LW-MIH: GENERATIVE MULTI-IMAGE HIDING WITH VAE PACKING AND LEARNABLE WHITENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-image hiding conceals several secret images in a single image sent over a public channel, but the two existing families leave a gap: cover-based methods reach high capacity yet modify the cover and leave traces of the secrets, whereas coverless generative methods are secure yet hide only a few images. Neither states how many images one cover can carry, nor how to trade that number against per-image fidelity. We propose LW-MIH, which answers both questions. It encodes every secret with a frozen foundation VAE — FLUX, SD1.5 or Sana, used only as secret encoders — and packs the latents into one tensor of $256\times256\times3 = 196{,}608$ elements, so the geometric capacity follows from the encoder geometry: up to 12 images with FLUX, 48 with SD1.5 and 96 with Sana. It then fills that fixed tensor in one of three ways — full capacity, high-frequency dual-slot and upsampling — which turns the number of hidden images and the fidelity of each into a discrete set (FLUX 12/6/3, SD1.5 48/24/12/3, Sana 96/48/24/6). Our main contribution is a *learnable invertible whitener* (LIW): an invertible neural network maps the packed tensor to the noise end of a separate AFHQ-Cat rectified flow, and LIW then aligns that field with the flow's Gaussian prior without adding representable capacity. Every learned map is exactly invertible, so the recovered images are bounded only by the frozen encoder and by 8-bit cover quantization. At full capacity the 16-bit PSNR of the recovered secrets is 35.03, 28.17 and 26.89 dB for FLUX, SD1.5 and Sana — within 0.15, 0.07 and 0.06 dB of the ceiling of the encoder that produced them — and rises to 41.03, 37.08 and 37.91 dB at the fewest-image operating point; the corresponding 8-bit values are 24.12, 22.77 and 20.18 dB. On SD1.5, LIW reaches 22.77 vs. 22.88 dB for the unmixed pipeline, so it keeps that quality while avoiding the loss of a random mixer; LIW gains 3.97 dB of 8-bit PSNR over such a mixer on FLUX and 3.36 dB on SD1.5, and at full capacity its worst per-group deficits from the encoder ceiling are 0.30, 0.41 and 0.15 dB. A random mixer instead leaves one Sana group 2.3 dB below its own ceiling, and with no mixer a FLUX group falls 12.4 dB and a Sana group 4.7 dB below theirs.

1 INTRODUCTION

Image hiding conceals secret information inside an image sent over a public channel (Fridrich, 2009). Following the coverless literature, we call the transmitted image the *cover* and the concealed images the *secrets*. Hiding several secrets together is called multi-image hiding (Guan et al., 2023); it suits confidential batch delivery, copyright protection and integrity verification, where sending several files separately is expensive or conspicuous. Existing methods are either cover-based or coverless.

Cover-based methods embed the secrets into a natural image, in the spatial or frequency domain or through deep invertible networks. All of them share one shortcoming: capacity is bounded by how much a natural image can be perturbed, and every perturbation leaves a measurable trace of the secrets on the transmitted image.

Coverless methods avoid modification altogether. Mapping-based methods transmit database images that match the secret's hash, so capacity is limited by the database size and the hash length. Generation-based methods synthesize the cover from the secret with a generative model (Yu et al., 2023; Yang et al., 2024; Zhu et al., 2026; Qian et al., 2026; Xu et al., 2025), which improves secu-

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.

rity because no natural image is modified, but the number of images hidden at once stays low: the earliest methods hide a single image, and coverless generation-based hiding of several images has only recently become possible (Chen et al., 2026).

Moving from one hidden image to many raises two questions that no existing method answers. First, how many images can one cover carry? Since the cover has a fixed pixel size and a frozen VAE encodes each secret into a fixed-size latent, that number follows from the encoder geometry; it changes only with the resolution at which the secrets are encoded, or with the rule used to fill the available positions. Second, how should that number be traded against the fidelity of each recovered image? More than one filling rule is admissible, so the trade-off is realised as a set of operating points from which the user selects one.

The generator raises a third question. Such a model integrates an ordinary differential equation from a sample of $\mathcal{N}(0, I)$, so the field carrying the secrets must itself resemble such a sample; yet the invertible network that produces it is trained by a likelihood that constrains only the per-coordinate distribution, so the field approaches the prior in its marginals but not in its geometry. For a minority of packed fields that deviation pushes the integration endpoint outside the range $[-1, 1]$ that pixelization can represent, and since every module is invertible the clipping is carried back through the pipeline and damages a few secret images far more than the rest.

We propose LW-MIH, a VAE-based generative framework that answers all three questions. FLUX, SD1.5 and Sana serve only as frozen secret encoders, while a separate rectified flow (Liu et al., 2023; Lipman et al., 2023), trained on AFHQ-Cat (Choi et al., 2020), synthesizes the 256×256 cover. The framework packs the encoder latents into one tensor of fixed size, fills that size under three hiding schemes, and aligns the packed field with the flow prior through a *learnable invertible whitener* (LIW). It hides up to 96 images in one cover without the wholesale failures of the untrained pipeline.

Contributions.

- **How many images one cover can carry, and the schemes that trade that number for fidelity.** The number is fixed by the frozen encoder and the tensor size through $B = n(L/f)^2C$, so one framework serves FLUX (n=12), SD1.5 (n=48) and Sana (n=96). Full capacity, high-frequency dual-slot and upsampling fill that fixed size differently, turning image count and per-image fidelity into a discrete, user-selectable set (Table 1).
- **A learnable invertible whitener (LIW)** built from a public permutation, a block-shared Cayley factor Q with $\det Q$=1, and a diagonal scale. The composite W is a well-conditioned bijection of the packed field and adds no representable capacity, so the recovery error remains bounded by the frozen encoder and the bit depth. Relative to a random orthogonal mixer of the same form, LIW raises 8-bit recovery by 3.97 dB on FLUX and 3.36 dB on SD1.5, and leaves no group more than 0.41 dB below its own encoder ceiling (§3.2, §4.3).
- **A paired ablation and a field analysis of the whitener** over three encoders and n from 3 to 96 (§4.2–§4.4), which separates what orthogonal mixing achieves from what learning adds: mixing alone removes the overflow that collapses whole evaluation groups, whereas only learning the amplitude and the tails of the field converts that into 8-bit fidelity.

2 Related Work

A. Cover-based multi-image hiding. Cover-based methods write the secrets into a natural image, so their capacity is bounded by the perturbation a natural cover tolerates. Spatial-domain methods hide the payload in the image array itself (Baluja, 2017); frequency-domain methods use DCT, DWT or DFT coefficients; and deep invertible methods such as HiNet (Jing et al., 2021), ISN (Lu et al., 2021) and DeepMIH (Guan et al., 2023) treat hiding and recovery as exact inverses and extend them to several secret images. All of them modify the cover and retain traces of the secrets.

B. Coverless generative multi-image hiding. Coverless methods synthesize the cover instead of modifying one. Generation-based methods produce the cover from the secret itself: CRoSS (Yu et al., 2023) and DiffStega (Yang et al., 2024) control a diffusion model with keys derived from the

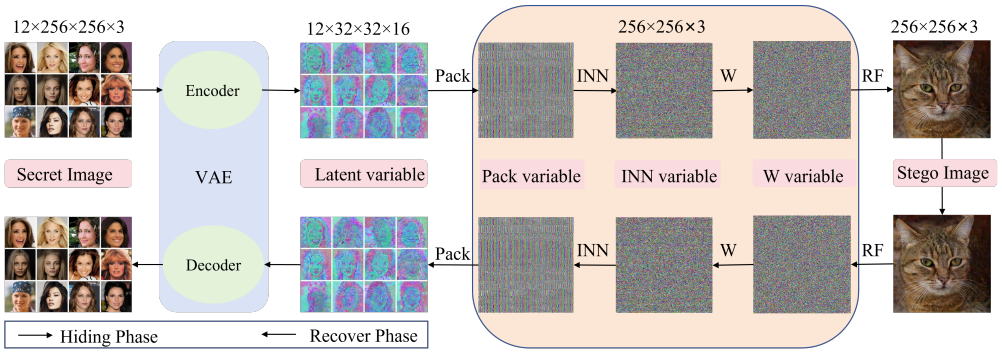

Figure 1: The hiding phase (top) and the exact-inverse recovery phase (bottom) of LW-MIH. Secrets pass through one of three hiding schemes (full capacity, upsampling, high-frequency dual-slot), are encoded by a frozen VAE, packed into the fixed-size tensor $B = 256\times256\times3$, mapped by an invertible neural network and whitened to match the generator's prior, and synthesized into a cover Y by a frozen rectified flow; recovery inverts every step. The orange border marks the learnable invertible whitener W, the only trained module, and the cat depicts the AFHQ-Cat generative prior rather than a modifiable natural cover.

