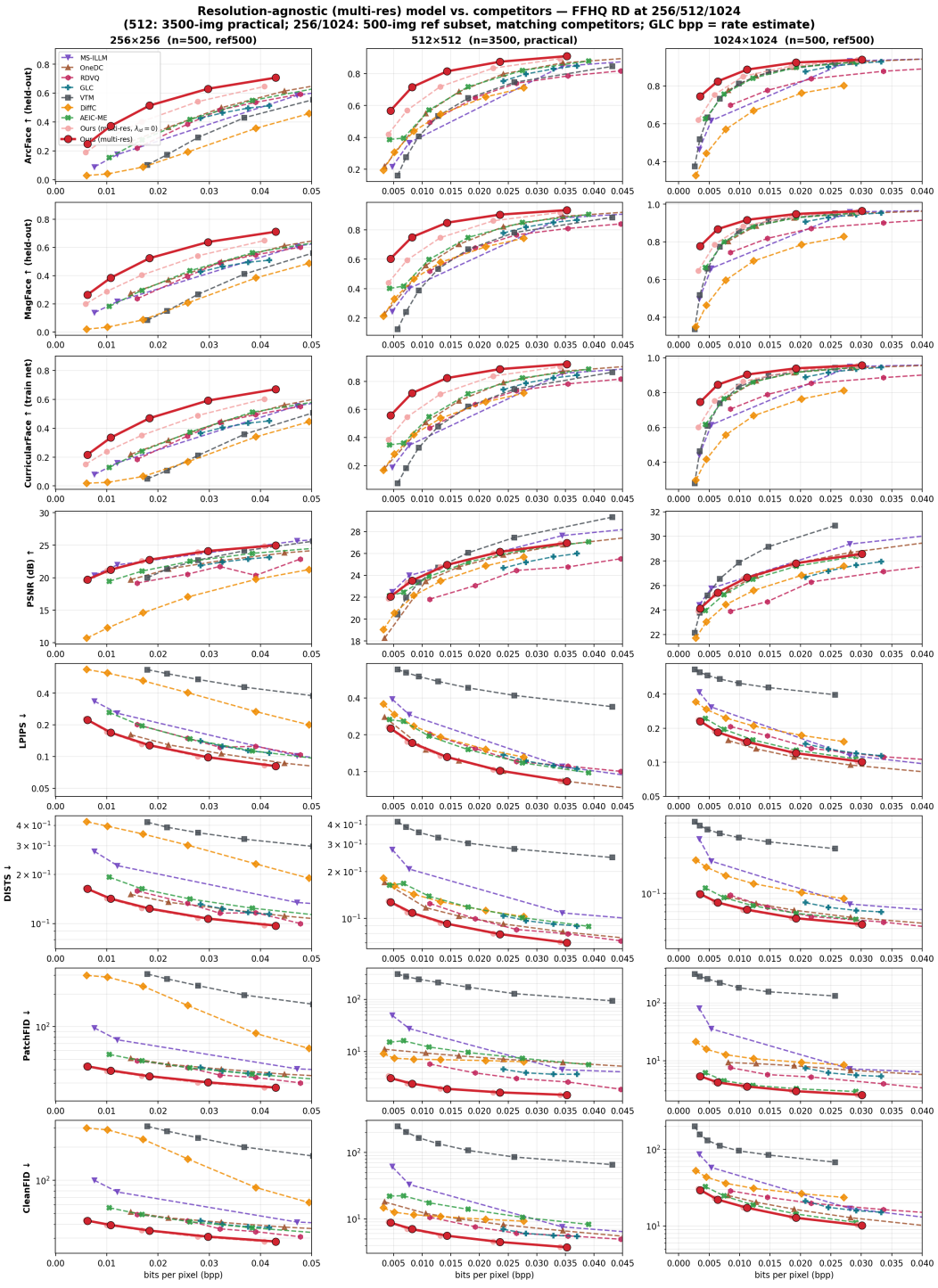


Figure 29: **Resolution-agnostic (multi-res) model vs. competitors — FFHQ rate–distortion at** $256/512/1024$**.** Held-out identity (ArcFace, MagFace), the training-net CurricularFace, and distortion/perceptual metrics against bitrate; red is our single multi-resolution model and the lighter-red dashed curve is the *same* model trained without the identity loss ($\lambda_{\text{id}}=0$): the loss lifts identity (largest at low rate) while leaving PSNR within 0.2 dB and the perceptual/distributional metrics on the id-off curve — near-free at every resolution. Sample basis per panel as in the text (512: 3,500-image practical; $256/1024$: 500-image reference subset matching the competitors). Fine-tuned competitor variants are excluded; GLC's bitrate is its rate estimate (App. E).

