# mAVE: A WATERMARK FOR JOINT AUDIO-VISUAL GENERATION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Watermarking joint audio-visual generation supports vendor copyright protection and content provenance. However, independently valid audio and video watermarks do not establish a shared generation session. An adversary can splice watermarked modalities from different sessions, causing the pair to be mistaken for the vendor's original joint output. We introduce mAVE (Manifold Audio-Visual Entanglement), a training-free watermarking framework that strengthens vendor attribution through session binding in native joint audio-visual diffusion transformers. mAVE separates public record retrieval from secret session authentication: a fixed public index locates the server record, while a randomized payload binds audio bits to a session-keyed video grid through a cryptographic digest. One prompt-conditioned joint inversion supports provider-assisted verification of both modalities against a session record, without modifying generator weights or training auxiliary watermark networks. Our analysis establishes implementation-matched distribution preservation and a full-initialization routing/clipping budget, alongside adaptive session-pool security and stable local-perturbation bounds. Experiments on LTX-2 and MOVA show comparable generation quality. mAVE achieves 99.8% true-positive rate and 0% observed false-positive rate in the evaluated swap test, and retains 99.2% true-positive rate under FrameAvg temporal averaging. Same-prompt and similarity-selected swaps further test session authentication beyond perceptual compatibility.

## 1 INTRODUCTION

Generative AI is making realistic audiovisual content easier to produce and distribute. Commercial video-generation platforms (OpenAI, 2025; Google DeepMind, 2025; Kuaishou Technology, 2026; ByteDance Seed, 2026) and openly available models (HaCohen et al., 2026; SII-OpenMOSS Team et al., 2026) expand creative applications while increasing opportunities to repurpose and misattribute synthetic media. Watermarking provides a proactive signal for synthetic-content identification and source attribution, supporting vendor copyright protection beyond judgments based on visual or auditory realism (Fernandez et al., 2024; San Roman et al., 2024). As generation moves from individual modalities to synchronized audio-video outputs, that protection must also account for how the modalities were produced together.

Native joint generators denoise audio and video within one inference request, producing a paired output with a shared generation history. Yet valid video and audio watermarks need not certify that history. An adversary can combine watermarked video from one session with watermarked audio from another, causing the pair to be mistaken for the vendor's original joint output. The individual marks survive while the attributed pairing is false (Figure 1).

Let $Video_{wm}$ and $Audio_{wm}$ denote valid modality-level detections. Neither accepting either mark ($Video_{wm} \vee Audio_{wm}$) nor requiring both ($Video_{wm} \wedge Audio_{wm}$) resolves this gap. Perceptually plausible swaps can also pass synchronization checks (Chung & Zisserman, 2016). Reliable attribution therefore requires protection at the level of the *generation-session pair*, beyond two separately valid modalities.

We introduce **mAVE** (Manifold Audio-Visual Entanglement), which places this session constraint in a joint generator's initial latents. A video-grid digest determines audio binding bits; inverse-transform sampling embeds the randomized payload, and one prompt-conditioned joint inversion



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

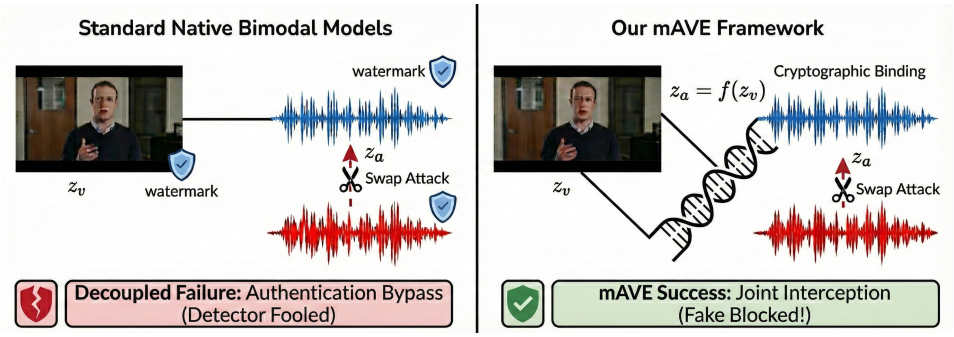


Figure 1: **Session binding with mAVE.** Left: independently valid watermarks can survive cross-session AV splicing. Right: mAVE binds the initial latent pair, enabling the verifier to test shared session provenance.

recovers both modalities. A public index locates the server record for verification. mAVE requires neither generator fine-tuning nor an auxiliary watermark network, enabling integration with existing joint generators through initialization and verification.

On LTX-2 and MOVA, experiments show comparable generation quality. In the evaluated swap test, mAVE achieves 99.8% true-positive rate (TPR) and 0% observed false-positive rate (FPR). mAVE also performs well under temporal attacks: FrameAvg retains 99.2% TPR with 0% observed FPR. Authentic-pair TPR remains 99.0% after MP4 compression and 98.1% after diffusion-based regeneration with audio reconstruction.

Our contributions are threefold:

- **Threat formulation:** We formalize cross-session AV swaps and distinguish modality-level watermark detection from same-session pair authentication.
- **Method and theory:** We introduce training-free session binding without changing generator weights, combining public record retrieval and one joint inversion with implementation-aware sampling guarantees, an explicit full-prior distortion budget, and adaptive session-pool security bounds.
- **Evaluation:** Quality remains comparable on LTX-2 and MOVA. The evaluated swap test yields 99.8% TPR and 0% observed FPR; FrameAvg, MP4 compression, and regeneration retain 99.2%, 99.0%, and 98.1% TPR, respectively. Matched server-reference controls assess the binding construction.

## 2 Related Work

**Joint AV generation.** LTX-2 (HaCohen et al., 2026) and MOVA (SII-OpenMOSS Team et al., 2026) jointly denoise audio and video, exchanging information during generation. Their coupled initialization offers a common point for embedding a session constraint; flow-based sampling (Liu et al., 2022) supports numerical inversion for extraction. Extended background and solver comparisons appear in Sections B and I.

**Unimodal watermarking.** Post-processing methods embed signals in decoded video or audio (Luo et al., 2025; Fernandez et al., 2024; Chen et al., 2023; San Roman et al., 2024). In-processing methods such as Gaussian Shading (Yang et al., 2024), VideoShield (Hu et al., 2025a), and VideoMark (Hu et al., 2025b) instead use generative latents. SIGMark (Zhu et al., 2026) addresses scalable blind extraction; SKeDA (Yang et al., 2026) improves temporal recovery through key permutation and differential attention. MarkDiffusion (Pan et al., 2025) benchmarks generative watermarks. mAVE extends modality-level copyright attribution to session-level AV authentication.

**Trainable video watermarking.** LVMark (Jang et al., 2024) learns latent-decoder modulation and watermark recovery; VIDSIG (Huang et al., 2025) partially fine-tunes the latent decoder with temporal alignment. FlowMark (Asnani et al., 2026) learns mask-guided, content-adaptive embedding. These methods optimize video attribution and integrity. mAVE introduces a complementary

generation-session relation between audio and video through the unchanged joint generator, without watermark-specific training.

**Content-derived signals.** Audio fingerprints (Wang, 2003) and video hashes (Facebook, 2019) match reference media. SyncNet (Chung & Zisserman, 2016) measures lip–speech alignment; ImageBind/CLAP (Girdhar et al., 2023; Wu et al., 2023) measure compatibility. AVFF (Oorloff et al., 2024) learns AV correspondences for deepfake detection. mAVE adds a keyed generation-session decision to these retrieval and forensic signals.

**Cross-modal media watermarking.** $V^2$A-Mark (Zhang et al., 2024) embeds localization/copyright signals into video frames and audio with trained watermarking modules, then combines the recovered copyright bits. Kim et al. (2025) use trainable invertible networks to hide audio in visual frames for audio recovery and tamper localization. Their media-domain copyright and localization functions complement mAVE's native session constraint in a joint diffusion transformer (DiT). Appendix I.5 compares their training and verification objectives.

**Cryptographic protocols.** Thietke et al. (2025) formalize sampling and key-management requirements. Gaussian Shading++ (Yang et al., 2025) develops dual-channel seed recovery and signature-based third-party verification for image watermarks. mAVE addresses a complementary objective: provider-assisted authentication of a joint AV generation session (Section I).

**Provenance infrastructure.** C2PA Content Credentials (C2PA, 2026) combine signed manifests with asset bindings, including watermark/fingerprint soft bindings. mAVE's embedded session signal can connect AV authentication to such provenance records; Section I.4 outlines the integration.

## 3 Threat Model and Problem Formulation

**Provider-assisted verification.** The provider controls latent initialization and retains session records. Each inference request has public index $I$ and secret $K_{priv} = m$. External callers submit verification requests through the provider's interface without receiving these secrets. The decision authenticates shared generation-session provenance; Section 4.3 specifies the prompt and record-retrieval workflow.

The adversary has generator weights and generation/detection API access, but no $K_{priv}$ (Tolosana et al., 2020; Mirsky & Lee, 2021). It can request outputs and adaptively choose or splice them. Detection exposes only accept/reject, never the secret score or gradient. We consider: **(1) Removal:** erasing a modality's watermark through compression, blur, or noise. **(2) Swap:** pairing video $\mathbf{x}_v^{(A)}$ and audio $\mathbf{x}_a^{(B)}$ from distinct sessions $A \neq B$, forming $\tilde{\mathbf{x}} = (\mathbf{x}_v^{(A)}, \mathbf{x}_a^{(B)})$. Individual watermark validity and perceptual compatibility do not establish shared provenance. Section 4.4 analyzes adaptive selection and local perturbations within the allowed swap family.

**Authentication objective.** Under $H_0$, both modalities originate from the indexed session; under $H_1$, they come from independent sessions. We seek low false acceptance under $H_1$ and high recall for distorted $H_0$ samples, accounting for inversion errors.

## 4 Method

mAVE constructs bound grids, embeds them in joint latents, and verifies them after inversion (Sections 4.1 to 4.3); Section K collects notation.

### 4.1 Constructing the Authentic Manifold

Let $\mathcal{Z}_v, \mathcal{Z}_a$ be the video/audio initial latent spaces, $\mathcal{Z} = \mathcal{Z}_v \times \mathcal{Z}_a$, and $(\mathbf{z}_v, \mathbf{z}_a)$ a latent pair. The session key $K_{sess}$ defines an Authentic Manifold $\mathcal{M}$ through a binding predicate Verify (Section 4.3):

$$\mathcal{M} = \{(\mathbf{z}_v, \mathbf{z}_a) \in \mathcal{Z} \mid \text{Verify}(\mathbf{z}_a, \text{SHA-256}(\mathbf{z}_v, K_{sess})) = \text{True}\}. \tag{1}$$

Here the latent-level notation summarizes the bit-grid construction below; hashes operate on discrete grids. The $N = |\mathcal{I}_{bind}|$ binding slots form a sparse constraint relative to $\dim(\mathcal{Z})$.

**Session-Aware Key Derivation.** Let $L_I, L_m$ be the bit lengths of public index $I \in \{0,1\}^{L_I}$ and secret $m \in \{0,1\}^{L_m}$. Only $I$ is embedded; $m = K_{priv}$ remains on the server. For original prompt

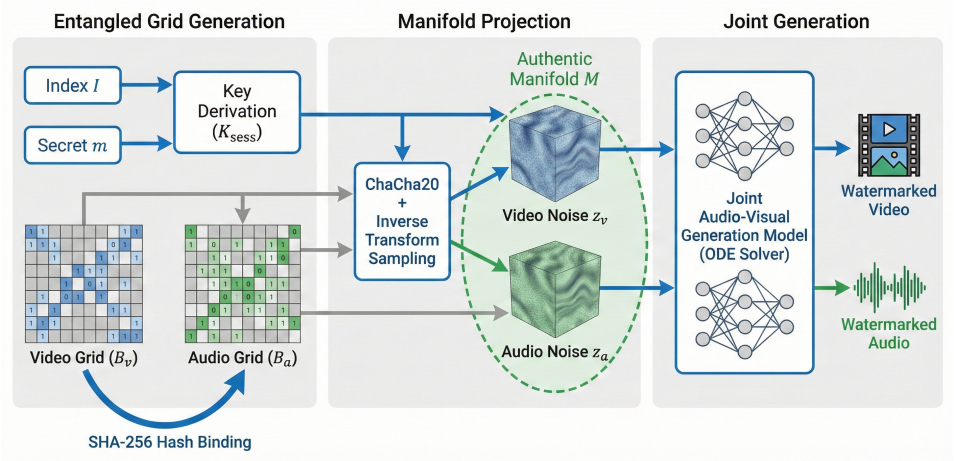


Figure 2: **mAVE pipeline.** A digest binds the audio grid to the video grid. Repetition, keyed randomization, and inverse-CDF sampling map these grids to joint initial latents, which the unchanged AV model denoises.