secret, Plug-and-Hide (Zhu et al., 2026) and CIF (Qian et al., 2026) make extraction provable or constrained, and HIS (Xu et al., 2025) splits the secrets by importance so that the more important ones survive a distorted channel. All of them hide a single secret image, and because the cover is conditioned on the secret it also tends to retain the secret's appearance; making secret and cover visually dissimilar is a separate design axis (Xu et al., 2026). For several images, GCL-MIH (Chen et al., 2026) compresses multiple secrets into feature vectors, normalizes the vectors to a standard normal distribution and generates a face image with an invertible generative network, reporting 33.18 dB at $n\leq4$. MIDAS (Bae & Lee, 2026) instead encrypts each secret latent with a public-seed orthonormal basis, fuses the encryptions into a single latent and generates the cover with a frozen diffusion model, which yields multi-image hiding with access control.

Two ingredients are shared within this family. The first is the invertible network: RealNVP (Dinh et al., 2017) and Glow (Kingma & Dhariwal, 2018) build exact-likelihood models from invertible coupling layers, the 1×1 convolution of (Kingma & Dhariwal, 2018) is a learned invertible linear map, and the Cayley parameterization gives the same guarantee for orthogonal recurrent networks (Helfrich et al., 2018). The second is the alignment of the latent payload with the prior of the generator: MIDAS (Bae & Lee, 2026) mixes multi-image latents with a public-seed orthonormal basis, whose seed also carries the access control, and classical whitening (Kessy et al., 2018) is the non-learned special case of that operation. LW-MIH belongs to this family: it adds to the public-seed mixing of (Bae & Lee, 2026) a permutation, block-shared orthogonal mixing and a diagonal scale, and *trains* the mixer to match the prior of the generator. Together with the fixed tensor size, that is what lets all 96 images be recovered rather than a few of them collapsing. Table 12 (Appendix E) places these families in a single taxonomy.

3 Proposed Method

As Figure 1 shows, LW-MIH couples a frozen foundation VAE (the secret encoder), an invertible neural network (INN) and a frozen AFHQ-Cat rectified flow (the cover generator), in two phases: the *hiding phase*, which packs the latents and embeds them generatively, and the *recovery phase*, which inverts every learned map and reconstructs the secrets.

3.1 A. Problem Formulation

Let $X_{\text{sec}} = \{x_1, \ldots, x_n\}$ denote n secret images of size $L \times L$, $\mathcal{E}$ the encoder and $\mathcal{D}$ the decoder of the frozen foundation VAE. The packed latent tensor holds a fixed number $B = H\times W\times3 = 256\times256\times3 = 196{,}608$ of elements, and a secret encoded by a VAE with downsampling factor f and latent channel count C occupies $(L/f)^2C$ elements. The fixed size is filled exactly when

$$B \;=\; n\cdot\big(L/f\big)^2\cdot C, \tag{1}$$

Algorithm 1 Hiding phase of LW-MIH

Require: n secret images $X_{\text{sec}} = \{x_1, \dots, x_n\}$, hiding scheme s, frozen VAE encoder $\mathcal{E}$, INN T, whitener W, frozen flow Φ
Ensure: Cover Y, scale σ
1: Form the latent slots $\{f_j\}$ from $\{x_i\}$ and encode them, $f_j = \mathcal{E}(\cdot)$ ▷ FC: $R{=}n$; HFD: $R{=}2n$; UP: Lanczos to L first
2: $Z \leftarrow \Pi(\{f_j\})$, reshaped to $256 \times 256 \times 3$ ▷ packing: Z fills the tensor of Eq. (1)
3: $z \leftarrow T(Z)$ ▷ INN: field at the noise end of the flow
4: $w \leftarrow W(z)$; $\sigma \leftarrow \text{std}(w)$; $\zeta \leftarrow w/\sigma$ ▷ whitening and rescaling
5: $x_{\text{end}} \leftarrow \Phi_{0\to T}(\zeta)$; $Y \leftarrow \psi(x_{\text{end}})$ ▷ frozen flow, then pixelization
6: **return** Y, σ

which holds by construction: once the VAE and the tensor size are fixed, so is the number of hidden images. Writing $R = B/\big((L/f)^2 C\big)$ for the full capacity, the three encoders used here give FLUX ($f{=}8$, $C{=}16$) $R{=}12$, SD1.5 ($f{=}8$, $C{=}4$) $R{=}48$ and Sana ($f{=}32$, $C{=}32$) $R{=}96$. Equation (1) also shows the only two ways in which the number of hidden images can change: encoding the secrets at a different resolution, or letting one secret occupy more than one latent slot.

Two requirements govern the design of the pipeline. First, *latent–prior compatibility*: the cover is produced by integrating a frozen rectified flow from a sample of $\mathcal{N}(0, I)$, so the packed tensor must be mapped to a variable ζ close to such a sample. Second, *faithful recovery*: the INN and the whitener are inverted exactly, while image recovery remains lossy. We write σ for a per-cover scalar transmitted with the cover, recover the secrets as $\hat{X}_{\text{sec}} = \mathcal{R}(Y, \sigma)$, and measure the error of the recovered set by

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{n} \big\| x_i - \hat{x}_i \big\|_1. \tag{2}$$

The whitener is the only module trained once the INN and the flow are fixed, and it is fitted on the packed field rather than on $\mathcal{L}_{\text{rec}}$ of (2) (§3.2). Because the VAE is frozen, $\hat{x}_i$ cannot be closer to x_i than the encoder-decoder pair is to itself, which we call the *VAE ceiling* of the setting and always report next to the PSNR. The VAE, the INN and the rectified flow are frozen once trained and are public.

3.2 B. Hiding Phase: Packing, Invertible Mapping and Whitening

The hiding phase transforms the secret image set into a single generated cover. Its steps are summarized in Algorithm 1.

1) VAE latent packing and the three hiding schemes. Packing Π concatenates the latent slots into a single tensor $Z \in \mathbb{R}^B$ that is reshaped as $256{\times}256{\times}3$. Because the size is fixed, hiding a different number of images means filling it differently: the three schemes below fill it in different ways and thereby trade image count against recovery fidelity.

- **Full capacity (FC).** Each secret is encoded at its native resolution $L{=}256$ and takes one latent slot, so $n{=}R$: 12, 48 or 96 depending on the encoder. FC hides the most images and is the setting used in all ablations.
- **High-frequency dual-slot (HFD).** Half of the latent slots carry a low-pass base, $z_{\text{b}} = \mathcal{E}(G{*}x)$, and half a detail residual, normalized by $\text{norm}(r) = \text{clip}\big(\frac{1}{2} + r/(2s_r), 0, 1\big)$ with $r = x - G{*}x$ and $s_r{=}0.12$, into the input domain of the frozen VAE (G is a Gaussian filter, $\sigma_g{=}1.2$); recovery is $\hat{x} = \text{clip}(\hat{x}_b + (2s_r)(\hat{x}_r - \frac{1}{2}), 0, 1)$. Two slots are spent per secret, so $n{=}R/2$.
- **Upsampling (UP).** Secrets are upscaled to $L \in \{512, 1024\}$ with Lanczos interpolation before encoding, and recovered images are Lanczos-downsampled to 256. Each secret then takes a finer latent that occupies more positions of the tensor, so fewer images are hidden, but each is bounded by a higher ceiling.

The three schemes reuse the same INN, whitener and flow and differ only in how the fixed size is filled and unpacked, so they are interchangeable settings of one framework rather than separate pipelines. Because each scheme admits finitely many resolutions, the admissible numbers of hidden

images form a discrete set rather than a continuous range: FLUX admits 12, 6 or 3; SD1.5 admits 48, 24, 12 or 3; Sana admits 96, 48, 24 or 6. All of these capacity–fidelity configurations are measured in §4.2.