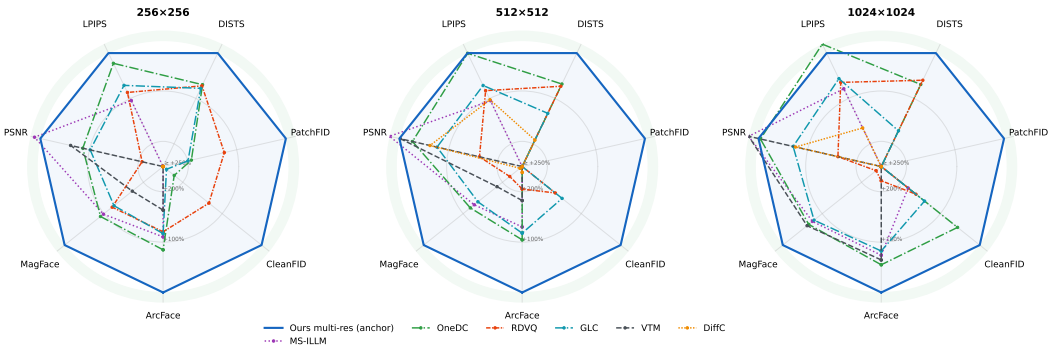


Figure 30: **Per-resolution BD-rate radar, multi-res model as anchor.** Each axis is a metric; the blue ring is the multi-resolution model (0%). A competitor plotted *inside* the ring needs *more* bitrate to match our quality on that axis; outside means less (capped at +250%). Held-out identity and perceptual axes are dominated at every resolution; PSNR is the only axis where competitors can sit outside the ring, consistent with our identity-first objective (we trade pixel fidelity for identity). BD-rate is the Bjøntegaard integral of log-bitrate over the measured quality overlap, interpolated with a *monotone piecewise-cubic Hermite* (PCHIP) spline rather than a single global cubic: on the competitors' sparse 4–5-point rate ladders a global cubic overshoots (Runge) when integrated across a measurement gap, whereas the shape-preserving PCHIP stays between the measured points (on densely sampled axes the two agree). An axis with no measured competitor operating point inside the overlap is omitted rather than extrapolated.

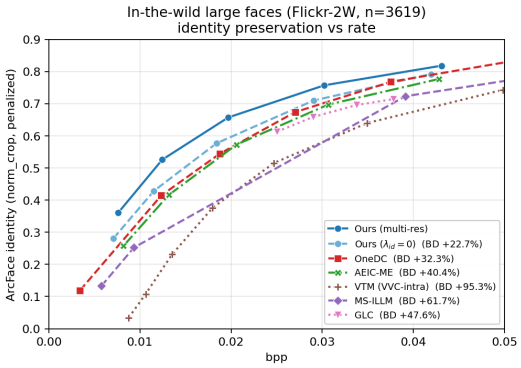


Figure 31: **In-the-wild large-face rate–distortion (Flickr-2W,** n=3,619 **single-face, native** 256–1024**).** Orig-vs-reconstruction ArcFace identity vs. bitrate under the registration-invariant norm_crop protocol (each reconstruction re-detected and similarity-aligned independently; non-detections penalised); blue is our single resolution-agnostic model. It leads every competitor across the operating range; per-competitor BD-rate vs. ours (more bitrate to match our identity) is in the legend. The light-blue dashed curve is the same model without the identity loss (λ_{id}=0): the loss adds a 22.7% identity BD-rate margin, yet even the id-off model already beats every competitor.