$P$, let $E_P = \mathrm{Norm}(P)$ denote lowercase conversion and whitespace trimming. With $\|$ denoting concatenation and $\mathrm{Prefix}_{\lambda_{key}}$ retaining a fixed $\lambda_{key}$ hash bits independently of binding length $N$, derive

$$K_{sess} = \text{SHA-256}(\mathrm{Prefix}_{\lambda_{key}}(\text{SHA-256}(m)) \,\|\, \text{SHA-256}(E_P)). \tag{2}$$

Purpose-separated subkeys $K_{grid}, K_{mask}, K_{perm}$ serve grid construction, masking, and block permutation (Section E). HMAC (hash-based message authentication code) derives modality sub-keys as $\mathrm{HMAC}(K_{grid}, \text{modality})$; shared base entropy is seeded from $m$ (Section E).

**Entangled Bit Grid Generation** We construct one-dimensional watermark grids $B_v, B_a \in \{0,1\}^L$ of length $L$, subsequently reshaped and repeated into the audio/video latent tensors, incorporating *Time-Templates* and *Hash-Binding* (see Figure 2, Left).

**1. Video Grid ($B_v$).** The grid has three disjoint regions: (i) a **Time-Template** $\mathbf{t}$ (channel $c_0$) for synchronization, (ii) the index $I$ in fixed, unencrypted repetition slots, and (iii) HMAC-derived base bits. Here $\mathcal{I}_{inject}$ inserts the template and index into designated slots, $S_{shared}$ is the shared pseudorandom bit stream, and $\oplus$ is bitwise XOR:

$$B_v = \mathcal{I}_{inject}(\text{HMAC}(K_{sess}, \text{“video”}) \oplus S_{shared},\ \mathbf{t},\ I)\,. \tag{3}$$

**2. Audio Grid with Binding ($B_a$).** To enforce dependency, we embed a digest of the video bits, $h_v = \text{SHA-256}(B_v)$, at fixed logical binding slots $\mathcal{I}_{bind}$ (128 distinct digest bits by default). Let $i$ index a grid bit, $\mathcal{I}_{time}$ contain template slots, $\phi$ select distinct digest bits, and $\mathrm{Base}_a$ denote the audio sub-key-derived base grid:

$$B_a[i] = \begin{cases} \mathbf{t}[i] & \text{if } i \in \mathcal{I}_{time} \quad \text{(Sync)}, \\ h_v[\phi(i)] & \text{if } i \in \mathcal{I}_{bind} \quad \text{(Binding)}, \\ \mathrm{Base}_a[i] & \text{otherwise.} \end{cases} \tag{4}$$

The resulting $\mathcal{M}$ is a union of hash-selected latent orthants, rather than a smooth lower-dimensional surface. We denote the randomized inverse-CDF projection from grids to Gaussian latents by $\Pi$.

### 4.2 Embed: Inverse Transform Sampling on Manifolds

A critical challenge in latent watermarking is mapping a low-dimensional payload to a high-dimensional latent tensor $\mathbf{z} \in \mathbb{R}^{C\times T\times H\times W}$ with channel, temporal, height, and width sizes $C, T, H, W$. We use three stages: *Diffusion*, *Randomization*, and *Inverse Sampling* (Figure 2, Middle).

**Watermark Diffusion and Randomization** Let $B \in \{0,1\}^L$ be the entangled bit grid constructed in Section 4.1. We first apply **Watermark Diffusion**. The discrete watermark bits are replicated block-wise to form a diffused tensor $B^{diff}$ matching the target latent dimensions.

**Watermark Randomization** masks repetition patterns in the authentication payload while keeping the routing index recoverable. Using the masking subkey $K_{mask}$, we generate a pseudo-random keystream $\mathcal{K}$ via the ChaCha20 (Bernstein, 2008) stream cipher. Let $\mathcal{J}_I$ be the fixed public index-copy coordinates and $\mathcal{J}_R$ their complement (all audio coordinates lie in $\mathcal{J}_R$). A secret uniform block-level permutation acts on $\mathcal{J}_R$, followed by independent per-copy masks; $M_{rand}$ is:

$$M_{rand}[j] = \begin{cases} B^{diff}[j], & j \in \mathcal{J}_I, \\ B^{diff}[j] \oplus \mathcal{K}[j], & j \in \mathcal{J}_R. \end{cases} \tag{5}$$

Randomization makes $M_{rand}|_{\mathcal{J}_R}$ pseudorandom, while the fixed public copies support direct index recovery (Thietke et al., 2025).

**Inverse Transform Sampling** We map the binary stream $M_{rand}$ to the continuous Gaussian latent $\mathbf{z}$ using the Inverse Probability Integral Transform. Let $\Phi(\cdot)$ be the CDF of the standard normal distribution $\mathcal{N}(0,1)$, and let ppf $= \Phi^{-1}$. For a standard uniform variable $u_i \sim \mathcal{U}(0,1)$, the latent value $z_i$ is sampled as:

$$z_i = \text{ppf}\left(\frac{u_i + M_{rand}[i]}{2}\right). \tag{6}$$

This effectively maps bit 0 to the negative half-Gaussian and bit 1 to the positive half-Gaussian.

**Implementation-aware prior preservation.** The sampler preserves the randomized payload distribution through the complete numerical pipeline, including clipping and finite precision. The reference uses the same public routing information and numerical operations, with independent uniform authentication signs.

**Theorem 1** (Implementation-matched distribution preservation)**.** *Let $\Gamma$ contain the public index, prompt, and sampler configuration, and let $M_R = M_{rand}|_{\mathcal{J}_R}$. Suppose the joint distinguishing advantage between $(\Gamma, M_R)$ and $(\Gamma, R)$ is at most $\epsilon_{rand}$, where $R$ is an independent uniform bit string of the same length. For the common numerical sampler, optionally followed by the joint generator, let $Y^s$ and $Y^{ref}$ denote its outputs under these respective inputs. Every efficient tester $\mathcal{A}$ satisfies*

$$\left|\Pr[\mathcal{A}(\Gamma, Y^s) = 1] - \Pr[\mathcal{A}(\Gamma, Y^{ref}) = 1]\right| \leq \epsilon_{rand}. \tag{7}$$

*Proof.* A tester for the pipeline outputs induces a bit-stream tester by applying the same numerical sampler and generator to its challenge, using independent sampling coins. The two distinguishing advantages coincide. In the ideal independent-mask experiment, XOR with any prepared payload yields uniform authentication signs, so the two pipeline distributions agree exactly. Purpose-separated key derivation and secure masking supply the computational replacement (Section A.1). □

**Full-initialization budget.** Let $D$ be the total latent dimension and $\eta_{AV}$ the public-coordinate fraction. For the ideal-mask, continuously evaluated clamped reference $P_\delta^{ref}$, Section A.1 proves

$$\frac{W_2(P_\delta^{ref}, \mathcal{N}(0, I_D))}{\sqrt{D}} \leq \sqrt{2\eta_{AV} + (1 - \eta_{AV})\chi_\delta}, \tag{8}$$

where $W_2$ is quadratic Wasserstein distance and $\chi_\delta$ is the per-coordinate mean-square displacement of the implemented clamp. Thus public routing and clipping receive an explicit distributional budget.

### 4.3 Detect: Joint Inversion & Verification

**Verification workflow.** A caller supplies the AV pair and either its original prompt $P$ or generation-request metadata from which the provider retrieves $P$. This supplies condition $c$ before inversion. The provider then recovers $\hat{I}$ from fixed public slots, retrieves $m$, derives $K_{sess}$ using $E_P$, and verifies the payload and binding. Session secrets remain server-side; the interface returns the authentication decision (Section E.5).

**Joint Flow Inversion** Let $\mathbf{x}_{query}$ be the queried AV pair in the model's joint data-latent representation, $\mathbf{z}_t$ its flow state at time $t$, and $v_\theta$ the velocity field with model parameters $\theta$. One backward ODE pass from $t = 1$ to the minimum timestep $\delta_t > 0$ estimates the initial latent $\tilde{\mathbf{z}}_0$:

$$\tilde{\mathbf{z}}_0 = \mathbf{x}_{query} + \int_1^{\delta_t} v_\theta(\mathbf{z}_t, t, c)dt. \tag{9}$$

Five-step inversion retains strong extraction performance in our ablation; solver details are in Section B. Joint inversion takes 2.07 s/item versus 4.16 s/item for two separate inversions on four RTX 5090 GPUs, with matching BA and reported peak/average VRAM of 32/22 GB (Section D).

**Decoding and verification.** Zero-thresholding, key-dependent derandomization, and majority voting recover the repeated bits. After inversion, majority voting on $\mathcal{J}_I$ recovers $\hat{I}$; lookup yields $m$ and hence $K_{sess}$. Only then is $\mathcal{J}_R$ derandomized (Section E.5). The provider reconstructs the noise-free video grid and hashes it to obtain the stable reference $H_{ideal}$. The score $S_{bind}$ is the fraction of recovered audio binding bits matching $H_{ideal}$ (Eq. (A.16)). For query $\mathbf{x}$, detector $\mathcal{D}$ accepts when video/audio bit accuracies $\mathrm{BA}_v, \mathrm{BA}_a$ exceed $\tau_{acc}$ and binding score exceeds $\tau_{bind}$:

$$\mathcal{D}(\mathbf{x}) = (\mathrm{BA}_v > \tau_{acc}) \wedge (\mathrm{BA}_a > \tau_{acc}) \wedge (S_{bind} > \tau_{bind}). \tag{10}$$

High mean BA alone does not determine this sample-level decision. Decoding and local temporal perturbations are examined in Sections C and G.

### 4.4 Adaptive Binding Security

For a fixed pool of $n_s$ sessions, let $Q_{ab} = (\mathbf{x}_v^{(a)}, \mathbf{x}_a^{(b)})$, $a \neq b$, be an ordered swap with fixed prompt and metadata. In the ideal recovered-bit model, each fixed pair's decoded $N$-bit audio binding vector is uniform conditional on its target record. For $1/2 < \tau_{bind} < 1$, define minimum match count $\kappa = \lfloor N\tau_{bind} \rfloor + 1$, mismatch radius $r_\tau = N - \kappa$, and the Hamming-ball probability

$$\varepsilon_N(b) = 2^{-N} \sum_{j=0}^{\min\{N, r_\tau + b\}} \binom{N}{j}, \qquad b \in \{0, \ldots, N\}. \tag{11}$$

**Theorem 2** (Adaptive session-pool security)**.** *Let $M = n_s(n_s - 1)$. Suppose that, except on an event of probability $\beta_g$, all allowed local perturbations of all $M$ swaps preserve their target records and change at most $b$ recovered binding bits. For adaptive search within these neighborhoods, the false-acceptance event* Forge *obeys*

$$\Pr[\mathrm{Forge}] \leq \min\{1, M\varepsilon_N(b) + \beta_g\}. \tag{12}$$

*Candidate choices and verification outcomes may be arbitrarily correlated.*

*Proof.* For any fixed pair, let $U_0, U_\delta$ be its baseline and perturbed decoded binding vectors and $h$ its target digest. On the stability event, acceptance implies $d_H(U_\delta, h) \leq r_\tau$ and $d_H(U_\delta, U_0) \leq b$, hence $d_H(U_0, h) \leq r_\tau + b$. Here $d_H$ is Hamming distance. The baseline uniform law gives probability $\varepsilon_N(b)$ for this enlarged ball. Union over the $M$ fixed pairs and add the global failure probability $\beta_g$. Every successful adaptive search must find a candidate in this union. □

**Consequences.** Pure swaps take $b = 0$, $\beta_g = 0$. For $N = 128$ and $\tau_{bind} = 0.8$, the exact single-test tail is $\varepsilon_N(0) \approx 9.67 \times 10^{-13}$; the bound for all ordered swaps from $n_s = 1{,}000$ sessions is $9.66 \times 10^{-7}$. Section A.2 proves a bound $\min\{1, q\varepsilon_{single}\}$ for $q$ binary attack queries, where $\varepsilon_{single}$ bounds the best no-verification-feedback constructor, and derives decoding-margin certificates connecting local input changes to $b$.

## 5 Experiments

### 5.1 Implementation Details.

We evaluate LTX-2 (HaCohen et al., 2026) and MOVA-720p (SII-OpenMOSS Team et al., 2026). Clips are 5 s at $256 \times 256$ and 24 fps. The $L = 512$ payload bits span the full latent tensor,

with channel/time/height/width repetition factors $(k_c, k_t, k_h, k_w) = (3, 1, 4, 4)$ and a 32-bit public index $I$. Audio grids are reshaped/repeated to their latent dimensions. I2AV images use Stable Diffusion (Rombach et al., 2022). Key derivation, ChaCha20 randomization (Bernstein, 2008), and clipped inverse sampling are detailed in Section E.

**Evaluation Prompts.** We curate 250 text prompts from VBench (Huang et al., 2024) spanning 8 semantic categories. Each prompt is paired with 4 random seeds, yielding 1,000 audio-visual samples.

**Evaluation Protocol and Metrics.** All detection thresholds are selected once on a held-out clean validation split and then frozen for swap and distortion tests, without attack-specific tuning. Extraction BA measures correct bits; watermark detection measures samples passing the BA gates; final authentication additionally checks binding (Eq. (10)). True-positive/false-negative rates (TPR/FNR) measure acceptance/rejection of same-session pairs; true-negative/false-positive rates (TNR/FPR) measure rejection/acceptance of swaps. Each swap group in Tables 3 and 5 contains 1,000 independent trials. mAVE records 0/1,000 false acceptances per group (one-sided 95% FPR upper limit: 0.30%).