2) Invertible neural network. The INN T turns the packed latents into a field the generator can use. We adopt a PackRealNVP flow (Dinh et al., 2017; Kingma & Dhariwal, 2018) on the $256\times256\times3$ field: three 2×2 squeeze operations followed by 8, 8 and 6 steps of an ActNorm layer, an invertible 1×1 convolution and an affine coupling layer. The network is trained by maximum likelihood,

$$-\log p(Z) = \tfrac{1}{2}\big(\|T(Z)\|_2^2 + D\log 2\pi\big) - \log\big|\det \partial T/\partial Z\big|, \qquad D{=}B, \tag{3}$$

so that the field $z = T(Z)$ has unit-variance coordinates: the first term of (3) is a sum of squares over coordinates and the second is a logarithmic volume. The output of T therefore matches the prior in its marginals but not in its geometry, because the packing operator lays each secret's latent into a contiguous region of the tensor, so positions inside a region stay correlated and the field retains the block structure of the layout. On the untrained pipeline the packing-aware correlation statistic of $T(Z)$ is 0.09–0.13, against about 0.019 for any orthogonal mixing.

3) Learnable invertible whitener. That structure matters because the cover generator is a rectified flow (Liu et al., 2023; Lipman et al., 2023) trained only on samples of $\mathcal{N}(0, I)$, so a field that matches the prior in its marginals but not in its geometry is not a draw from it. For a small fraction of packed fields the endpoint x_{end} then leaves the interval $[-1, 1]$ that pixelization can represent, and since every module is invertible the clipping is carried back through the whole pipeline, damaging a few secret images far beyond the average. LIW removes that deviation before the field reaches the generator: it receives $z = T(Z)$ and produces

$$w = W(z), \qquad \sigma = \mathrm{std}(w), \qquad \zeta = w/\sigma, \tag{4}$$

where

$$W(z) = \mathrm{blk}(Q, s) \circ P(z), \tag{5}$$

P is a fixed public permutation of all D coordinates and $\mathrm{blk}(Q, s)$ acts on the first $n_{\text{mix}} = n_{\text{blk}}K$ coordinates after P, with $n_{\text{blk}} = \lfloor \gamma D/K \rfloor$. On a block $u \in \mathbb{R}^K$ it applies the diagonal scale first and the orthogonal factor second, $\mathrm{blk}(Q, s)u = Q(s \odot u)$; the remaining $D - n_{\text{mix}}$ coordinates (39,424 at $\gamma{=}0.8$, $K{=}256$) are left unchanged. The orthogonal factor is produced by the Cayley transform,

$$Q = (I - A)(I + A)^{-1}, \qquad A = -A^\top, \tag{6}$$

which keeps Q orthogonal for any real skew-symmetric A: $(I - A)^\top = I + A$, so $Q^\top = (I - A)^{-1}(I + A)$, and because $I \pm A$ commute, $Q^\top Q = I$. Orthogonality is therefore a property of the parameterization, not something a training penalty enforces. The eigenvalues of A are $\{0, \pm i\mu_k\}$, so $\det Q = \prod_k (1 - i\mu_k)/(1 + i\mu_k) = 1$; the unit determinant belongs to Q, while the permutation contributes ±1 and the scale is shared by all blocks, so $|\det W| = (\prod_j s_j)^{n_{\text{blk}}}$, independent of the input. As a bijection of $\mathbb{R}^D$, W amplifies a perturbation by at most $1/\min_j s_j$ and introduces no recovery-side gain, unlike the public-seed fusion of (Bae & Lee, 2026), whose convex combination of the packed latent with a reference latent amplifies the field at recovery. Above all, it adds no representable degrees of freedom beyond the fixed size B: it reshapes the geometry of the field but cannot create positions. Because one Q is shared across all blocks, the trainable part of LIW is $K(K{-}1)/2 + K$ free scalars — 32 896 at $K{=}256$ — regardless of D, while P mixes distant lattice sites so that packing-induced structure cannot hide inside a block.

LIW is trained after the INN has been frozen, on the packed fields the network produces:

$$\begin{aligned}\mathcal{L}(\theta) = \lambda_1 \sum_{\ell \in \Lambda} \big|\overline{\rho}_\ell(w)\big| + \lambda_2 \,\mathrm{Var}_\omega\big[\log S(\omega; w)\big] + \lambda_3\big(\overline{m}^2 + (\hat{\sigma}_w - 1)^2\big) \\ + \lambda_4\,\mathrm{NLL}(w) + \lambda_5\,\mathcal{R}_{\text{tail}}(w).\end{aligned} \tag{7}$$

The five terms target the ways in which a packed field deviates from the prior. The first penalizes the magnitude of the spatial autocorrelation at lags $\Lambda{=}\{1, 2, 4\}$ px and the second the spectral colouring, through the variance of the azimuthally averaged log power spectrum $S(\omega; w)$; the third penalizes deviations of the field's mean $\overline{m}$ and its empirical standard deviation $\hat{\sigma}_w$ from 0 and 1; the fourth, a Gaussian negative log-likelihood, keeps the marginals close to $\mathcal{N}(0, 1)$; and the fifth, $\mathcal{R}_{\text{tail}} = \mathbb{E}[\max(0, |w| - \tau)^2]$ with $\tau{=}3$, penalizes the heavy tails whose overflow becomes the permanent damage described above.

Algorithm 2 Recovery phase of LW-MIH

Require: Cover Y, scale σ, hiding scheme s, whitener W, INN T, VAE decoder $\mathcal{D}$, frozen flow Φ
Ensure: Recovered secrets $\hat{X}_{\text{sec}} = \{\hat{x}_1, \ldots, \hat{x}_n\}$
1: $x_{\text{end}} \leftarrow \psi^{-1}(Y)$; $\zeta \leftarrow \Phi_{T\to 0}(x_{\text{end}})$ ▷ undo pixelization, backward ODE
2: $w \leftarrow \sigma\,\zeta$; $z \leftarrow W^{-1}(w)$
3: $Z \leftarrow T^{-1}(z)$
4: $\{f_j\}_{j=1}^{R} \leftarrow \Pi^{-1}(Z)$
5: Decode the R slots with $\mathcal{D}$ and invert scheme s to obtain n images ▷ HFD: recombine; UP: Lanczos to 256
6: **return** $\hat{X}_{\text{sec}}$

3.3 C. Recovery Phase: Inversion

Recovery inverts every step of the hiding phase, as summarized in Algorithm 2: the receiver undoes the pixelization, integrates the frozen flow backwards to obtain ζ, rescales it by the transmitted scalar σ, inverts the whitener, the INN and the packing, and decodes with the frozen VAE decoder $\mathcal{D}$. The maps T and W are analytically bijective, but end-to-end recovery is bounded by the VAE ceiling, the ODE error and the cover quantization.

Since W, T and Π are all invertible, the error of the recovered secrets decomposes as

$$\text{Err} \approx \text{Err}_{\text{VAE}} + \text{Err}_{\text{INN}} + \text{Err}_{\text{ODE}} + \text{Err}_{\text{quant}}, \tag{8}$$

with $\text{Err}_{\text{INN}} \approx 0$, because the round-trip maximum absolute error of the INN and LIW is about 10^{-5}, far below $\text{Err}_{\text{quant}}$ at either bit depth. The remaining error therefore comes from the VAE ceiling and from the bit depth. At 16-bit precision the transmitted cover reproduces the packed field almost exactly, and the recovered secrets sit at the limit of the encoder; at 8-bit precision the quantization step is large enough to interact with the statistics of the field, which is what makes the whitener matter in practice.

3.4 D. Design Consequences of the Whitener

Three consequences follow. First, because the whitener cannot create positions in the tensor, a 16-bit PSNR above the VAE ceiling of its encoder would contradict the frozen encoder, so that ceiling is the reference against which every reported number is read. Second, because the whitener is exactly invertible and well conditioned, its own error is negligible: a difference between two mixers reflects the statistics of the fields they produce, not any loss inside the mixer. Third, its effect is asymmetric in the bit depth: at 16-bit the quantization step is far below the typical spacing of the field, so avoiding overflow suffices, whereas at 8-bit it is comparable to the scale of the field itself, so the amplitude and the tails matter.