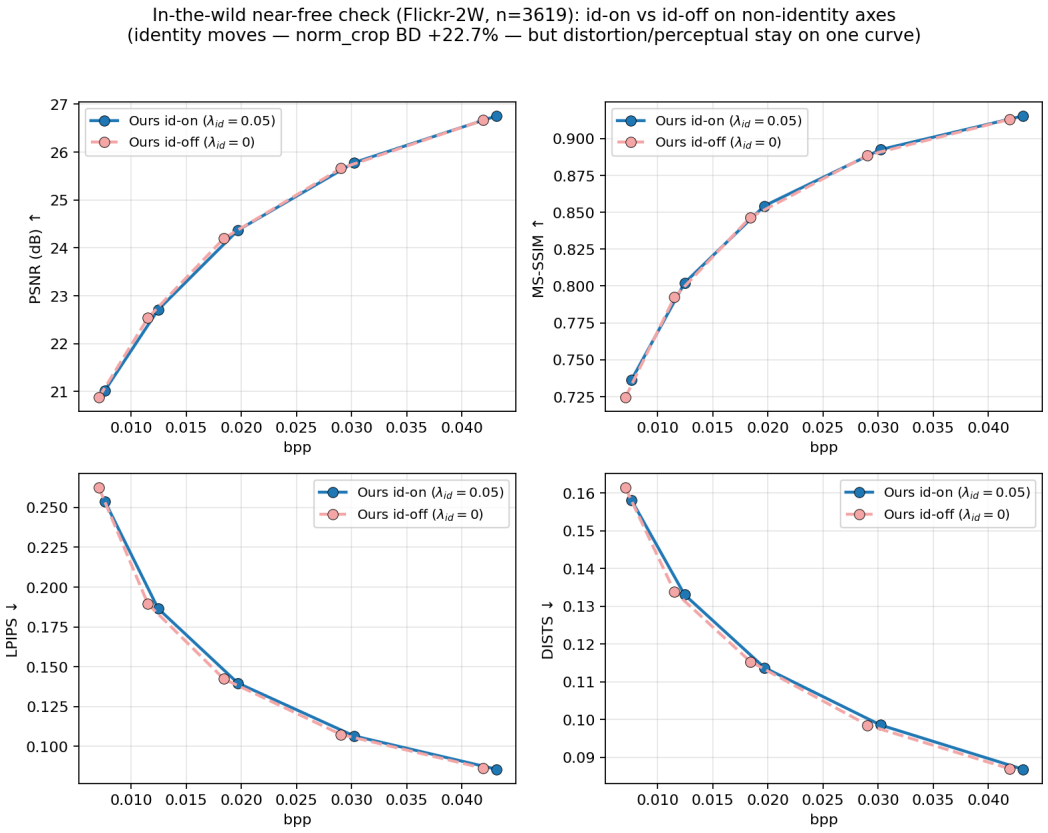


Figure 32: **Near-free in-the-wild (Flickr-2W, n=3,619): identity loss on vs. off on the *non-identity* axes.** PSNR/MS-SSIM/LPIPS/DISTS vs. bitrate for the resolution-agnostic model with (λ_{id}=0.05) and without (λ_{id}=0) the identity loss. The two curves overlap on every quality axis ($|\Delta\text{PSNR}|\leq 0.17$ dB, $|\Delta\text{LPIPS}|\leq 0.009$, $|\Delta\text{DISTS}|\leq 0.003$), while identity alone moves (BD +22.7%, Fig. 31) — the near-free property holds on large, unaligned in-the-wild faces, not only on aligned FFHQ.



Figure 33: **In-the-wild side-view examples.** Three large unaligned faces (rows); each column is a codec at comparable ultra-low bitrate ($\approx$0.01 bpp). Our codec preserves the side-view identity-bearing structure (eyes, nose bridge, jawline), whereas OneDC and AEIC-ME re-frontalise the head into a plausible but *different* person, VTM collapses into blocking, and MS-ILLM blurs the face away. GLC is omitted from this visual comparison because its lowest available rate is more than $2\times$ that of the other codecs, which precludes a rate-matched panel. The per-image bitrate and orig-vs-recon ArcFace cosine (norm_crop) are printed below each cell; Ours has the highest identity in every row.