### 5.2 Extraction Performance

Table 1 reports majority-vote payload BA for mAVE and unimodal baselines (Chen et al., 2023; Fernandez et al., 2024; Hu et al., 2025a; Liu et al., 2024; San Roman et al., 2024); AudioSeal instead reports a detection score. Tasks are text-to-AV (T2AV), image-to-AV (I2AV), and text-and-image-to-AV (TI2AV). BA below 1 does not imply failed detection: a sample passes when its BA exceeds the threshold. Detection rates require per-sample counts (Section F).

Table 1: **Watermark Detection and Extraction.** Det. V/A = mAVE video/audio detection rates (%). VS = VideoShield; VSeal = VideoSeal; AS = AudioSeal score. Remaining entries are extraction BA, not detection rates.

| | mAVE | Video BA | | | Audio BA / score | | | |
|---|---|---|---|---|---|---|---|---|
| Model / task | Det. V/A | VS | VSeal | mAVE | AS | WavMark | Timbre | mAVE |
| LTX-2 / T2AV | 100/100 | .953 | 1.000 | .936 | 1.000 | .994 | .993 | .915 |
| LTX-2 / I2AV | 100/99.8 | .946 | .995 | .934 | 1.000 | .995 | 1.000 | .917 |
| MOVA-720p / TI2AV | 100/99.9 | .952 | .998 | .949 | 1.000 | .993 | .999 | .928 |

### 5.3 Generation Fidelity

We evaluate deployed-system quality with both public routing and randomized authentication present. This complements the implementation-matched guarantee and full-prior budget in Theorem 1 and Eq. (8). We employ VBench metrics for video quality: *Subject Consistency* (Sub), *Motion Smoothness* (Mot), *Dynamic Degree* (Dyn), *Background Consistency* (Back), and *Imaging Quality* (Img). For audio-visual quality, we use **CLAP Score** (Wu et al., 2023) (Audio-Text relevance) and **ImageBind** (Girdhar et al., 2023) (AV alignment). Synchronization is measured via **SyncNet** (Chung & Zisserman, 2016) confidence.

Table 2: **Fidelity Comparison.** Video quality, audio–text relevance, and AV alignment. Avg averages the five video-quality metrics.

| **Method** | **Sub** ↑ | **Mot** ↑ | **Dyn** ↑ | **Back** ↑ | **Img** ↑ | **Avg** | **CLAP** ↑ | **ImBind** ↑ | **Sync** ↑ |
|---|---|---|---|---|---|---|---|---|---|
| Clean (No WM) | 0.983 | 0.988 | **0.715** | 0.966 | 0.524 | **0.835** | **0.442** | 0.117 | 0.965 |
| Uncoupled (Base) | 0.994 | 0.988 | 0.684 | 0.982 | 0.454 | 0.820 | 0.425 | 0.098 | 0.965 |
| **mAVE (Ours)** | **0.998** | **0.991** | 0.669 | **0.985** | **0.529** | 0.834 | 0.426 | **0.132** | **0.966** |

**Results.** Table 2 shows comparable quality, including synchronization scores of 0.966 and 0.965. A blind mean-opinion-score study ($N = 15$) reports clean/watermarked visual means of 4.20/4.27 and audio means of 3.47/3.53. The automatic metrics and human ratings indicate comparable generation quality.

### 5.4 BINDING SECURITY

We compare independent VideoShield + AudioSeal, consistency checks, and Shared Session-ID (decoded-tag agreement). To separate hash dependency from access to a noise-free reference, we add server-referenced random binding (SR-Rand) and shared ID (SR-ID). They retain the embedding budget and frozen calibration rule; SR-Rand also retains mAVE's binding gate (Section J).

Table 3: **Watermark Detection versus Pair Authentication.** WM det. is the authentic-pair rate passing both modality detectors before consistency/binding checks. Each method uses 1,000 independently sampled swaps. All rates use frozen thresholds. Acc = $(\mathrm{TPR} + \mathrm{TNR})/2$.

| Method | WM det. ↑ | TPR (Auth) ↑ | FNR (Auth) ↓ | TNR (Swap) ↑ | FPR (Swap) ↓ | Acc ↑ |
|---|---|---|---|---|---|---|
| Weak Base (Uncoupled) | 100% | **100%** | **0%** | 0% | 100% | 50.0% |
| Uncoupled + SyncNet | 100% | 96.2% | 3.8% | 76.2% | 23.8% | 86.2% |
| Uncoupled + ImageBind/CLAP | 100% | 97.8% | 2.2% | 65.4% | 34.6% | 81.6% |
| Shared Session-ID | 99.8% | 98.9% | 1.1% | 88.7% | 11.3% | 93.8% |
| SR-Rand | 99.7% | 99.7% | 0.3% | 93.2% | 6.8% | 96.5% |
| SR-ID | 99.7% | 99.7% | 0.3% | 95.7% | 4.3% | 97.7% |
| mAVE (Ours) | 99.8% | 99.8% | 0.2% | **100%** | **0%** | **99.9%** |

Table 3 reports 6.8%/4.3% FPR for SR-Rand/SR-ID at 99.7% TPR, versus mAVE's 0% FPR at 99.8% TPR. Figure 3 shows the original configurations.

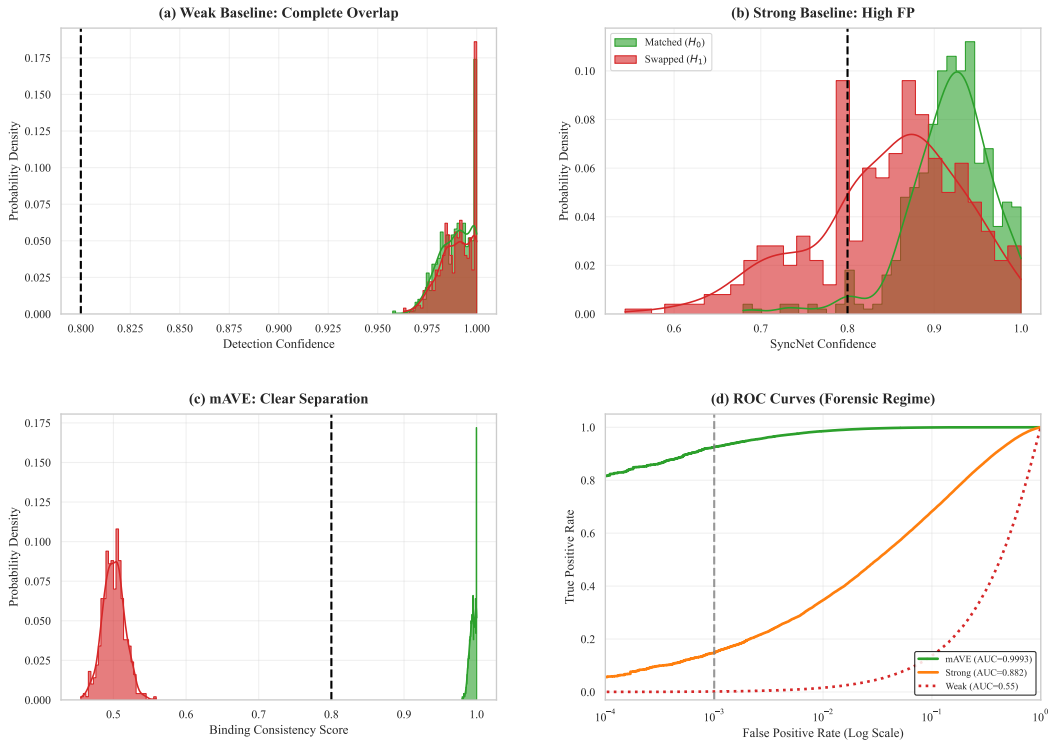


Figure 3: **Binding-security analysis.** (a) Independent watermarks overlap across authentic and swapped pairs. (b) SyncNet leaves residual overlap. (c) mAVE separates the evaluated pairs. (d) ROC curves compare the three original configurations; ImageBind/CLAP results are in Table 3.

### 5.5 ROBUSTNESS

Table 4 reports extraction and final authentication. mAVE performs well under temporal attacks: FrameAvg achieves 0.927 video BA, 99.2% TPR, and 0% observed FPR. Across the tested attacks, TPR ranges from 97.4% to 99.5%; MP4 and regeneration retain 99.0% and 98.1% TPR, respectively. Attack protocols and further temporal results appear in Sections G and H.1.

Table 4: **Robustness by Attack Domain.** Extraction reports BA, except AudioSeal (AS) scores; VS = VideoShield. TPR/FPR measure full mAVE pair authentication (%) at frozen thresholds. FNR $= 100 - \mathrm{TPR}$ and TNR $= 100 - \mathrm{FPR}$.

**(a) Video attacks**

| Attack / parameter | Extraction VS | Extraction mAVE | Auth. (%) TPR ↑ | Auth. (%) FPR ↓ |
|---|---|---|---|---|
| FrameAvg ($N = 3$) | .948 | .927 | 99.2 | 0 |
| Noise ($\sigma = .1$) | .907 | .892 | 98.9 | 0 |
| G-Blur ($r = 2$) | .947 | .939 | 99.5 | 0 |
| M-Blur ($k = 3$) | .952 | .938 | 98.6 | 0 |
| S&P ($p = .1$) | .863 | .855 | 97.7 | 0 |
| Resize (.3×) | .951 | .936 | 99.4 | 0 |
| Brightness ($f = 3$) | .943 | .914 | 99.3 | 0 |

**(b) Audio attacks**

| Attack / parameter | Extraction AS | Extraction mAVE | Auth. (%) TPR ↑ | Auth. (%) FPR ↓ |
|---|---|---|---|---|
| MP3 (64k) | .97 | .85 | 98.5 | 0.1 |
| Compression | 1.00 | .89 | 99.1 | 0 |
| Lowpass (8kHz) | 1.00 | .86 | 98.9 | 0 |
| Resample (16k→8k) | 1.00 | .88 | 99.0 | 0 |
| Noise (20dB) | .98 | .80 | 97.4 | 0.2 |
| Quantization (8-bit) | .92 | .83 | 99.1 | 0 |

**(c) Joint delivery / regeneration**

| Attack | Video extraction VS | Video extraction mAVE | Audio extraction AS | Audio extraction mAVE | Auth. (%) TPR ↑ | Auth. (%) FPR ↓ |
|---|---|---|---|---|---|---|
| MP4 (H.264 + AAC) | .932 | .940 | 1.00 | .942 | 99.0 | 0 |
| Regeneration[†] | .867 | .859 | .979 | .871 | 98.1 | 0 |

[†] Diffusion purification + EnCodec reconstruction.

### 5.6 Ablations

**Parameter ablations.** For $N \in \{16, 32, 64, 128, 256\}$, video BA exceeds 95% and audio BA is 87–91%. Five inversion steps retain over 95% video and 85% audio BA; one step yields 85.7%/78.0% (Section B).

**Binding and hard swaps.** Across three swap constructions, SR-Rand/SR-ID yield 6.6–7.1%/8.1–9.2% FPR at 99.7% TPR, versus mAVE's 0% FPR at 99.8% TPR (Table 5). Protocols and ablation definitions appear in Section J.

Table 5: **Hard-swap Controls** (%). 1,000 independent swaps per group; separate from Table 3.

| Method / control | Authentic TPR (%) ↑ | Swap FPR (%) ↓ Cross-prompt | Swap FPR (%) ↓ Same-prompt | Swap FPR (%) ↓ Selected |
|---|---|---|---|---|
| Uncoupled + SyncNet | 96.2 | 23.8 | 29.5 | 33.1 |
| Uncoupled + ImageBind/CLAP | 97.8 | 34.6 | 33.7 | 41.8 |
| Shared Session-ID | 98.9 | 11.3 | 11.2 | 10.8 |
| mAVE: no binding gate | 99.8 | 0.1 | 2.4 | 3.8 |
| mAVE: no hash or binding gate | 99.8 | 0 | 6.7 | 5.9 |
| SR-Rand | 99.7 | 6.9 | 6.6 | 7.1 |
| SR-ID | 99.7 | 8.2 | 9.2 | 8.1 |
| mAVE (full) | 99.8 | 0 | 0 | 0 |

## 6 Conclusion

mAVE advances copyright attribution from independently valid audio and video marks to authentication of their shared generation session. It binds the two modalities at latent initialization and verifies them through one joint inversion, without changing generator weights or training auxiliary watermark networks. Experiments on LTX-2 and MOVA demonstrate comparable generation quality, with 99.8% TPR and 0% observed FPR in the evaluated swap test. Strong authentication under FrameAvg, MP4 compression, and regeneration further supports its use beyond unmodified outputs. Together, these results establish session-level pairing as a practical protection objective for native joint audio-visual generation.

## AI USE STATEMENT

Generative AI tools assisted with language polishing, manuscript organization, literature discovery and summarization, and feedback on experimental design and protocol descriptions, including the additional session-binding controls, clarification of the index protocol and theoretical scope, formalization and checking of adaptive-security proofs, and design of matched server-reference controls. The authors directed these revisions and retain responsibility for the final text, claims, references, and experimental protocols.