4 Experiments

4.1 A. Experimental Settings

Frozen components and data. We freeze the three foundation VAEs — FLUX (Black Forest Labs, 2024), SD1.5 (Rombach et al., 2022) and Sana (Xie et al., 2025; Chen et al., 2025) — the INN and a rectified flow (Liu et al., 2023; Lipman et al., 2023) trained on AFHQ-Cat (Choi et al., 2020); only LIW is trained, separately per (VAE, hiding scheme) pair. Secrets come from CelebA-HQ (Karras et al., 2018), COCO (Lin et al., 2014), DIV2K (Agustsson & Timofte, 2017) and VOC2012/2007 (Everingham et al., 2015), centre-cropped to 256×256, so the content spans faces, objects, natural scenes and mixtures; unless stated otherwise we use CelebA-HQ.

Metrics and protocol. We report the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) (Wang et al., 2004) of the recovered secrets at 16-bit and 8-bit precision, the fraction of ODE endpoints outside $[-1, 1]$ (overflow), and the VAE ceiling of each setting (Table 9). Every comparison fixes the cover prior, tensor size, frozen VAE and INN and varies only the module placed after the INN, on the same eight groups, so contrasts are paired at group level; we report the paired mean difference with a 2×10^4-resample bootstrap confidence interval and a two-sided exact Wilcoxon signed-rank test. A full-capacity setting uses eight groups; the dual-slot and upsampling

Table 1: Every operating point of LW-MIH, on CelebA-HQ secrets with LIW. Each cell gives n together with the secret PSNR in dB (16-bit / 8-bit) for a single cover of fixed size $B = 196\,608$. FC and HFD use $L{=}256$, UP uses $L{=}512$ or 1024, and each value is bounded by its own VAE ceiling (Table 9). Full-capacity cells use eight groups; the HFD and UP cells use four.

Hiding scheme	FLUX	SD1.5	Sana
FC ($L{=}256$, $n{=}R$)	$n{=}12$: 35.03/24.12	$n{=}48$: 28.17/22.77	$n{=}96$: 26.89/20.18
HFD ($n{=}R/2$)	$n{=}6$: 34.97/23.40	$n{=}24$: 29.43/23.42	$n{=}48$: 27.06/20.90
UP ($L{=}512$)	$n{=}3$: 41.03/25.28	$n{=}12$: 32.89/24.68	$n{=}24$: 32.36/24.97
UP ($L{=}1024$)	–	$n{=}3$: 37.08/29.03	$n{=}6$: 37.91/27.33

settings use four, at $L{=}256$ and at $L \in \{512, 1024\}$. A group is called *collapsed* when its 16-bit PSNR falls more than 2 dB below its own VAE ceiling, a convention under which collapsed groups fall below that ceiling by 2.3–12.4 dB while every other group falls short of it by at most 1.8 dB.

Compared variants. Three post-INN modules are compared under one protocol: (a) the INN alone; (b) a random orthogonal mixer of the same form with random signs, the MIDAS-style construction (Bae & Lee, 2026); and (c) LIW. A permutation-only ablation ($Q{=}I$) on all three encoders is in Appendix B.

4.2 B. Trade-off Between the Number of Hidden Images and Per-Image Recovery Quality

Table 1 lists every operating point the framework can realise and Figure 2 shows one secret recovered at each. Two properties stand out.

The number of hidden images is discrete. Within one encoder it cannot be varied continuously: each operating point is a specific way of filling the tensor, and the framework cannot interpolate between neighbouring entries, so a user selects a point rather than tuning a continuous rate.

Quality and quantity trade off along that set. Moving from full capacity to the smallest operating point raises the 16-bit PSNR of each recovered secret from 35.03 to 41.03 dB on FLUX, from 28.17 to 37.08 dB on SD1.5 and from 26.89 to 37.91 dB on Sana, at the cost of hiding fewer images. The 8-bit column follows the same pattern, spanning 23.40–25.28, 22.77–29.03 and 20.18–27.33 dB. The two intermediate schemes, high-frequency dual-slot (HFD) and upsampling (UP), occupy the interior of each range, with one exception: on FLUX the dual-slot point lies below full capacity at both bit depths (0.06 dB at 16-bit, 0.7 dB at 8-bit), because FLUX already carries $C{=}16$ latent channels, so splitting off a detail residual is largely redundant and the extra recombination step only adds error, whereas Sana, whose latent is spatially much smaller, does benefit from it. Two further points illustrate the range: FLUX reaches 34.97 dB at 6 images with dual-slot, and Sana reaches 27.06 dB at 48 images with dual-slot.

The extreme points lie close to their own VAE ceiling at 16-bit — 35.03 against 35.18 at full capacity, and 41.03 against 41.29 for FLUX at $n{=}3$, $L{=}512$ — so when many images are hidden, fidelity is set by the encoder and the bit depth rather than by how many images share the cover. The same holds in SSIM: 16-bit recovery stays within 0.001 of the frozen VAE's own round trip on all three encoders, and except at the Sana dual-slot point, which is 1.6 dB below its ceiling, every operating point stays within 0.3 dB of its ceiling at 16-bit (Table 5). Which of the two levers matters more then depends on the encoder. On FLUX, moving from full capacity to the fewest images costs 10.91 dB at 16-bit, and the 8-bit column moves by only 1.16 dB; on SD1.5 and Sana the same move raises 8-bit recovery by 6.26 and 7.15 dB, against 16-bit gains of 8.91 and 11.02 dB, at a 16-bit cost of essentially nothing (Figure 4).

4.3 C. Ablation: the Learnable Invertible Whitener

Table 2 compares the three modules at full capacity, and Figure 4 places all of the operating points on one set of axes. Without any mixer the pipeline fails on whole evaluation groups: $2/8$ of the FLUX groups and $4/8$ of the Sana groups fall more than 2 dB below their own ceiling, with worst-case deficits of 12.4 and 4.7 dB, because those fields overflow and are clipped. Here and below a deficit is measured against the group's own ceiling; the aggregate ceiling is only a mean over groups.

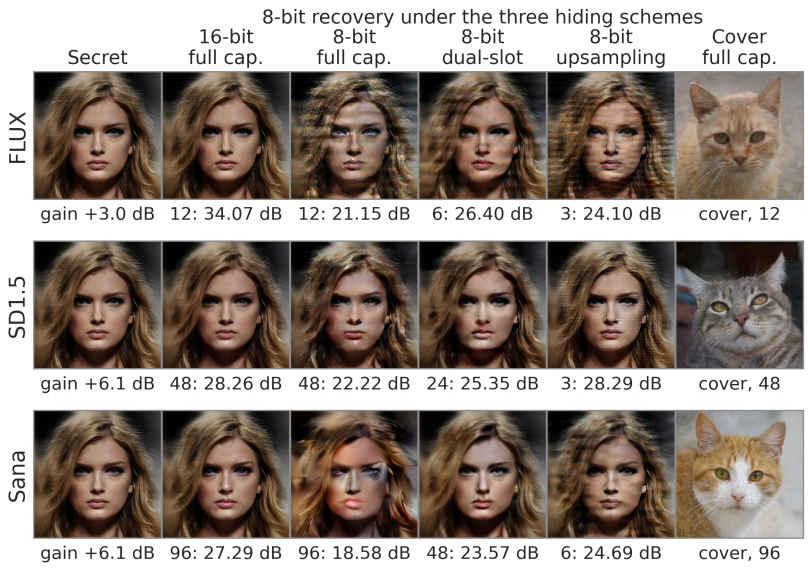

Figure 2: The same secret through the whole framework, for the three frozen encoders. Columns: the secret; its 16-bit recovery at full capacity; and, grouped under "8-bit recovery under the three hiding schemes", its 8-bit recovery at full capacity, with the high-frequency dual-slot scheme and under upsampling at that encoder's fewest images (L=512 for FLUX, L=1024 for SD1.5 and Sana; the remaining upsampling points are in Table 1); and the transmitted cover. A row fixes the encoder, the tensor size and the secret, so the number of hidden images falls from left to right across the 8-bit columns. Each tile prints the number of hidden images and the PSNR of that group; the tile under the secret additionally prints the 8-bit gain to that encoder's fewest images. Table 1 gives the means over groups.