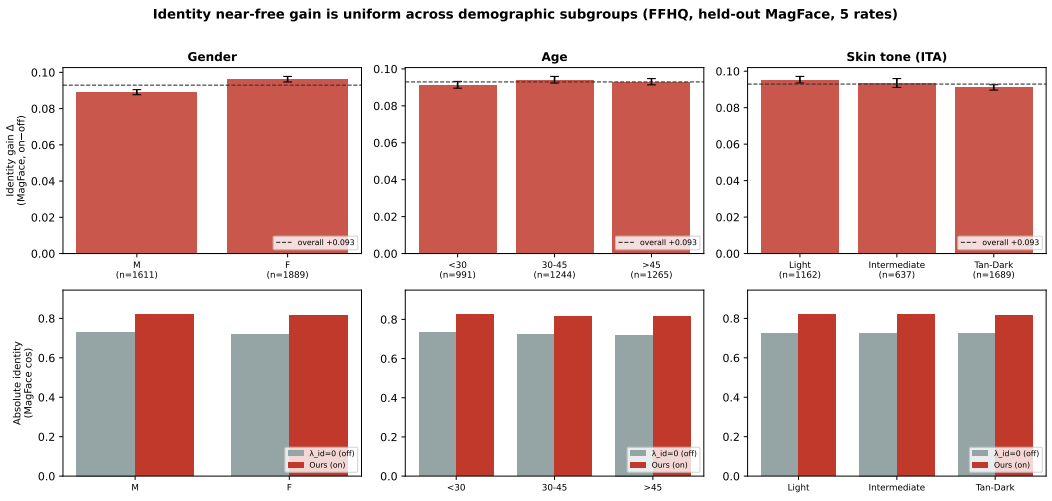


Figure 34: **Per-subgroup identity gain and absolute identity** (held-out MagFace). *Top:* the gain Δ (on$-$off, averaged over five rates) with 95% bootstrap CIs; every subgroup sits at the overall $+0.093$ (dashed) within 0.005. *Bottom:* absolute identity of the $\lambda_{\text{id}}{=}0$ anchor (grey) and our model (red); the loss lifts every subgroup by $\approx$0.09 and leaves no subgroup behind.

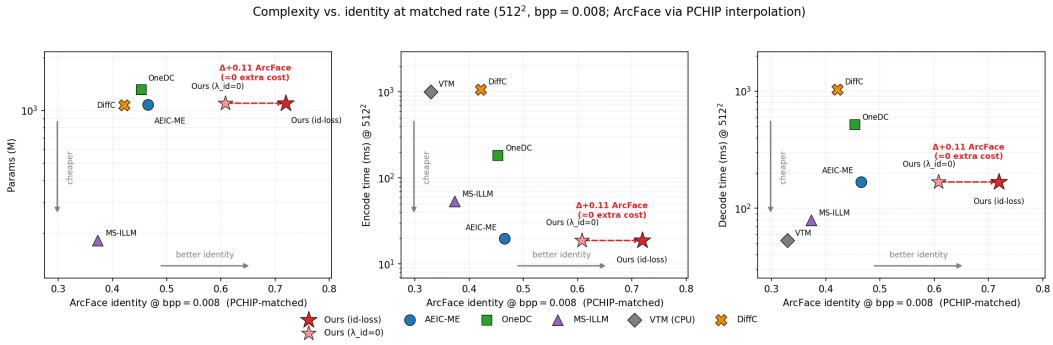


Figure 35: **Complexity vs. identity at a matched rate** (512^2)**.** For each codec we read ArcFace off its rate–distortion curve at a common bitrate (0.008 bpp) by shape-preserving PCHIP interpolation (exact bitrate matching is impossible across codecs), then plot it against parameters, encode time, and decode time. Up-and-to-the-right is worse; the ideal codec is low (cheap) and right (high identity). Our model attains the highest identity at low latency; the red arrow marks the identity loss lifting ArcFace by $+0.11$ at *identical* parameters and latency ($\lambda_{\text{id}}{=}0 \rightarrow 0.05$).