## ETHICS STATEMENT

mAVE supports copyright protection and content provenance by authenticating the shared generation session of audio and video. It is intended to help providers and users identify misleading cross-session combinations and strengthen accountability for synthetic media. Responsible deployment should protect session secrets and prompt metadata, respect the licenses and usage conditions of underlying models and media, and communicate verification results accurately.

## REPRODUCIBILITY STATEMENT

The main paper and appendices document the method, implementation choices, and evaluation protocols to support reproducibility. Sections 4 and 5 describe the embedding and verification workflow, model configurations, and experimental setup. The appendices provide theoretical assumptions and full proofs (Section A), inversion and decoding procedures (Sections B and C), key derivation and encoding/verification pseudocode (Section E), metric definitions (Section F), prompt examples and attack protocols (Sections G and H), and session-binding controls (Section J).

## REFERENCES


Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. WAVES: Benchmarking the robustness of image watermarks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1456–1492. PMLR, 2024. URL https://proceedings.mlr.press/v235/an24a.html.

Vishal Asnani, Shruti Agarwal, and John Collomosse. FlowMark: Mask-guided video watermarking. *arXiv preprint arXiv:2607.05261*, 2026. URL https://arxiv.org/abs/2607.05261.

Daniel J. Bernstein. ChaCha, a variant of Salsa20, 2008. URL https://cr.yp.to/chacha/chacha-20080120.pdf.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. URL https://arxiv.org/abs/2311.15127.

ByteDance Seed. Seedance 2.0 official launch, 2026. URL https://seed.bytedance.com/en/blog/seedance-2-0-official-launch. February 12, 2026. Accessed: 2026-09-08.

C2PA. Content credentials: C2PA technical specification. Version 2.4, 2026. URL https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html. Accessed: 2026-09-18.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Guangyu Chen, Yu Wu, Shujie Liu, Tao Liu, Xiaoyong Du, and Furu Wei. WavMark: Watermarking for audio generation. *arXiv preprint arXiv:2308.12770*, 2023. URL https://arxiv.org/abs/2308.12770.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ivCd8z8zR2.

Facebook. Open-sourcing photo- and video-matching technology to make the internet safer, 2019. URL https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

Pierre Fernandez, Hady Elsahar, I. Zeki Yalniz, and Alexandre Mourachko. Video seal: Open and efficient video watermarking. *arXiv preprint arXiv:2412.09492*, 2024. URL https://arxiv.org/abs/2412.09492.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Girdhar_ImageBind_One_Embedding_Space_To_Bind_Them_All_CVPR_2023_paper.html.

Google DeepMind. Veo: A text-to-video generation system. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Accessed: 2026-09-08.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. URL https://doi.org/10.1080/01621459.1963.10500830.

Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. VideoShield: Regulating diffusion-based video generation models via watermarking. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=uzz3qAYy0D.

Xuming Hu, Hanqian Li, Jungang Li, Yu Huang, Shuliang Liu, Qi Zheng, Junhao Chen, and Aiwei Liu. VideoMark: A distortion-free robust watermarking framework for video diffusion models. *arXiv preprint arXiv:2504.16359*, 2025b. URL https://arxiv.org/abs/2504.16359.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 2023. URL https://proceedings.mlr.press/v202/huang23i.html.

Yu Huang, Junhao Chen, Shuliang Liu, Hanqian Li, Jungang Li, Qi Zheng, Aiwei Liu, Yi R. Fung, and Xuming Hu. Video signature: Implicit watermarking for video diffusion models. *arXiv preprint arXiv:2506.00652*, 2025. URL https://arxiv.org/abs/2506.00652.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Youngdong Jang, MinHyuk Jang, JaeHyeok Lee, Feng Yang, Gyeongrok Oh, Jongheon Jeong, and Sangpil Kim. LVMark: Robust watermark for latent video diffusion models. *arXiv preprint arXiv:2412.09122*, 2024. URL https://arxiv.org/abs/2412.09122.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.00364.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Int. Conf. Comput. Vis.*, pp. 5148–5157, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

Minyoung Kim, Sehwan Park, Sungmin Cha, and Paul Hongsuck Seo. Cross-modal watermarking for authentic audio recovery and tamper localization in synthesized audiovisual forgeries. In *Interspeech 2025*, pp. 5103–5107, 2025. doi: 10.21437/Interspeech.2025-316. URL https://www.isca-archive.org/interspeech_2025/kim25e_interspeech.html.

Hugo Krawczyk and Pasi Eronen. HMAC-based extract-and-expand key derivation function (HKDF). Technical Report 5869, RFC Editor, 2010. URL https://www.rfc-editor.org/rfc/rfc5869.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director, 2026. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be. February 5, 2026. Accessed: 2026-09-08.

De Zhang Lee, Han Fang, Hanyi Wang, and Ee-Chien Chang. Removal attack and defense on AI-generated content latent-based watermarking. *arXiv preprint arXiv:2509.11745*, 2025. URL https://arxiv.org/abs/2509.11745.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. AMT: All-pairs multi-field transforms for efficient frame interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9801–9810, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_AMT_All-Pairs_Multi-Field_Transforms_for_Efficient_Frame_Interpolation_CVPR_2023_paper.html.

Chang Liu, Jie Zhang, Tianwei Zhang, Xi Yang, Weiming Zhang, and Nenghai Yu. Detecting voice cloning attacks via timbre watermarking. In *Network and Distributed System Security Symposium (NDSS)*, 2024. doi: 10.14722/ndss.2024.24200. URL https://www.ndss-symposium.org/wp-content/uploads/2024-200-paper.pdf.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. $I^2$sb: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ai.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. URL https://arxiv.org/abs/2209.03003.

Xiyang Luo, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang. DVMark: A deep multiscale framework for video watermarking. *IEEE Transactions on Image Processing*, 34:4371–4385, 2025. doi: 10.1109/TIP.2023.3251737. URL https://doi.org/10.1109/TIP.2023.3251737.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41, 2021. doi: 10.1145/3425780. URL https://doi.org/10.1145/3425780.

Yoav Nir and Adam Langley. ChaCha20 and Poly1305 for IETF protocols. Technical Report 8439, RFC Editor, 2018. URL https://www.rfc-editor.org/rfc/rfc8439.html.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27102–27112, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Oorloff_AVFF_Audio-Visual_Feature_Fusion_for_Video_Deepfake_Detection_CVPR_2024_paper.html.

OpenAI. Sora 2 is here, 2025. URL https://openai.com/index/sora-2/. Accessed: 2026-09-08.

Leyi Pan, Sheng Guan, Zheyu Fu, Luyang Si, Huan Wang, Zian Wang, Hanqian Li, Xuming Hu, Irwin King, Philip S. Yu, Aiwei Liu, and Lijie Wen. MarkDiffusion: An open-source toolkit for generative watermarking of latent diffusion models, 2025. URL https://arxiv.org/abs/2509.10569.

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 *Int. Conf. Mach. Learn.*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43180–43196. PMLR, 2024. URL https://proceedings.mlr.press/v235/san-roman24a.html.

SII-OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. MOVA: Towards scalable and synchronized video-audio generation, 2026. URL https://arxiv.org/abs/2602.08794.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pp. 402–419, 2020. URL https://doi.org/10.1007/978-3-030-58536-5_24.

Jonas Thietke, Andreas Müller, Denis Lukovnikov, Asja Fischer, and Erwin Quiring. Towards a correct usage of cryptography in semantic watermarks for diffusion models. *arXiv preprint arXiv:2503.11404*, 2025. URL https://arxiv.org/abs/2503.11404.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. doi: 10.1016/j.inffus.2020.06.014.

Avery Li-Chun Wang. An industrial-strength audio search algorithm. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, pp. 7–13, 2003. URL https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. URL https://arxiv.org/abs/2308.06571.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969. URL https://doi.org/10.1109/ICASSP49357.2023.10095969.

Yang Yang, Xinze Zou, Zehua Ma, Han Fang, and Weiming Zhang. SKeDA: A generative watermarking framework for text-to-video diffusion models. *arXiv preprint arXiv:2603.00194*, 2026. URL https://arxiv.org/abs/2603.00194.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12162–12171, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Gaussian_Shading_Provable_Performance-Lossless_Image_Watermarking_for_Diffusion_Models_CVPR_2024_paper.html.

Zijin Yang, Xin Zhang, Kejiang Chen, Kai Zeng, Qiyi Yao, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading++: Rethinking the realistic deployment challenge of performance-lossless image watermark for diffusion models. *arXiv preprint arXiv:2504.15026*, 2025. URL https://arxiv.org/abs/2504.15026.

Xuanyu Zhang, Youmin Xu, Runyi Li, Jiwen Yu, Weiqi Li, Zhipei Xu, and Jian Zhang. $V^2$A-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9818–9827, 2024. doi: 10.1145/3664647.3680904. URL https://arxiv.org/abs/2404.16824.

Xinjie Zhu, Zijing Zhao, Hui Jin, Qingxiao Guo, Yilong Ma, Yunhao Wang, Xiaobing Guo, and Weifeng Zhang. SIGMark: Scalable in-generation watermark with blind extraction for video diffusion. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tKyAD2LhnI.

# A Full Proofs of Theoretical Guarantees

## A.1 Implementation-Matched Distribution Preservation (Theorem 1)

The argument separates three objects: the exact continuous sampling law, the complete numerical implementation with a matched reference, and the reference's distance from a standard Gaussian prior. This separates cryptographic randomization strength from the costs of public routing and numerical sampling.

### A.1.1 Exact Joint Law of Inverse-Transform Embedding

Let $M \in \{0,1\}^d$ be any sign-bit vector, allowing arbitrary dependence between its coordinates. Independently draw $U_j \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$ and set $Z_j = \Phi^{-1}((U_j + M_j)/2)$. Write $s(z)_j = \mathbb{1}[z_j > 0]$ and $\varphi_d(z) = \prod_j \varphi(z_j)$, where $\varphi$ is the standard Gaussian density. Denote the standard Gaussian law by $\gamma_d = \mathcal{N}(0, I_d)$.

**Lemma 1** (Joint sign-mixture identity)**.** *The untruncated continuous sampler has joint density*

$$p_Z(z) = 2^d \Pr[M = s(z)]\varphi_d(z) \tag{A.1}$$

*almost everywhere. Moreover,*

$$\mathrm{TV}(P_Z, \gamma_d) = \mathrm{TV}(P_M, \mathrm{Unif}(\{0,1\}^d)), \tag{A.2}$$

$$D_{KL}(P_Z \| \gamma_d) = d \ln 2 - H(M), \tag{A.3}$$

*where* TV *is total variation distance, $D_{KL}$ is relative entropy, and $H$ uses natural logarithms.*

*Proof.* Given $M = m$, independent uniform sampling gives

$$p(z \mid m) = 2^d \varphi_d(z) \mathbb{1}[s(z) = m].$$

Summing over $m$ proves Eq. (A.1). Each orthant has Gaussian mass $2^{-d}$. The likelihood ratio is constant on that orthant and equals $2^d \Pr[M = m]$. Integrating the absolute density difference therefore yields

$$\tfrac{1}{2} \int |p_Z - \varphi_d| = \tfrac{1}{2} \sum_m |\Pr[M = m] - 2^{-d}|.$$

Similarly, integrating the log-likelihood ratio gives

$$D_{KL}(P_Z \| \gamma_d) = \sum_m \Pr[M = m] \ln(2^d \Pr[M = m]) = d \ln 2 - H(M),$$

with $0 \ln 0 = 0$. These are identities for the joint law, regardless of dependence among the bits. □

Independent uniform sign bits recover the complete Gaussian product law. Correlated signs are characterized by the same identities: for example, two equal uniform signs yield Gaussian marginals but joint KL divergence $\ln 2$ and TV distance $1/2$. These are information-theoretic distribution statements; computational pseudorandomness is handled by the implementation reduction below.

**General $l$-bit partitions.** For $M \in \{0, \ldots, 2^l - 1\}^d$, use $Z_j = \Phi^{-1}((U_j + M_j)/2^l)$. Define $s_l(z)_j = \lfloor 2^l \Phi(z_j) \rfloor$ away from the zero-probability endpoints. Conditional on symbol $i$, one coordinate has density

$$p(z \mid M_j = i) = 2^l \varphi(z) \mathbb{1}[\Phi^{-1}(i/2^l) < z \le \Phi^{-1}((i+1)/2^l)]. \tag{A.4}$$

Its CDF is 0 below the interval, 1 above it, and

$$F(z \mid M_j = i) = 2^l \Phi(z) - i \tag{A.5}$$

inside it. Solving $u = F(z \mid M_j = i)$ gives

$$z = \Phi^{-1}\left(\frac{u + i}{2^l}\right), \qquad u \sim \mathcal{U}(0,1), \tag{A.6}$$

and the inverse bin map is

$$i = \lfloor 2^l \Phi(z) \rfloor. \tag{A.7}$$

For the full vector, each bin product has Gaussian mass $2^{-ld}$, so the preceding proof gives $p_Z(z) = 2^{ld} \Pr[M = s_l(z)]\varphi_d(z)$, the same TV identity relative to a uniform symbol vector, and $D_{KL} = dl \ln 2 - H(M)$. mAVE uses $l = 1$.