SD1.5 shows no collapsed group, so the failure reflects the encoder geometry rather than the pipeline alone (Figure 3 in Appendix B).

A random orthogonal mixer removes the overflow, but it is not sufficient: one Sana group still collapses with a 2.3 dB deficit and 0.20% mean overflow, and on SD1.5 it makes 8-bit recovery worse than no mixer at all (Table 6). LIW is the only variant with no collapsed group on any encoder, its worst per-group shortfall from that group's own ceiling being 0.30, 0.41 and 0.15 dB for FLUX, SD1.5 and Sana.

Relative to the random mixer, LIW recovers the 8-bit loss wherever random mixing causes it: +3.97 dB on FLUX and +3.36 dB on SD1.5 (p=0.016 in both, with 7 of 8 paired groups positive). On SD1.5 that gain is measured against a mixer that itself hurts 8-bit recovery: the whitened pipeline reaches 22.77 dB against 22.88 dB with no mixer, so the role of LIW there is to keep a single pipeline without that penalty. On Sana the contrast is +1.14 dB and not significant (p=0.148, interval $[-0.04, +2.39]$, 5 of 8 groups positive). At 16-bit precision the mean differences from the random basis are at most 0.49 dB (+0.49, p=0.039 on FLUX; +0.04, p=0.250 on SD1.5; +0.30, p=0.312 on Sana), which matches §3.4: orthogonal mixing restores 16-bit fidelity, while *learning* the mixer restores 8-bit fidelity. In 8-bit SSIM the same ordering appears: at full capacity LIW reaches 0.761, 0.717 and 0.634, against 0.611, 0.627 and 0.577 for coordinate permutation alone (Table 5).

4.4 D. Effect of the Whitener on the Packed Latent Distribution

Table 11 (Appendix D) reports four properties of the packed field at the checkpoints used for all other results, separating what mixing achieves from what learning adds.

What mixing achieves. The untrained field carries a packing-aware correlation (0.13 on FLUX, 0.10 on Sana) and a coloured spectrum (1.667 on FLUX, 1.664 on SD1.5 and 1.661 on Sana). Any orthogonal mixing collapses the first to 0.018–0.020 and the second to 1.643–1.648, and the two mixers are indistinguishable in both: on FLUX they differ by 0.0007 in correlation and 0.0025 in

Table 2: Ablation of the module placed after the INN (8 groups, CelebA-HQ, full capacity). "Degraded" counts groups more than 2 dB below their own VAE ceiling, "Worst" is the minimum per-group 16-bit PSNR, and overflow is the mean fraction of ODE endpoints outside $[-1, 1]$. 16-bit and 8-bit entries are secret PSNR in dB, mean±std over groups.

Encoder	Post-INN mixer	U16	U8	Overflow	Worst U16	Degraded
FLUX (n=12)	INN only	32.36±4.72	21.82±1.33	1.03%	22.83	2/8
	Random orthogonal Q	34.54±0.79	20.15±1.26	0.000%	33.33	0/8
	Learned W	35.03±0.73	24.12±2.40	0.001%	34.07	0/8
SD1.5 (n=48)	INN only	28.21±0.25	22.88±0.95	0.002%	27.93	0/8
	Random orthogonal Q	28.13±0.29	19.40±1.43	0.000%	27.77	0/8
	Learned W	28.17±0.30	22.77±1.79	0.002%	27.76	0/8
Sana (n=96)	INN only	24.46±1.49	17.74±0.61	0.88%	22.75	4/8
	Random orthogonal Q	26.59±0.92	19.04±0.95	0.20%	24.40	1/8
	Learned W	26.89±0.22	20.18±1.34	0.000%	26.60	0/8

spectral variance, less than the epoch-to-epoch spread of either statistic. Across the nine (encoder, mixer) cells neither statistic correlates with 8-bit recovery (r=0.07).

What learning adds. LIW also calibrates the amplitude of the field, bringing its standard deviation to 0.910–0.912 on all three encoders, where the untrained and the random variant stay at 0.983–0.997; the random variant leaves the standard deviation unchanged, as an orthogonal map must. The tail mass behaves the same way: on FLUX, $\Pr(|w| > 3)$ falls from 0.0024 (untrained) and 0.0023 (random) to 0.0011, and the reduction is comparable on SD1.5 and Sana. These two statistics, unlike the correlation and the spectrum, track 8-bit recovery with $r \approx -0.53$ across the same nine cells, consistent with the asymmetry of §3.4: at 16-bit a field need only avoid overflow, whereas at 8-bit its amplitude and its tails determine how much quantization noise the inverse mapping amplifies. The mechanism is therefore amplitude and tail calibration on top of marginal matching, not spectral flattening, and it corresponds to the third and fifth terms of objective (7).

4.5 E. Cross-Encoder Generalization

One framework, one whitener objective and one cover generator serve all three encoders: absolute quality follows the frozen ceilings (35.18, 28.24 and 26.95 dB), no group collapses, and the mean gaps to the 16-bit ceiling under LIW are 0.15, 0.07 and 0.06 dB. Table 7 reports the remaining secret datasets, and Appendix F scores the covers themselves. The two sensitivity alternatives, each evaluated on four groups, differ from each other by only 0.45 dB at 8-bit, but both lie 1.8–2.2 dB above the reported eight-group configuration (Table 10), so we keep the reported configuration.

4.6 F. Comparison with Reported Numbers

Table 8 (Appendix B) places LW-MIH next to published image hiding results. The compared settings differ in cover type, image count, dataset and metric, so the table is context rather than evidence; the controlled comparisons are Tables 1 and 2.

5 Conclusion

This paper presented LW-MIH, a generative framework for coverless multi-image hiding. Frozen foundation VAEs encode the secrets and a separate rectified flow synthesizes the cover, so fixing the size of the packed latent tensor lets one framework serve FLUX, SD1.5 and Sana: the three hiding schemes turn image count and per-image fidelity into discrete operating points, each bounded by the ceiling of its own encoder. The learnable invertible whitener then aligns the packed field with the prior of the generator without adding representable capacity. Mixing alone removes the overflow that collapses whole evaluation groups, and learning the amplitude and the tails turns that into 8-bit fidelity: no group falls more than 0.41 dB below its own ceiling, and FLUX gains 3.97 dB.

AI use statement

Generative AI tools were used (i) to aid and polish writing; (ii) for retrieval and discovery of related work; (iii) for research ideation and execution, including assistance with code that the authors reviewed, tested and ran themselves; and (iv) to draft and revise sections of the paper. We reviewed

all AI-assisted text and code; every number reported here came from the authors' verified code, and the authors take responsibility for the final content.

Reproducibility statement

Hyper-parameters, checkpoints and the training schedule are in Appendix A.

References

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

Minyeol Bae and Si-Hyeon Lee. Training-free coverless multi-image steganography with access control. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.09390.

Shumeet Baluja. Hiding images in plain sight: Deep steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2069–2079, 2017.

Black Forest Labs. FLUX.1 models. Model release and model card, 2024. `https://github.com/black-forest-labs/flux`.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Liang Chen, Xianquan Zhang, Chunqiang Yu, Xinpeng Zhang, Ching-Nung Yang, and Zhenjun Tang. GCL-MIH: A generative-based coverless multi-image hiding method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6720–6734, 2026. doi: 10.1109/TPAMI.2026.3658731.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. doi: 10.1007/s11263-014-0733-5.

Jessica Fridrich. *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, Cambridge, U.K., 2009.

Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):372–390, 2023.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled Cayley transform. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1969–1978, 2018.

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.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. HiNet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4713–4722, 2021. doi: 10.1109/ICCV48922.2021.00469.

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.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018. doi: 10.1080/00031305.2016.1277159.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1×1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L. Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10811–10820, 2021. doi: 10.1109/CVPR46437.2021.01067.

Yuqi Qian, Yun Cao, Meiyang Lv, and Haocheng Fu. CIF: A constrained inversion framework for reliable message extraction in diffusion-based generative steganography. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2026. doi: 10.1145/3785353.3815092. arXiv:2508.00434.

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. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Enze Xie, Junsong Chen, Junyi Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

Hongxin Xu, Jian-Ping Mei, Can Wang, and Defang Chen. Secret-stego dissimilarity as a design axis: Invertible coverless image steganography with diffusion models. *arXiv preprint arXiv:2608.13597*, 2026.