### A.1.2 Complete Numerical Pipeline and Purpose Separation

Let $\Gamma$ collect the index $I$, prompt $P$, fixed public routing layout, latent shapes, and numerical sampling configuration. The grid, permutation, and mask material are generated in separate purpose domains. Let $B^*$ be the repeated, permuted authentication payload in physical coordinate order, so $M_R = B^* \oplus \mathcal{K}_{mask}$. Public index coordinates remain unmasked. The payload may contain repeated bits, time templates, and any deterministic digest relation.

The cryptographic premise is joint pseudorandomization: replacing $M_R$ by an independent uniform vector $R$ while preserving $\Gamma$ changes the success probability of any efficient bit-stream distinguisher by at most $\epsilon_{rand}$. A sufficient hybrid first separates the mask-purpose key from the prepared grid/permutation state, with joint advantage $\epsilon_{sep}$, and then replaces its masking stream by uniform bits, with advantage $\epsilon_{stream}$. This yields $\epsilon_{rand} \le \epsilon_{sep} + \epsilon_{stream}$. The key-separation premise includes all state used to prepare the payload, including shared-seed material. Different application labels are the standard mechanism for separating KDF purposes (Krawczyk & Eronen, 2010); masks for different physical copies and modalities occupy distinct stream positions (Nir & Langley, 2018).

In the ideal hybrid, condition on all payload and permutation preparation. Independent uniform mask bits make $B^* \oplus R$ uniform, whatever the dependence within $B^*$. Thus this sampling result treats the binding digest as arbitrary payload data and does not require a random-oracle model for that digest.

Write $\mathcal{S}_{num}(\Gamma, M_R; \omega)$ for the actual full sampler, where $\omega$ denotes its independent random coins. It includes inverse-CDF evaluation, sample-then-clamp, finite random-number precision, floating-point operations, reshaping, and public-index insertion. The matched reference replaces only the authentication sign vector by $R$ and uses the identical numerical procedure and public header. A common generator conditioned on $P$, including its output decoder and any independent randomness, can be composed with $\mathcal{S}_{num}$.

*Proof of Theorem 1.* Suppose an efficient tester $\mathcal{A}$ distinguishes the complete pipeline outputs. Construct $\mathcal{B}$ that receives $(\Gamma, c)$, samples the pipeline's independent coins, and returns $\mathcal{A}(\Gamma, \mathcal{P}_{num}(\Gamma, c))$, where $\mathcal{P}_{num}$ is either the full sampler or the sampler followed by the generator. For $c = M_R$ this is the real experiment; for $c = R$ it is the matched-reference experiment. The distinguishing advantage is identical and hence at most $\epsilon_{rand}$. Clipping, rounding, and generation are common efficient postprocessing, so no exact-arithmetic assumption is needed. With truly independent uniform masks the two input laws agree, and so do the complete output laws. □

The information-theoretic reference in the remaining analysis uses ideal uniform masking. The real finite-key implementation is connected to that reference through Eq. (7), rather than by identifying pseudorandom bits with statistically uniform bits.

### A.1.3 Exact Cost of Public Routing in the Continuous Prior

Let $D = d_I + d_R$ be the full latent dimension, where $d_I = |\mathcal{J}_I| = L_I r_I$ counts physical public-index copies and $d_R$ the complementary coordinates. Let $C(I) \in \{0,1\}^{d_I}$ be the injective repeated index code. In the ideal untruncated experiment the full sign vector is $(C(I), R)$, with $R$ independent uniform and independent of $I$. Therefore $H(C(I), R) = H(I) + d_R \ln 2$, and Lemma 1 gives

$$D_{KL}(P_Z^{ref,0} \| \gamma_D) = d_I \ln 2 - H(I). \tag{A.8}$$

For a fixed index this equals $d_I \ln 2$. For an index uniform over all $2^{L_I}$ values, it equals $(d_I - L_I)\ln 2$, and

$$\mathrm{TV}(P_Z^{ref,0}, \gamma_D) = 1 - 2^{L_I - d_I}. \tag{A.9}$$

Indeed, the legal index patterns occupy $2^{L_I}$ of $2^{d_I}$ possible public sign patterns, with uniform mass on the legal set. Conditioning on a particular public index instead gives TV $1 - 2^{-d_I}$. These formulas explicitly account for routing redundancy in the full latent distribution. The next result measures its mean-square transport cost together with the implemented clamp.

### A.1.4 The Implemented Half-Interval Clamp

The implemented rule in Section E first draws $u \sim \mathcal{U}(1/2, 1)$ and then clamps it to $[1/2 + \delta, 1 - \delta]$, with $0 < \delta < 1/4$. Define

$$\ell_\delta = \Phi^{-1}(1/2 + \delta), \qquad a_\delta = \Phi^{-1}(1 - \delta), \qquad H_\delta(x) = \min\{a_\delta, \max\{\ell_\delta, x\}\}, \quad x \geq 0.$$

For a uniform sign, the continuously evaluated output is distributed as $T_\delta(G) = \operatorname{sgn}(G) H_\delta(|G|)$, $G \sim \mathcal{N}(0, 1)$. Thus both near-zero magnitudes and extreme tails are clamped. In one dimension the law agrees with the Gaussian density on $(-a_\delta, -\ell_\delta) \cup (\ell_\delta, a_\delta)$, and has mass $\delta$ at each of $-a_\delta, -\ell_\delta, \ell_\delta, a_\delta$.

For $d$ independent uniform-sign coordinates, their common continuous interior mass is $(1 - 4\delta)^d$, and the remaining clamped mass is supported on boundaries of zero Gaussian mass. Consequently,

$$\mathrm{TV}(\mathcal{L}(T_\delta(G_1), \ldots, T_\delta(G_d)), \gamma_d) = 1 - (1 - 4\delta)^d. \tag{A.10}$$

In this continuous comparison the forward KL divergence is infinite because of the boundary atoms. The implementation-matched theorem still applies exactly; mean-square displacement provides the useful quantitative comparison to the continuous prior.

**Lemma 2** (Clamping transport cost)**.** *For the continuously evaluated clamp, define*

$$\begin{aligned} \chi_\delta &= \mathbb{E}[(H_\delta(|G|) - |G|)^2] \\ &= 2\int_0^{\ell_\delta} (\ell_\delta - x)^2 \varphi(x)\, dx + 2\int_{a_\delta}^{\infty} (x - a_\delta)^2 \varphi(x)\, dx. \end{aligned} \tag{A.11}$$

*Then* $W_2^2(\mathcal{L}(T_\delta(G_1), \ldots, T_\delta(G_d)), \gamma_d) = d\chi_\delta$. *Also* $v_\delta = \mathbb{E}[H_\delta(|G|)^2] \leq 1$.

*Proof.* The scalar map $T_\delta$ is nondecreasing, so coupling $G$ with $T_\delta(G)$ is the monotone optimal coupling for squared distance. Its cost is the two integrals in Eq. (A.11). Coordinatewise coupling attains their sum for the product laws; every joint coupling has cost at least the sum of its optimal scalar marginal costs. Hence equality holds in $d$ dimensions.

For the second moment, using $\Pr[|G| \leq \ell_\delta] = \Pr[|G| \geq a_\delta] = 2\delta$ gives

$$v_\delta = 1 - 4\delta + 2\delta(\ell_\delta^2 + a_\delta^2) + 2\ell_\delta \varphi(\ell_\delta) - 2a_\delta \varphi(a_\delta).$$

Since $\ell_\delta' = 1/\varphi(\ell_\delta)$ and $a_\delta' = -1/\varphi(a_\delta)$,

$$v_\delta' = 4\delta \left[ \frac{\ell_\delta}{\varphi(\ell_\delta)} - \frac{a_\delta}{\varphi(a_\delta)} \right] \leq 0.$$

The function $x/\varphi(x)$ is increasing for $x \geq 0$, and $\ell_\delta < a_\delta$. Together with $v_\delta \to 1$ as $\delta \downarrow 0$, this proves $v_\delta \leq 1$. □

An explicit expression useful for checking the numerical value is

$$\begin{aligned} \chi_\delta = 2[(1 + \ell_\delta^2)\delta + \ell_\delta \varphi(\ell_\delta) - 2\ell_\delta \varphi(0)] \\ + 2[(1 + a_\delta^2)\delta - a_\delta \varphi(a_\delta)]. \end{aligned} \tag{A.12}$$

For the reported $\delta = 0.001$, $\ell_\delta \approx 0.00250663$, $a_\delta \approx 3.09023231$, $\chi_\delta \approx 2.88895 \times 10^{-4}$, and $v_\delta \approx 0.996289$. The normalized authentication-region transport distance is therefore $\sqrt{\chi_\delta} \approx 0.016997$. These are analytical sampler quantities.

### A.1.5 A Full-Initialization Routing and Clipping Budget

Let $P_\delta^{ref}$ be the full ideal-mask reference law with the public index code and the preceding continuous clamp. Write $\eta_{AV} = d_I/D$ for the public-coordinate fraction.

**Lemma 3** (Full-initialization transport budget)**.** *For any distribution of the index independent of the sampling magnitudes,*

$$W_2^2(P_\delta^{ref}, \gamma_D) \leq d_I(1 + v_\delta) + d_R\chi_\delta \leq 2d_I + d_R\chi_\delta. \tag{A.13}$$

*Normalizing by D gives Eq.* (8)*.*

*Proof.* Draw $G \sim \gamma_D$ independently of $I$. At authentication coordinates, set $Z_j = T_\delta(G_j)$; at public coordinates set $Z_j = (2C(I)_j - 1)H_\delta(|G_j|)$. This constructs the required law, including any correlations among index copies. Authentication coordinates contribute $\chi_\delta$ each. At an index coordinate, symmetry of $G_j$ gives

$$\mathbb{E}[(Z_j - G_j)^2] = \mathbb{E}[H_\delta(|G_j|)^2] + \mathbb{E}[G_j^2] = v_\delta + 1,$$

since the cross term is odd in $G_j$ for either index sign. Summing these coupling costs bounds $W_2^2$ from above. Finally use $v_\delta \le 1$. Any unused coordinates sampled directly from the standard Gaussian can instead be coupled identically at zero cost, so counting them in the complementary region only loosens the bound. □

This separates two concrete contributions: the physical public-routing fraction and the clamp's scalar transport cost. At the specified clamp level, Eq. (8) becomes

$$\frac{W_2(P_\delta^{ref}, \gamma_D)}{\sqrt{D}} \le \sqrt{2\eta_{AV} + 2.88895 \times 10^{-4}(1 - \eta_{AV})}.$$

**Numerical evaluation and quality functionals.** Let $P_{num}^{ref}$ be the numerical matched-reference law in Theorem 1. If a coupling to the continuously evaluated $P_\delta^{ref}$ has normalized RMS discrepancy at most $\zeta_{num}$, the Wasserstein triangle inequality gives

$$\frac{W_2(P_{num}^{ref}, \gamma_D)}{\sqrt{D}} \le \zeta_{num} + \sqrt{2\eta_{AV} + (1 - \eta_{AV})\chi_\delta}. \tag{A.14}$$

Here $\zeta_{num}$ accounts for finite random-number precision and numerical evaluation under the chosen coupling. Unlike this comparison to a continuous Gaussian, Theorem 1 already uses identical finite-precision operations in both worlds.

For a bounded efficient output statistic $f \in [0, 1]$, let $F(z)$ be its expectation after generation from $z$, at the fixed prompt. If $F$ is $L_F$-Lipschitz with respect to $\|z - z'\|_2/\sqrt{D}$, coupling and the computational comparison yield

$$|\mathbb{E}f(X^s) - \mathbb{E}f(X_G)| \le \epsilon_{rand} + L_F\left[\zeta_{num} + \sqrt{2\eta_{AV} + (1 - \eta_{AV})\chi_\delta}\right], \tag{A.15}$$

where $X_G$ is generated from $\gamma_D$. Indeed, the same random-output test converts the bounded-statistic difference between real and matched-reference outputs to at most $\epsilon_{rand}$, while $|\mathbb{E}F(Z_{num}^{ref}) - \mathbb{E}F(G)| \le L_F\mathbb{E}\|Z_{num}^{ref} - G\|_2/\sqrt{D}$ follows from the coupling. Apply Cauchy–Schwarz and Eq. (A.14). This transfer statement parameterizes the generator/statistic sensitivity; the empirical quality results are reported directly in Section 5.3.

### A.2 Adaptive Session-Pool Security (Theorem 2)

The analysis has three complementary parts: exact single-test probabilities, adaptive search over a candidate family, and local perturbation stability. The first supplies a Hamming-ball probability; the latter two account for selection and modification without imposing independence between verification queries.

#### A.2.1 Verification Experiment and Exact Single-Test Probability

**Queries and references.** For an ordered cross-session pair, $Q_{ab} = (\mathbf{x}_v^{(a)}, \mathbf{x}_a^{(b)})$ with $a \ne b$, fix the original video prompt, request metadata, preprocessing, inversion configuration, and detection thresholds. The verifier recovers the public index, retrieves its session record, and rebuilds the ideal video grid. Let $h \in \{0, 1\}^N$ contain the $N$ distinct selected bits of its digest, and let $U \in \{0, 1\}^N$ be the audio binding vector after inversion, derandomization, and majority decoding. The score is

$$S_{bind} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[U_i = h_i] = 1 - \frac{d_H(U, h)}{N}. \tag{A.16}$$

The server-held ideal reference is fixed for a selected record, independently of perturbations to the recovered video bits. Full acceptance also requires both modality BA gates in Eq. (10).