Youmin Xu, Xuanyu Zhang, Xiandong Meng, Chong Mou, and Jian Zhang. Diffusion-based hierarchical image steganography. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209329. arXiv:2405.11523.

Yiwei Yang, Zheyuan Liu, Jun Jia, Zhongpai Gao, Yunhao Li, Wei Sun, Xiaohong Liu, and Guangtao Zhai. DiffStega: Towards universal training-free coverless image steganography with diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1579–1587, 2024. doi: 10.24963/ijcai.2024/175.

Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. CRoSS: Diffusion model makes controllable, robust and secure image steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Table 3: Public-constant σ versus the per-cover scale of the main tables (CelebA-HQ, full capacity, 8 groups). "Recover $\bar{\sigma}$" inverts oracle-generated covers with that encoder's public mean. "Recover FLUX 0.9033" uses the FLUX mean on SD1.5 and Sana. "Recover 1" uses unit scale. "Generate+ recover $\bar{\sigma}$" uses the encoder's public mean on both sides. Every FLUX row uses the same whitener checkpoint as the main tables; only the test-time σ protocol changes. "Per-cover" reproduces the setting of the main tables; the "generate+recover" rows replace the per-cover scale and show the effect of that substitution.

Encoder	Scale at test time	U16	U8
FLUX	Per-cover (main tables)	35.03±0.73	24.12±2.40
	Recover $\bar{\sigma}$=0.9033	35.01±0.69	24.11±2.24
	Recover σ=1	30.38±0.46	23.39±1.91
	Generate+recover 0.9033	31.86±4.99	23.19±5.80
SD1.5	Per-cover (main tables)	28.17±0.30	22.77±1.79
	Recover $\bar{\sigma}$=0.8531	28.16±0.28	22.77±1.68
	Recover FLUX 0.9033	27.02±0.25	22.01±1.61
	Recover σ=1	22.89±0.17	19.75±1.18
	Generate+recover 0.8531	28.00±0.50	22.79±2.79
Sana	Per-cover (main tables)	26.89±0.22	20.18±1.34
	Recover $\bar{\sigma}$=0.9692	26.87±0.22	20.18±1.25
	Recover FLUX 0.9033	26.22±0.16	20.28±1.19
	Recover σ=1	26.59±0.30	20.00±1.24
	Generate+recover 0.9692	23.96±5.74	19.91±5.73

Jiahao Zhu, Zixuan Chen, Jiali Liu, Weiqi Luo, Yi Zhou, and Xiaohua Xie. Plug-and-hide: Provable and adjustable diffusion generative steganography. *arXiv preprint arXiv:2409.04878*, 2026. Preprint; first version 2024.

A Implementation details

LW-MIH is implemented in Python 3.12 and trained on one NVIDIA RTX 3090. The rectified-flow integrator uses RK45 with relative and absolute tolerance 10^{-5}. LIW uses permutation seed 20260723 and mixing fraction γ=0.8; the per-encoder checkpoints carry block length K=256, with alternative configurations at K=128 and (γ=0.5, K=256) (Table 10). LIW is trained for 12 epochs with 256 packed fields per epoch, batch 4, AdamW at learning rate 2×10^{-3} with cosine decay and gradient-norm clipping at 5. The reported checkpoints are the validation-best checkpoint of that schedule for each encoder: epoch 6 of 12 for FLUX and Sana and epoch 11 of 12 for SD1.5. All field statistics reported here, in Table 11 and in Figure 5, are measured on those same checkpoints, each on the same 32 validation packs that the training schedule uses for validation. The loss weights of (7) are λ_1=1.0 (autocorrelation), λ_2=0.05 (spectral), λ_3=1.0 (moments), λ_4=0.2 (Gaussian NLL) and λ_5=0.5 ($\mathcal{R}_{\text{tail}}$, τ=3). The random-mixer variant uses QR factorization with sign fixing from the same public seed. Round-trip maximum absolute error is about 10^{-5}. After permutation, only the first n_{mix} coordinates are mixed; the rest follow the identity. The scalar σ of (4) accompanies each cover as side information: it is a single floating-point number, four bytes, or 0.002% of the 196 608 bytes of an 8-bit cover. It cannot be recomputed from the cover, because the division by σ removes it from the field, so the numbers in the main tables assume that it is delivered to the receiver with Y, as in $\mathcal{R}(Y, \sigma)$. The per-cover mean is encoder-specific: 0.9033 ± 0.0044 (FLUX), 0.8531 ± 0.0015 (SD1.5), 0.9692 ± 0.0094 (Sana). Table 3 substitutes a public constant at recovery or at both generation and recovery. Recovering with that encoder's mean changes the 8-bit mean by less than 0.02 dB on every encoder. Replacing the per-cover scale is not equivalent. At recovery, σ=1 gives 30.38, 22.89 and 26.59 dB against 35.03, 28.17 and 26.89 dB for the per-cover scale, so it is worse on every encoder and much worse on FLUX and SD1.5; replacing the scale at generation as well gives 31.86, 28.00 and 23.96 dB, which is now far worse on Sana.

B Additional results

This appendix collects the paired contrasts of §4.3, the per-group collapse plot, the remaining secret datasets, and a permutation-only ablation. Table 4 reports that ablation: a public permutation with Q=I, evaluated on all three encoders under the same protocol as Table 2. On FLUX it does not

Table 4: Permutation-only mixer ($Q{=}I$) versus the three mixers of Table 2, CelebA-HQ, full capacity, 8 groups.

Encoder	Post-INN mixer	U16	U8	Overflow	Worst U16	Degraded
FLUX ($n{=}12$)	INN only	32.36±4.72	21.82±1.33	1.03%	22.83	2/8
	Permutation only ($Q{=}I$)	33.11±3.32	20.26±1.75	0.62%	24.48	2/8
	Random orthogonal Q	34.54±0.79	20.15±1.26	0.000%	33.33	0/8
	Learned W	35.03±0.73	24.12±2.40	0.001%	34.07	0/8
SD1.5 ($n{=}48$)	INN only	28.21±0.25	22.88±0.95	0.002%	27.93	0/8
	Permutation only ($Q{=}I$)	26.82±2.69	20.06±1.71	1.92%	20.07	2/8
	Random orthogonal Q	28.13±0.29	19.40±1.43	0.000%	27.77	0/8
	Learned W	28.17±0.30	22.77±1.79	0.002%	27.76	0/8
Sana ($n{=}96$)	INN only	24.46±1.49	17.74±0.61	0.88%	22.75	4/8
	Permutation only ($Q{=}I$)	26.83±0.22	18.51±1.10	0.000%	26.57	0/8
	Random orthogonal Q	26.59±0.92	19.04±0.95	0.20%	24.40	1/8
	Learned W	26.89±0.22	20.18±1.34	0.000%	26.60	0/8

Table 5: SSIM of the recovered secrets at full capacity (8 groups, CelebA-HQ) for the learned whitener and for the permutation-only mixer. The first column is the round trip of the frozen VAE, which bounds both. The whitener is closer to that bound at both bit depths and on every encoder, whereas coordinate permutation alone does not reach it; the two configurations of each encoder are evaluated on the same groups.

Encoder	Post-INN mixer	VAE rec.	16-bit	8-bit
FLUX ($n{=}12$)	Learned W	0.952	0.951	0.761
	Permutation only ($Q{=}I$)	0.952	0.927	0.611
SD1.5 ($n{=}48$)	Learned W	0.831	0.830	0.717
	Permutation only ($Q{=}I$)	0.831	0.801	0.627
Sana ($n{=}96$)	Learned W	0.786	0.785	0.634
	Permutation only ($Q{=}I$)	0.786	0.784	0.577

Table 6: Paired per-group contrasts over the same groups (8 pairs per cell), as discussed in §4.3. Δ is the paired mean difference in dB, the confidence interval uses a $2{\times}10^4$-resample paired bootstrap of the mean, and p is the two-sided exact Wilcoxon signed-rank test by full enumeration.