**Baseline recovered-bit model.** For every fixed indexed pair, the ideal experiment assumes

$$\Pr[U = u \mid \text{target record}] = 2^{-N}, \qquad u \in \{0,1\}^N. \tag{A.17}$$

This is the baseline law for the decoded vector. The cryptographic construction uses independent session secrets and the PRF/random-oracle abstraction; Eq. (A.17) specifies the recovered-bit model to which the statistical bounds apply. It is imposed once per fixed base pair, rather than anew after each adaptive query. Joint distributions across different pairs may be dependent.

**Lemma 4** (Exact binding acceptance)**.** *For $1/2 < \tau_{bind} < 1$, let $\kappa = \lfloor N\tau_{bind} \rfloor + 1$ and $r_\tau = N - \kappa$. Under Eq.* (A.17)*,*

$$\Pr[S_{bind} > \tau_{bind}] = \varepsilon_N(0) = 2^{-N} \sum_{j=0}^{r_\tau} \binom{N}{j}, \qquad \Pr[\mathcal{D}(Q_{ab}) = 1] \le \varepsilon_N(0). \tag{A.18}$$

*Proof.* Fix the target record and hence $h$. Uniformity of $U$ makes the mismatch vector $U \oplus h$ uniform, so its Hamming weight has distribution $\mathrm{Binomial}(N, 1/2)$. The strict score threshold requires at least $\kappa$ matches, equivalently at most $r_\tau$ mismatches. There are $\sum_{j=0}^{r_\tau} \binom{N}{j}$ such vectors among the $2^N$ equiprobable vectors. This proves the equality. The BA gates can only reduce the set of accepted candidates, giving the full-verifier inequality. Averaging over target records preserves the bound. □

Hoeffding's inequality (Hoeffding, 1963) gives the convenient exponential form

$$\varepsilon_N(0) \le \exp[-2N(\tau_{bind} - 1/2)^2]. \tag{A.19}$$

At $N = 128$, $\tau_{bind} = 0.8$, the score needs 103 matches and allows 25 mismatches. Thus $\varepsilon_{128}(0) \approx 9.6667 \times 10^{-13}$, while the Hoeffding upper bound is approximately $9.8595 \times 10^{-11}$. These quantities describe the ideal binding experiment; the reported empirical FPR is defined separately in Section F.

Table A.1: **Analytical Single-Test Probabilities.** Exact binding acceptance and its Hoeffding upper bound at $\tau_{bind} = 0.8$.

| $N$ | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|
| $\varepsilon_N(0)$ | $1.06 \times 10^{-2}$ | $2.68 \times 10^{-4}$ | $2.28 \times 10^{-7}$ | $9.67 \times 10^{-13}$ | $2.33 \times 10^{-23}$ |
| Hoeffding bound | $5.61 \times 10^{-2}$ | $3.15 \times 10^{-3}$ | $9.93 \times 10^{-6}$ | $9.86 \times 10^{-11}$ | $9.72 \times 10^{-21}$ |

### A.2.2 Adaptive Selection from a Fixed Session Pool

Let $n_s$ be the number of sessions available before verification feedback and define the complete ordered-pair family $\mathcal{C} = \{Q_{ab} : 1 \le a, b \le n_s,\ a \ne b\}$, of size $M = n_s(n_s - 1)$. Each candidate has fixed prompt and metadata. The verifier is stateless and uses a fixed numerical verification map. The pool includes every session from which the adversary may construct a swap, not just a shortlist selected after inspecting content.

**Lemma 5** (Adaptive search over a finite pool)**.** *If each fixed pair in $\mathcal{C}$ satisfies Eq.* (A.17)*, any selection strategy restricted to $\mathcal{C}$ satisfies*

$$\Pr[\text{Forge}] \le \min\{1, M\varepsilon_N(0)\}. \tag{A.20}$$

*The strategy may inspect all generated content, rank similarities, and adapt its query order to previous binary responses.*

*Proof.* Let $E_{ab} = \{\mathcal{D}(Q_{ab}) = 1\}$. Success requires at least one accepted candidate in the fixed family, so

$$\{\text{Forge}\} \subseteq \bigcup_{a \ne b} E_{ab}, \qquad \Pr[\text{Forge}] \le \sum_{a \ne b} \Pr[E_{ab}] \le M\varepsilon_N(0).$$

The union bound uses only the individual event probabilities. Events may share target keys, video tracks, audio tracks, or decoding errors. Content-dependent ranking and feedback-dependent ordering change which candidate is submitted, but not the existence event bounded above. □

For a fixed target video, the candidate family has $n_s - 1$ donor audios, giving $(n_s - 1)\varepsilon_N(0)$. At $n_s = 1{,}000$, $N = 128$, and $\tau_{bind} = 0.8$, the all-pairs and fixed-target bounds are approximately $9.66 \times 10^{-7}$ and $9.66 \times 10^{-10}$, respectively. These bounds account for offline selection over the entire allowed family, even if few candidates are ultimately sent to the API.

### A.2.3 Binary Feedback and the All-Reject Path

A complementary reduction isolates the benefit of verification feedback. Let $K$ denote the fixed verification secrets and $\mathcal{D}_K(Q) \in \{0, 1\}$ a stateless binary verifier. Let $V$ be the adversary's side information, including model weights, generated outputs, prompts, and public indices. All queries in the following experiment are admissible attack attempts: an Accept is already a false acceptance and terminates the first-success game. The interface provides no score, error category, or other feedback. The budget $q$ counts every tested attack candidate, including any final candidate checked by the experiment.

Define $\varepsilon_{single}$ as the supremum of success probabilities over all allowed candidate constructors with the same side information, computational resources, and generation access, but no verification feedback. Generation access may be simulated with the same oracle available in the real experiment. This definition includes content-dependent offline candidate selection.

**Theorem 3** (Binary-feedback first-success bound)**.** *An adversary making at most $q$ attack-only verification queries satisfies*

$$\Pr[\text{Forge}] \le \min\{1, q\varepsilon_{single}\}. \tag{A.21}$$

*No independence of queries, or bound on their history-conditioned success probabilities, is required.*

*Proof.* Fix the adversary's random tape and couple any generation-oracle randomness between executions. Simulate a counterfactual interaction that answers every verification query with Reject. This generates an all-reject sequence

$$Q_t^{\perp} = \mathcal{A}_t(V, Q_1^{\perp}, 0, \ldots, Q_{t-1}^{\perp}, 0), \qquad t = 1, \ldots, q.$$

Here $\mathcal{A}_t$ is the next-query rule; random tapes and generation-oracle interactions are implicit. Suppose the real execution first succeeds on query $t$. All preceding responses are Reject, so its prefix agrees with the counterfactual execution, and $Q_t = Q_t^{\perp}$. Consequently,

$$\{\text{Forge}\} \subseteq \bigcup_{t=1}^{q} \{\mathcal{D}_K(Q_t^{\perp}) = 1\}.$$

For every fixed $t$, a no-verification-feedback constructor can simulate the adversary with $t-1$ dummy rejections, using the same permitted generation access, and output $Q_t^{\perp}$. Its success probability is at most $\varepsilon_{single}$. Taking probabilities over the secret environment, public view, and coupled random tapes, then applying the union bound, yields

$$\Pr[\text{Forge}] \le \sum_{t=1}^{q} \Pr[\mathcal{D}_K(Q_t^{\perp}) = 1] \le q\varepsilon_{single}.$$

The prefix argument holds even when the candidates are highly correlated or each is a modification of the previous one. □

Rejections can exclude possibilities: guessing a fresh $N$-bit password after one failed distinct guess has conditional success probability $1/(2^N - 1)$, exceeding $2^{-N}$. Nevertheless, $q$ distinct guesses succeed with probability $q/2^N$ for $q \le 2^N$. The reduction captures this distinction. The single-query quantity must bound every allowed no-feedback constructor, rather than only an unselected random swap. A $q\varepsilon_N(0)$ specialization follows if such constructors satisfy the stronger $\varepsilon_N(0)$ bound. With only the per-fixed-pair model, Lemma 5 supplies the pool-size guarantee directly. Genuine-positive diagnostic queries fall outside this attack-only first-success experiment.

### A.2.4 Adaptive Local Perturbations of a Swap

Fix a base swap $Q_0 = (\mathbf{x}_v^{(A)}, \mathbf{x}_a^{(B)})$, $A \neq B$, and candidates $Q_\delta = (\mathbf{x}_v^{(A)}, \mathbf{x}_a^{(B)} + \delta)$. Let $\Delta_{\epsilon_{atk}}$ be a prescribed perturbation set, for example $\|\delta\|_p \leq \epsilon_{atk}$ with any donor-content constraints. The budget is measured relative to the original donor audio, so successive small updates remain subject to one total budget. Prompts, request metadata, and numerical verification settings are fixed.

Let $U_0$ and $U_\delta$ be the recovered binding vectors. Define a good event $\mathcal{G}_b$ on which every $\delta \in \Delta_{\epsilon_{atk}}$ preserves the selected target record and satisfies $d_H(U_\delta, U_0) \leq b$, where $b$ is a fixed integer. Suppose $\Pr[\mathcal{G}_b^c] \leq \beta$. The server's ideal-grid reconstruction keeps the target digest $h$ fixed throughout this event.

**Lemma 6** (Adaptive neighborhood bound)**.** *Under the baseline law for $U_0$ and the stability event above,*

$$\Pr[\exists \delta \in \Delta_{\epsilon_{atk}} : \mathcal{D}(Q_\delta) = 1] \leq \min\{1, \varepsilon_N(b) + \beta\}. \tag{A.22}$$

*Proof.* On $\mathcal{G}_b$, an accepted perturbation has $d_H(U_\delta, h) \leq r_\tau$. The triangle inequality gives

$$d_H(U_0, h) \leq d_H(U_0, U_\delta) + d_H(U_\delta, h) \leq b + r_\tau.$$

Thus existence of an accepted perturbation implies either $d_H(U_0, h) \leq r_\tau + b$ or failure of $\mathcal{G}_b$. The original uniform law for $U_0$ gives the enlarged-ball probability $\varepsilon_N(b)$; the failure event contributes at most $\beta$. This union argument does not condition the baseline distribution on stability, and requires no independence between those events. □

The bound covers every candidate in the allowed neighborhood at once, so adaptive query count does not enter. Perturbed vectors may be biased and mutually dependent. When $\tau_{bind} - b/N > 1/2$, its statistical term also satisfies

$$\varepsilon_N(b) \leq \exp[-2N(\tau_{bind} - 1/2 - b/N)^2]. \tag{A.23}$$

To see this, $d_H(U_0, h) \leq r_\tau + b$ requires at least $\kappa - b > N\tau_{bind} - b$ baseline matches; apply Hoeffding above the mean $N/2$. Thus changing at most $b$ decoded bits reduces the effective threshold by $b/N$ in this exponential bound.

Table A.2: **Analytical Neighborhood Bounds.** Prescribed decoded-bit budgets with $N = 128$, $\tau_{bind} = 0.8$, and $\beta = 0$. These are parameter illustrations of Eq. (A.22).

| Bit-change budget $b$ | Radius $r_\tau + b$ | $\varepsilon_{128}(b)$ |
|---|---|---|
| 0 | 25 | $9.67 \times 10^{-13}$ |
| 4 | 29 | $1.92 \times 10^{-10}$ |
| 8 | 33 | $1.91 \times 10^{-8}$ |
| 16 | 41 | $2.94 \times 10^{-5}$ |

### A.2.5 From Numerical Inversion to Decoded-Bit Stability

Let $\mathcal{F}$ denote the complete deterministic numerical mapping from queried media to recovered joint latents, including preprocessing, latent encoding, and the implemented inversion solver. Suppose a certified bound gives

$$\sup_{\delta \in \Delta_{\epsilon_{atk}}} \|\mathcal{F}(Q_\delta) - \mathcal{F}(Q_0)\|_\infty \leq \eta. \tag{A.24}$$

A local operator-norm bound $L_\mathcal{F}$ for the actual map from input $\ell_p$ perturbations to output $\ell_\infty$ changes supplies $\eta = L_\mathcal{F}\epsilon_{atk}$. The map in this statement includes the implemented finite-step solver; the continuous ODE alone is not the numerical certificate.

Let $\widetilde{z}_0 = \mathcal{F}(Q_0)$. Any coordinate satisfying $|\widetilde{z}_0[j]| > \eta$ retains its sign throughout the neighborhood. For logical binding bit $i$, let $G_i$ be its physical repetition-copy set under the fixed recovered record, and let $s_j \in \{-1, +1\}$ be each baseline sign after XOR derandomization. Define the vote margin and the number of potentially changing copies as

$$\gamma_i = \left| \sum_{j \in G_i} s_j \right|, \qquad e_i(\epsilon_{atk}) = \#\{j \in G_i : |\widetilde{z}_0[j]| \leq \eta\}. \tag{A.25}$$

**Lemma 7** (Majority-margin certificate)**.** *If Eq.* (A.24) *holds and the target record is fixed, then*

$$b_{cert}(\epsilon_{atk}) = \#\{i : \gamma_i \le 2e_i(\epsilon_{atk})\} \tag{A.26}$$

*upper-bounds the number of decoded binding bits that can change under any allowed perturbation.*

*Proof.* Only copies with $|\widetilde{z}_0[j]| \le \eta$ can change sign. Each flipped copy changes its signed vote by at most 2, so bit $i$'s vote sum changes by at most $2e_i$. If $\gamma_i > 2e_i$, the sum cannot cross or reach zero, and its majority decision stays fixed. Counting all remaining groups gives Eq. (A.26). Groups at equality, including ties under the decoder's fixed tie rule, are conservatively counted as potentially changing. Fixed permutation only assigns copies to groups, and XOR only reverses their signs, so neither changes the counting argument. □

Apply the same margin test to the public, unmasked index-copy groups $\mathcal{J}_I(\ell)$ for $\ell = 1, \dots, L_I$. Certifying every index bit preserves the lookup result; fixed prompt and metadata then preserve the session key, permutation, and masks used in the binding certificate. The operational chain is therefore

$$\text{input budget} \implies \eta \implies \text{copy-sign margins} \implies (\text{stable index}, b_{cert}) \implies \text{binding bound}.$$

For probabilistic reporting, choose a fixed budget $b$ and bound the probability that the inversion certificate, index certificate, or inequality $b_{cert} \le b$ fails. This yields $\beta$ for a pair or $\beta_g$ jointly for a pool. A sample-dependent certificate is used through this failure event, rather than by conditioning the uniform baseline law on its observed value. No independence between certificate failures and initial binding vectors is needed.

### A.2.6 Composition over Session Pools and Perturbation Neighborhoods

Let $\mathcal{G}_b^{pool}$ be the event that every base pair in $\mathcal{C}$ and every permitted perturbation preserves its record and changes at most $b$ binding bits. Set $\Pr[(\mathcal{G}_b^{pool})^c] \le \beta_g$. This event controls the entire allowed search family, including switches between targets, donor audios, and local modifications.

*Proof of Theorem 2.* On $\mathcal{G}_b^{pool}$, acceptance anywhere in the search family implies $d_H(U_{ab}, h_{ab}) \le r_\tau + b$ for at least one fixed base pair. Hence

$$\{\text{Forge}\} \subseteq \bigcup_{a \ne b} \{d_H(U_{ab}, h_{ab}) \le r_\tau + b\} \cup (\mathcal{G}_b^{pool})^c.$$

Apply the baseline uniform law separately to each pair and union-bound the $M$ enlarged-ball events. Charge the global failure event once. This gives $M\varepsilon_N(b) + \beta_g$, capped at 1. Adaptive decisions may exploit content and any earlier accept/reject feedback because the bound concerns existence of a successful candidate anywhere in the allowed family. □

If only a per-pair failure bound $\beta$ is available, the corresponding bound is

$$\Pr[\text{Forge}] \le \min\{1, M[\varepsilon_N(b) + \beta]\}. \tag{A.27}$$

For pure swaps, $b = 0$ and $\beta_g = 0$, recovering Lemma 5. The finite-pool and neighborhood statements are statistical results in the ideal baseline model; Theorem 3 separately converts any established no-feedback success bound into a binary-query bound. Any real-to-ideal cryptographic reduction is accounted for through its distinguishing advantage: if it bounds the relevant adversary-success probability gap by $\nu$, add $\nu$ to the ideal bound and cap at 1. This keeps the statistical search factor and cryptographic reduction loss separate.

## B Inversion Dynamics and Solver Analysis

### B.1 ODE Solver Discretization and Truncation Drift

In this section only, $N$ counts solver steps, not binding bits; $\epsilon = \delta_t$ is the final integration time. The forward ODE for Rectified Flow (Liu et al., 2022) models define the velocity field:

$$\frac{d\mathbf{z}_t}{dt} = v_\theta(\mathbf{z}_t, t, c), \tag{B.1}$$

where $\mathbf{z}_0 \sim \mathcal{N}(0, I)$ is the initial noise and $\mathbf{z}_1$ is the generated sample. For inversion, we solve the ODE backwards from $t = 1$. We use the Euler method (Karras et al., 2022) with a uniform time schedule $\{t_N = 1, t_{N-1}, \ldots, t_1 = \epsilon\}$ where $N$ is the number of inversion steps:

$$\mathbf{z}_{t_{i-1}} = \mathbf{z}_{t_i} + (t_{i-1} - t_i) \cdot v_\theta(\mathbf{z}_{t_i}, t_i, c). \tag{B.2}$$

**The $\delta_t$ Truncation Gap.** In theoretical implementations, integrating exactly to $t = 0$ yields perfect inversion. However, to ensure generation quality and avoid the numerical singularity near pure noise ($t = 0$) where the velocity field $v_\theta$ may become unbounded, state-of-the-art joint audio-visual generation models (MOVA and LTX-2) enforce a strict minimum timestep $t = \delta_t > 0$ in their sampling pipelines. Consequently, the inversion process must be truncated to maintain stability. This discretization gap introduces a deterministic drift, preventing the raw Bit Accuracy from reaching a perfect 1.0.

### B.2 Inversion Fidelity: DDIM vs. Rectified Flow

Prior works utilizing DDIM (Song et al., 2021) solve a related ODE in the noise-prediction parameterization:

$$\mathbf{z}_{t_{i-1}} = \sqrt{\alpha_{t_{i-1}}} \left( \frac{\mathbf{z}_{t_i} - \sqrt{1 - \alpha_{t_i}} \cdot \epsilon_\theta(\mathbf{z}_{t_i}, t_i)}{\sqrt{\alpha_{t_i}}} \right) + \sqrt{1 - \alpha_{t_{i-1}}} \cdot \epsilon_\theta(\mathbf{z}_{t_i}, t_i). \tag{B.3}$$

To systematically understand the impact of the solver paradigm versus the aforementioned $\delta_t$ truncation, we compare inversion fidelity across three distinct experimental settings:

- **DDIM-based Inversion (VideoShield (Hu et al., 2025a)):** Evaluated on ModelScope (Wang et al., 2023) (T2V) and Stable-Video-Diffusion (Blattmann et al., 2023) (I2V), integrating fully to the noise boundary.
- **Ideal Rectified Flow Inversion:** Evaluated using the official Rectified Flow codebase on CIFAR-10 (Krizhevsky, 2009), utilizing an Euler ODE solver that integrates fully to $t = 0$.
- **mAVE (Ours) on MOVA-720p:** Operating under the enforced $\delta_t$ truncation inherently required by native joint generation pipelines.

Table B.1: **Inversion Accuracy vs. Number of Steps across Frameworks.** Both DDIM and ideal Rectified Flow achieve near-perfect inversion ($\sim 1.000$) given sufficient steps. In our mAVE pipeline, performance is marginally bottlenecked by the necessary $\delta_t$ truncation rather than the underlying solver paradigm, yet it remarkably maintains high robustness even down to 5 steps.

| **Framework / Model** | **Task** | **Inversion Steps ($N$)** 50 | 25 | 10 | 5 | 1 |
|---|---|---|---|---|---|---|
| *Baseline Paradigms (Full Integration to $t = 0$)* | | | | | | |
| DDIM (ModelScope) | T2V | 1.000 | 1.000 | 1.000 | - | - |
| DDIM (SVD) | I2V | 0.999 | 0.999 | 0.999 | - | - |
| Rectified Flow (CIFAR-10) | Uncond | 1.000 | 1.000 | 1.000 | - | - |
| *mAVE with $\delta_t$ Truncation (MOVA-720p)* | | | | | | |
| mAVE Video BA | TI2AV | 0.967 | 0.961 | 0.955 | 0.958 | 0.857 |
| mAVE Audio BA | TI2AV | 0.913 | 0.892 | 0.901 | 0.865 | 0.780 |

As demonstrated in Table B.1, both DDIM and Flow Matching naturally provide excellent inversion guarantees in practice when integrated fully. For mAVE, the slight performance gap is fundamentally a product of the $\delta_t$ truncation mechanism, rather than an inherent weakness of the Euler solver.

Crucially, despite this truncation drift, the straight-line transport property of Rectified Flow allows mAVE to remain highly resilient under extremely constrained computational budgets. The accuracy is solidly maintained when scaled down to just 5 inversion steps, with significant degradation only occurring at an extreme 1-step inversion. This property is highly advantageous for efficient, real-world deployment.

# C Decoding Method Comparison: Median versus Zero Thresholding

After inversion, the recovered latent $\tilde{\mathbf{z}}_0$ exhibits minor drift due to the discretization gap. We investigated two decoding strategies:

**Zero-Thresholding (Standard).** The simplest approach: each bit is decoded as:

$$\tilde{b}[i] = \mathbb{1}[\tilde{\mathbf{z}}_0[i] > 0]. \tag{C.1}$$

This leverages the fundamental asymmetry of the half-Gaussian constraint: watermarked coordinates are guaranteed to have the correct sign in the noise-free case, and the redundancy from blockwise repetition absorbs moderate sign flips.

**Median Thresholding.** This approach estimates a per-block threshold by computing the median of all $k_{all}$ copies of each bit, then decodes relative to this median:

$$\mu_j = \text{median}\{\tilde{\mathbf{z}}_0[i] : i \in \text{Block}_j\}, \quad \tilde{b}[j] = \mathbb{1}[\mu_j > 0]. \tag{C.2}$$

The motivation is to correct for a potential global bias in the recovered latents. However, since Rectified Flow inversion introduces very little systematic bias (the drift is primarily random noise), the median and zero thresholds produce nearly identical results.

Table C.1: **Decoding Strategy Comparison.** Bit Accuracy on LTX-2 T2AV with 25 inversion steps. The negligible difference confirms that zero-thresholding is sufficient.

| Decoding Method | Video BA | Audio BA |
|---|---|---|
| Zero-Thresholding | 0.936 | 0.915 |
| Median Thresholding | 0.940 | 0.913 |

As shown in Table C.1, the performance difference is marginal and inconsistent across modalities. We therefore adopt the simpler Zero-Thresholding in all experiments to avoid introducing an unnecessary hyperparameter.

# D The Performance Analysis of Joint Inversion

A key architectural advantage of mAVE is that both embedding and extraction operate through the *same* joint denoising / inversion pass used by the underlying audio-visual model. In this section we empirically verify that this single joint inversion recovers watermark bits as accurately as performing two dedicated, modality-specific inversions, and we discuss the resulting efficiency gain.

## D.1 Setup

We compare two extraction protocols on LTX-2 T2AV (25 inversion steps, zero-thresholding) using the same four NVIDIA RTX 5090 GPUs and timing protocol. Joint inversion has reported peak/average VRAM usage of 32/22 GB. The timing comparison is between one joint inversion and two separate modality inversions:

- **Joint Inversion (default).** A single backward ODE pass is performed through the bimodal model, simultaneously recovering both the video latent $\tilde{\mathbf{z}}_v$ and the audio latent $\tilde{\mathbf{z}}_a$.
- **Separate Inversion.** Two independent backward ODE passes are performed: one using only the video branch to recover $\tilde{\mathbf{z}}_v$, and one using only the audio branch to recover $\tilde{\mathbf{z}}_a$. Each pass uses the same 25 inversion steps and solver configuration.

## D.2 Results

As shown in Table D.1, the two protocols yield virtually identical Bit Accuracy for both modalities, confirming that the joint denoising trajectory preserves per-modality fidelity during inversion. This is expected: in Rectified-Flow-based joint models, the audio and video latents share the same ODE velocity field but occupy orthogonal sub-spaces of the joint latent, so inverting them together introduces no additional cross-modal interference compared with inverting them separately.

Table D.1: **Joint vs. Separate Inversion (LTX-2 T2AV).** Bit Accuracy and wall-clock time comparison. Joint inversion matches the extraction quality of two separate passes while halving the computational cost.

| Extraction Protocol | Video BA | Audio BA | Time (s/item) |
|---|---|---|---|
| Joint Inversion (ours) | 0.936 | 0.915 | 2.07 |
| Separate Inversion | 0.936 | 0.915 | 4.16 |

### D.3 Efficiency Implications

The practical consequence is a significant reduction in detection cost relative to baseline combinations. Table D.2 summarises the pipeline-level comparison.

Table D.2: **Extraction Pipeline Comparison.** mAVE uses one joint inversion; baseline combinations additionally invoke an audio extractor. Pass counts alone do not establish wall-clock speedups across different extractors.

| Method | Video Extraction | Audio Extraction | Binding Check | Total Passes |
|---|---|---|---|---|
| VideoShield + AudioSeal | ODE inversion ($N$ steps) | AudioSeal encoder (forward) | — | 2 |
| VideoShield + WavMark | ODE inversion ($N$ steps) | WavMark encoder (forward) | — | 2 |
| **mAVE (ours)** | Joint ODE inversion ($N$ steps) | | $\tau_{bind}$ | **1** |

The measured latency decreases from 4.16 to 2.07 s/item, a $2.01\times$ speedup (50.2% reduction), while video/audio BA remains 0.936/0.915. This supports replacing two separate inversions with one joint pass on the measured setup. It does not establish the same speedup relative to VideoShield plus AudioSeal or WavMark, whose neural-extractor costs were not measured here. Likewise, the reported hardware and VRAM figures characterize this deployment rather than multi-GPU scaling or single-GPU feasibility.