Encoder	Contrast	U16 PSNR Δ [95% CI], p	U8 PSNR Δ [95% CI], p
FLUX	Learned W − random Q	$+0.49$ [$+0.19$, $+0.80$], 0.039	$+3.97$ [$+1.92$, $+6.27$], 0.016
FLUX	random Q − INN	$+2.18$ [-0.58, $+5.68$], 0.844	-1.67 [-2.86, -0.56], 0.055
SD1.5	Learned W − random Q	$+0.04$ [-0.09, $+0.13$], 0.250	$+3.36$ [$+1.82$, $+4.85$], 0.016
SD1.5	random Q − INN	-0.08 [-0.14, -0.04], 0.016	-3.48 [-4.66, -2.36], 0.008
Sana	Learned W − random Q	$+0.30$ [-0.04, $+0.88$], 0.312	$+1.14$ [-0.04, $+2.39$], 0.148
Sana	random Q − INN	$+2.13$ [$+1.19$, $+3.13$], 0.008	$+1.30$ [$+0.48$, $+2.13$], 0.039

remove the collapsed groups ($2/8$ still fall more than 2 dB below their own ceiling). On SD1.5 it *creates* two collapsed groups that the untrained INN did not have. On Sana the 16-bit mean already matches that of LIW, but the 8-bit mean remains 1.67 dB lower and the 8-bit SSIM is 0.577 against 0.634. Block-orthogonal mixing is therefore required, and a permutation alone is not sufficient.

Table 8 quotes published image hiding results against the number of hidden images n; as its caption states, each row is capped by the quality ceiling of the system that reports it, so the table is context for §4.6 rather than evidence.

C HIDING SCHEMES: REMAINING TABLES

Table 9 lists the ceiling of the frozen VAE at every operating point of Table 1. Table 10 and Figure 4 complete the capacity–fidelity view.

Table 7: Full-capacity packing with LIW on the remaining secret datasets (8 groups; mean±std over groups).

Dataset	Enc.	n	VAE	16-bit	8-bit
DIV2K	FLUX	12	26.78	26.58±0.78	18.71±1.69
	SD1.5	48	20.70	20.70±0.48	19.12±0.65
	Sana	96	19.48	19.46±0.43	16.48±0.43
VOC2012	FLUX	12	27.97	27.80±1.18	19.20±1.41
	SD1.5	48	22.36	22.35±0.52	19.56±1.25
	Sana	96	21.16	21.13±0.44	17.45±0.78
VOC2007	FLUX	12	28.14	28.08±0.77	20.99±2.04
	SD1.5	48	22.05	22.02±0.31	19.12±1.46
	Sana	96	21.32	21.28±0.40	18.02±1.40
COCO	FLUX	12	34.24	33.79±1.58	20.70±1.44
	SD1.5	48	23.26	23.23±0.70	20.03±1.46
	Sana	96	21.65	21.63±0.26	18.55±1.19

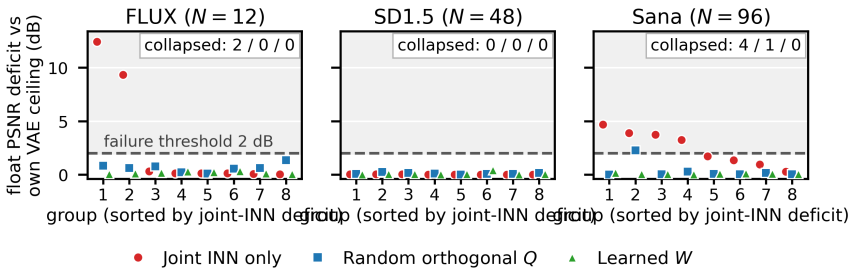

Figure 3: Recovery deficit of each group against its own VAE ceiling, groups sorted by the deficit of the untrained pipeline, as discussed in §4.3. The dashed line marks the 2 dB failure threshold and the shaded band marks deficits above it; the y-axis is the deficit of the 16-bit PSNR. Without any mixer the pipeline loses whole groups on FLUX and Sana instead of degrading uniformly; a random orthogonal mixer removes the FLUX tail but leaves one Sana group; LIW leaves every group's deficit below 0.5 dB, its worst being 0.41 dB on SD1.5. Insets give the collapsed counts for three of the four arms of Table 4, the permutation-only arm being omitted here.

Table 8: Reported secret PSNR (dB) against the number of hidden images n. The numbers are quoted rather than reproduced, the setting behind each of them differs in cover type, host content and dataset, and every PSNR is capped by the quality ceiling of the system that reports it, so each row should be read on its own; the source column names the table or abstract it is quoted from, and each quoted PSNR is the mean over the recovered secrets under the correct key. "Natural" covers are natural photographs; "prompt" covers are generated from a text prompt derived from the secret; "RF" is a rectified flow.

Method	Cover	n	PSNR	Source
HiNet	natural	1	46.15	Tab. 1 of (Yang et al., 2024)
CRoSS	prompt	1	21.25	Tab. 1 of (Yang et al., 2024)
DiffStega	prompt	1	23.29	Tab. 1 of (Yang et al., 2024)
MIDAS	prompt	2	23.90	Tab. 2 of (Bae & Lee, 2026)
GCL-MIH	Flow-GAN face	4	33.18	abstract of (Chen et al., 2026)
LW-MIH (FLUX)	AFHQ-Cat RF	3–12	41.03–35.03	this work
LW-MIH (SD1.5)	AFHQ-Cat RF	3–48	37.08–28.17	this work
LW-MIH (Sana)	AFHQ-Cat RF	6–96	37.91–26.89	this work

D EXTENDED FIELD DIAGNOSTICS

Figure 5 documents the effect of LIW on a single FLUX packed field, and Table 11 reports the corresponding statistics for all three encoders.

Table 9: VAE ceiling of each operating point (16-bit PSNR in dB), for reference against the entries of Table 1.

Setting	FLUX	SD1.5	Sana
FC (L=256)	35.18	28.24	26.95
HFD (n=$R/2$)	35.15	29.51	28.63
UP (L=512)	41.29	32.91	32.40
UP (L=1024)	–	37.14	38.07

Table 10: Sensitivity of LIW to its block length and mixing fraction (CelebA-HQ, FLUX). The alternatives were evaluated at four groups.

Configuration	Groups	U16	U8	Overflow
K=128, γ=0.8	4	34.88±0.47	26.36±1.05	0.0000%
K=256, γ=0.5	4	34.85±0.51	25.91±2.17	0.0001%
K=256, γ=0.8 (reported)	8	35.03±0.73	24.12±2.40	0.0008%

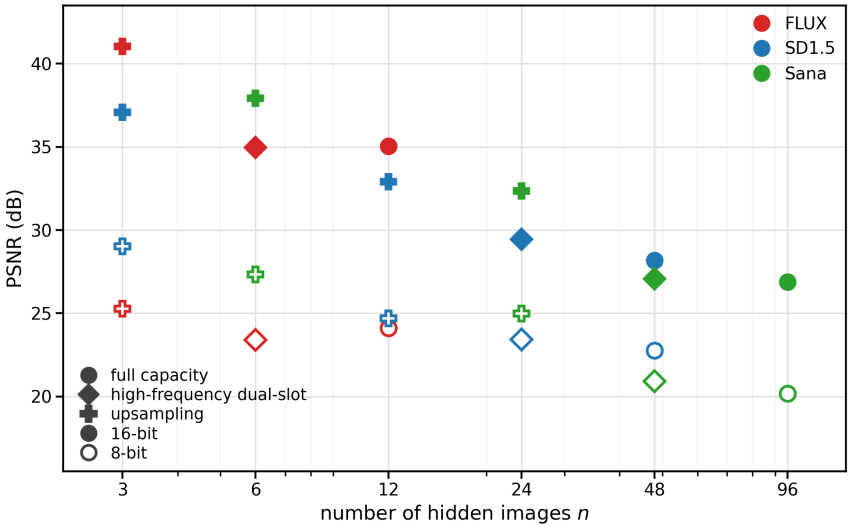

Figure 4: Capacity–fidelity working points on CelebA-HQ with LIW. Colour encodes the frozen encoder (red: FLUX; blue: SD1.5; green: Sana) and marker shape the hiding scheme (circle: full capacity; plus sign: upsampling; diamond: high-frequency dual-slot); the horizontal axis is the number of hidden images n on a logarithmic scale and the vertical axis is PSNR in dB. Filled markers are 16-bit and open markers 8-bit, and all points share the fixed tensor size. The points are the cells of Table 1 (full capacity: eight groups; HFD and UP: four).

E Taxonomy of related work

F Cover realism under the synthetic-cat protocol

The public channel carries an AFHQ-Cat rectified-flow sample, not a natural photograph. Table 13 therefore compares, for each post-INN map, the 24 8-bit covers it produces at full capacity (three encoders × eight CelebA-HQ groups) against the same reused AFHQ-Cat samples. The row marked *joint training* is a reference point rather than one of the maps: it is an earlier model in which the INN and the mixer were trained in a single stage rather than in two, and it is scored on a different, smaller cover set, so its FID is not a controlled contrast. FID (Heusel et al., 2017) is computed with `pytorch-fid`. A 64×64 patch CNN (15 epochs, image-level 70/30 split, 40 patches per image) is trained to label cover versus stego; $|\mathrm{AUC} - 0.5|$ is the detection strength of that classifier, not a claim of undetectability on natural hosts.

INN-only covers are the farthest from the generator's typical set (FID 82.69, AUC 0.831), consistent with the overflow-prone groups of Table 2. LIW returns the covers to the same FID band as the jointly trained checkpoint (59.89 versus 60.41). Among the 24-cover rows the random mixer records the lowest FID (53.07); the mixed-dataset row at 48 covers reaches 40.85, but it is scored against a

Table 11: Measured properties of the packed field, computed on the 32 validation packs of the reported checkpoint of each encoder (epoch 6 of 12 for FLUX and Sana, epoch 11 of 12 for SD1.5), together with the 8-bit recovery of the same three configurations on the same groups, as discussed in §4.4. Each cell is *INN / random orthogonal mixer / LIW*. *corr.* is the packing-aware multi-lag spatial-correlation statistic of the whitener's training objective, i.e. the quantity minimized by that objective; "Spec. var." is the variance of the log azimuthally-averaged power and $\Pr(|w|>3)$ the tail mass. All four statistics are measured before the division by σ.

Encoder	corr.	Spec. var.	std	$\Pr(\|w\|>3)$	U8 (dB)
FLUX	0.1324/0.0185/0.0178	1.6667/1.6462/1.6437	0.9833/0.9833/0.9107	0.0024/0.0023/0.0011	21.82/20.15/24.12
SD1.5	0.1124/0.0196/0.0192	1.6641/1.6471/1.6425	0.9975/0.9975/0.9124	0.0027/0.0026/0.0012	22.88/19.40/22.77
Sana	0.0955/0.0179/0.0180	1.6606/1.6478/1.6435	0.9952/0.9952/0.9100	0.0027/0.0026/0.0011	17.74/19.04/20.18

Table 12: Taxonomy of related work. "Cross-VAE" means evaluation across multiple frozen foundation VAEs.

Method	Coverless	Multi-image	Learned mixing	Cross-VAE
HiNet	no	no	no	no
ISN / DeepMIH	no	yes	no	no
CRoSS / DiffStega	yes	no	no	no
CIF / Plug-and-Hide	yes	no	no	no
GCL-MIH	yes	yes ($n\leq 4$)	no	no
MIDAS random basis	yes	yes	no	no
LW-MIH (ours)	yes	yes	yes	FLUX/SD1.5/Sana

Table 13: Cover realism under one protocol: each post-INN map is scored on the 8-bit covers it produces, at full capacity, against the reused AFHQ-Cat rectified-flow samples (24 covers unless noted, so three encoders $\times$ eight CelebA-HQ groups). The number of covers is given per row; the *joint training* row comes from an earlier, single-stage model and is a reference point, not a controlled contrast. FID is deterministic for a fixed image set, whereas the Patch-CNN AUC varies across retrainings by a few hundredths.

Post-INN map	#covers	FID	Patch-CNN AUC	$\lvert\mathrm{AUC}-0.5\rvert$
Joint training	24	60.41	0.550	0.050
INN only	24	82.69	0.831	0.331
Random orthogonal Q	24	53.07	0.413	0.087
Learned W	24	59.89	0.372	0.128
Learned W, mixed datasets	48	40.85	0.613	0.113

different cover set, so the two are not comparable. None of the mixers yields $|\mathrm{AUC}-0.5|\approx 0$ on this split; a closer FID to synthetic cats is not the same as a natural-host cover.

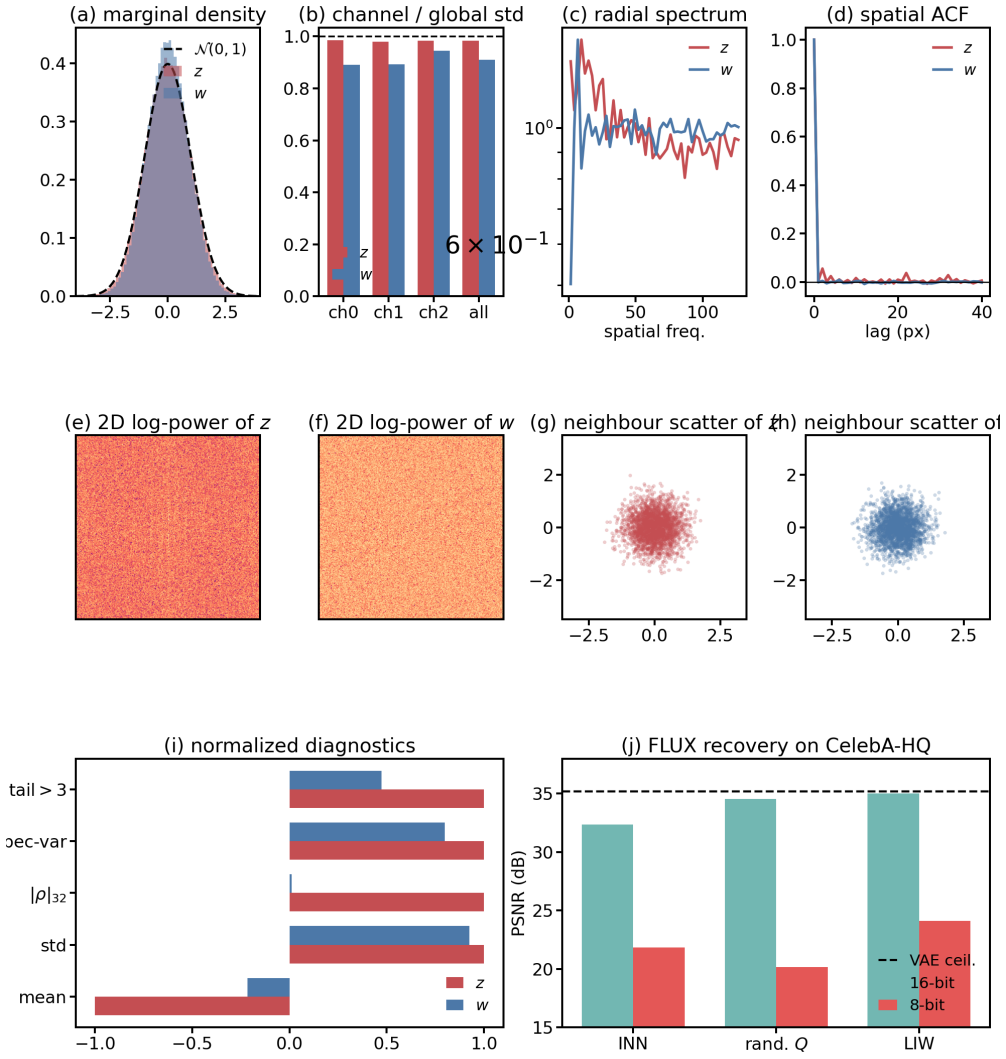

Figure 5: Multi-dimensional effect of LIW on a FLUX packed field. (a) marginal density and (b) per-channel standard deviation; (c) radial spectrum, (d) spatial autocorrelation; (e–f) 2D log-power of z and of w; (g–h) neighbour scatter of z and of w; (i) normalized diagnostics; (j) FLUX recovery on CelebA-HQ. Red denotes $z{=}T(Z)$ and blue $w{=}W(z)$; the dashed curve in (a) is the target prior $\mathcal{N}(0,1)$.