## E Complete Cryptographic Key Derivation

This section provides the end-to-end derivation of the session-specific randomization, from payload to continuous latent initialization.

The consolidated symbol glossary is in Section K.

### E.1 Payload Structure: Index and Secret

The overall watermark payload comprises two functionally distinct segments:

- **Plaintext Index** $I \in \{0,1\}^{L_I}$ ($L_I = 32$ bits) *(embedded)*: A non-confidential session identifier repeated at fixed public video coordinates $\mathcal{J}_I$. These coordinates bypass both secret permutation and ChaCha20 masking. Their inverse-sampled signs permit majority-vote index recovery without a key, after prompt-conditioned inversion. No digest, template, or secret payload shares these coordinates.
- **Secret Payload** $m \in \{0,1\}^{L_m}$ *(non-embedded)*: A high-entropy secret stored exclusively on the server-side database, indexed by $I$. The payload $m$ is *never* embedded in the latent; it participates only in key derivation.

### E.2 Session Key Computation

Given the secret payload $m$ and the generation prompt $P$, the session key is computed as:

$$K_{sess} = \text{SHA-256}\big(\text{Prefix}_{\lambda_{key}}(\text{SHA-256}(m)) \parallel \text{SHA-256}(\text{Norm}(P))\big), \tag{E.1}$$

where $\text{Prefix}_{\lambda_{key}}(\cdot)$ retains a fixed key-security prefix; $\lambda_{key}$ is independent of the binding length $N$, $\parallel$ denotes concatenation, and $\text{Norm}(\cdot)$ applies lowercase conversion and whitespace trimming to the

prompt string. The prompt binds the key derivation to the generation condition; it is not a password or a source of secret entropy. The adversary may know it. Even when prompts coincide after normalization, fresh independent session secrets produce distinct session keys with overwhelming probability. Changing the prompt changes derived material, but does not by itself prove resistance to general copy attacks.

At generation time, the provider stores the original prompt $P$, public index $I$, and secret $m$ in the session record. Before inversion, the original $P$ is either supplied by the caller or retrieved by the provider using the generation-request metadata available to the verification service. This service-side request reference is distinct from the embedded index $I$. The verifier conditions the joint ODE inversion on $P$, recovers $\hat{I}$ from the resulting video latent, and uses $\hat{I}$ to retrieve $m$ and derive $K_{sess}$ with $E_P = \mathrm{Norm}(P)$. Thus prompt acquisition precedes inversion and watermark-index extraction; $P$ is not inferred from the received content. Key secrecy rests on the server-held $m$, not on hiding $P$. Normalization affects key derivation only; the generator is conditioned on the original prompt. Metadata and $I$ locate records rather than authenticate content. The protocol assumes trusted session records; acceptance is decided by recovered payload and binding checks.

**Purpose-separated keys.** Let KDF denote the purpose-separated derivation and let $\mathrm{label}_x$ be distinct fixed application labels. We use

$$K_x = \mathrm{KDF}(K_{sess}, \mathrm{label}_x), \qquad x \in \{grid, mask, perm\}.$$

Grid construction uses the grid-purpose subkey together with the shared seed defined below; block-shuffle randomness and per-copy masking use their own domains. Purpose-label separation follows the standard KDF design principle (Krawczyk & Eronen, 2010). The joint randomization premise and its reduction are specified in Section A.1. The fixed $\lambda_{key}$ controls key derivation throughout the binding-length ablation; changing $N$ changes digest coverage, not key-prefix length.

### E.3 Shared Seed Extraction

To ensure correlated base entropy across the audio and video modalities, we derive a shared seed from the secret payload $m$:

$$\mathrm{seed}_{shared} = \text{SHA-256}(m)[:8], \tag{E.2}$$

*i.e.*, the first 8 bytes (64 bits) of the hash of the secret payload $m$. This shared seed initializes both the audio and video pseudo-random streams, ensuring that their bit grids are structurally aligned for the binding verification.

### E.4 ChaCha20 Stream Generation

The ChaCha20 stream cipher (Bernstein, 2008) is initialized with:

- **Key:** Mask-purpose subkey $K_{mask}$ (256-bit).
- **Nonce:** A fixed zero nonce (96-bit all-zeros). This fixed nonce requires a fresh masking key for each generation request. Reusing a prompt does not authorize reusing the session secret or a key/nonce pair; video/audio masks occupy disjoint stream slices.
- **Counter:** Starting from 0.

ChaCha20 (Nir & Langley, 2018) generates the masking byte stream $\mathcal{S}$ from $K_{mask}$, with disjoint video/audio slices and a distinct pseudorandom mask bit for every physical copy. Separate randomness derived from $K_{perm}$ drives the uniform shuffle of eligible blocks; $\pi$ is its induced coordinate map. Public index copies remain fixed. At $j \in \mathcal{J}_I$, the sign bit is the unmasked index bit; at $j \in \mathcal{J}_R$, it is the repeated, permuted payload bit XOR its per-copy mask.

### E.5 Public Index, Encoding, and Verification

**Public layout.** Let $r_I$ be the number of repetitions per index bit and $\mathcal{J}_I(b)$ the fixed set of $r_I$ video coordinates assigned to index bit $b \in \{1, \ldots, L_I\}$. The sets are disjoint and known before generation; $\mathcal{J}_I = \bigcup_b \mathcal{J}_I(b)$. They do not depend on $I$, the prompt, or a secret. All keyed permutations and mask slices act only on $\mathcal{J}_R$. Other unused coordinates, if any, are sampled independently from

the standard Gaussian. The public index provides indexed record retrieval before secret-dependent decoding, avoiding a search over all stored session keys. This routing function is separate from payload authentication: knowing an address does not establish a valid binding.

**Coordinate accounting.** Let $d_v$ and $d_a$ be the numbers of scalar initial-latent coordinates in video and audio. Public index copies are disjoint, so $|\mathcal{J}_I| = L_I r_I$. Define the video and joint public-coordinate fractions as

$$\eta_v = \frac{L_I r_I}{d_v}, \qquad \eta_{AV} = \frac{L_I r_I}{d_v + d_a}.$$

The complementary joint fraction is $1 - \eta_{AV}$, and audio contains no public index copies. These counts concern physical latent coordinates, not the ratio of logical index bits to payload bits. The payload repetition product is $k_{all} = k_c k_t k_h k_w = 48$ for the reported $(3, 1, 4, 4)$ setting; $r_I$ is the separate index repetition count. Repeat fixed logical payload fields first, uniformly permute the resulting blocks within the eligible region of each modality, then apply a distinct pseudorandom mask bit to every physical copy. Even copies of the same payload bit do not reuse a mask bit. Audio reshaping and any additional replication must be accounted for in $d_a$ and the corresponding decoding map.

**Binding positions and digest bits.** The logical fields for base bits, templates, and binding bits have fixed allocation. Audio binding positions $\mathcal{I}_{bind}$ are disjoint from template positions $\mathcal{I}_{time}$; their physical blocks are relocated by the secret uniform block permutation. Slot selection is thus fixed in logical space and secret in physical space. The fixed digest selector $\phi$ uses 128 different SHA-256 bits by default, or $N$ distinct bits in the binding-length ablation ($N \leq 256$); repeated physical copies provide decoding redundancy and do not increase this number. This separates logical payload allocation, physical coordinate mapping, and the number of comparisons used in the probability bound.

**Encoding pseudocode.**

1. Allocate an unused index $I$ and a fresh independent secret $m$; store $(I, m, P)$ in the provider's session record. Derive $K_{sess}$ as in Eq. (2), then the grid, mask, and permutation subkeys.
2. Construct $B_v$ with its template, index, and base bits. Compute its digest and construct $B_a$ with the designated binding bits. Repeat the grids to obtain $B^{diff}$.
3. Keep index copies at $\mathcal{J}_I(b)$. Reconstruct the keyed block permutation and disjoint video/audio per-copy mask slices on $\mathcal{J}_R$. Set $M_{rand}[j] = I_b$ for $j \in \mathcal{J}_I(b)$ and $M_{rand}[j] = B^{diff}[j] \oplus \mathcal{K}[j]$ on the permuted complement.
4. Draw independent uniform variates and apply Eq. (6); the practical version uses the clipping specified below. Generate the AV pair with the original prompt $P$.

**Verification pseudocode.**

1. Obtain the original prompt from the caller or the provider's request metadata. Jointly invert the received AV pair once to recover approximate initial latents. No watermark key is needed for this model inversion.
2. Read signs at the public positions and decode each index bit as $\hat{I}_b = \mathbb{1}[\sum_{j \in \mathcal{J}_I(b)} \mathbb{1}[\tilde{z}_j > 0] > r_I/2]$; ties decode to zero. No keyed permutation or XOR is applied at this stage.
3. Query the session database with $\hat{I}$ to obtain $m$, then derive $K_{sess}$ and its purpose subkeys from $m$ and the normalized original prompt. This is the first step requiring the session secret.
4. Reconstruct the block permutation and per-copy mask slices on $\mathcal{J}_R$; undo masking and permutation, then majority-decode the payload. Rebuild the ideal grids and video digest from the retrieved record.
5. Apply the video/audio BA gates and the binding gate in Eq. (10). Successful index recovery alone does not authenticate either modality or their pairing.

**Routing, authentication, and fidelity.** Public repetition enables direct record lookup; secret randomization protects the authentication payload. Theorem 1 covers the complete numerical sampler

and generated output against a reference with the same routing header. Section A.1 additionally gives the exact untruncated joint law, the index entropy cost, and the full routing/clipping transport budget. Tables 2 and E.1 report complete-system quality and index recovery.

**Exact index recovery.** Let $n_I$ be the number of evaluated samples. The exact-index recovery rate is $R_I = n_I^{-1} \sum_{k=1}^{n_I} \mathbb{1}[\hat{I}^{(k)} = I^{(k)}]$, where equality requires all $L_I = 32$ bits to match. The denominator includes every attempted sample, not only successful lookups. Index BA and payload BA cannot substitute for this rate. On the clean sets in Table E.1, exact index recovery is 100% for each task. This separately measures record recovery before payload detection and binding acceptance.

Table E.1: Exact 32-bit index recovery on clean samples. Every index bit must match; all attempted samples are included. These rates do not describe attacked samples.

| Model / task | Exact index recovery (%) |
|---|---|
| LTX-2 / T2AV | 100 |
| LTX-2 / I2AV | 100 |
| MOVA-720p / TI2AV | 100 |

### E.6 Numerical Inverse Sampling

For each watermark coordinate $i \in \mathcal{I}_w$ with assigned sign bit $y_i$, draw an independent $u_i \sim \mathcal{U}(0.5, 1)$ and apply the actual sample-then-clamp rule

$$u_i^\delta = \min\{1-\delta, \max\{0.5+\delta, u_i\}\}, \qquad z_i = \Phi^{-1}\big(y_i u_i^\delta + (1-y_i)(1-u_i^\delta)\big). \tag{E.3}$$

We use $\delta = 0.001$, giving $|z_i| \in [\Phi^{-1}(0.501), \Phi^{-1}(0.999)] \approx [0.00250663, 3.09023231]$. Sampling precedes clamping, so the boundary values carry probability mass. Theorem 1 includes this numerical operator, and Section A.1 quantifies its near-zero and tail contributions to the prior budget. Non-watermark coordinates $j \notin \mathcal{I}_w$ are sampled from the standard Gaussian.

## F Formal Metric Definitions

### F.1 Bit Decision Rule (Majority Voting)

Here $m_j$ denotes a generic payload bit, not the server secret $m$; $k_{all} = k_c k_t k_h k_w$ is its repetition count, and $\pi(j,l)$ locates copy $l$ of bit $j$. Hats denote decoded estimates and tildes denote recovered values. The block-wise repetition scheme embeds each payload bit $m_j$ into $k_{all}$ copies across the latent tensor. After inversion, the recovered copies are $\{\tilde{b}_{j,1}, \tilde{b}_{j,2}, \ldots, \tilde{b}_{j,k_{all}}\}$ where $\tilde{b}_{j,l} = \mathbb{1}[\tilde{\mathbf{z}}_0[\pi(j,l)] > 0]$. The final decision for bit $j$ is:

$$\hat{m}_j = \mathbb{1}\left[\sum_{l=1}^{k_{all}} \tilde{b}_{j,l} > \frac{k_{all}}{2}\right]. \tag{F.1}$$

That is, $\hat{m}_j = 1$ if the majority of the $k_{all}$ copies decode to 1, and $\hat{m}_j = 0$ otherwise.

### F.2 Bit Accuracy

Given the ground-truth payload $m \in \{0,1\}^L$ and the recovered payload $\hat{m} \in \{0,1\}^L$ (after majority voting), the Bit Accuracy is defined as:

$$\text{BA} = \frac{1}{L} \sum_{j=1}^{L} \mathbb{1}[\hat{m}_j = m_j]. \tag{F.2}$$

Bit Accuracy of 1.0 indicates perfect recovery, while 0.5 corresponds to random guessing. BA measures the fraction of correctly recovered bits, not the fraction of samples in which a watermark is detected. A sample need not recover every bit to pass the watermark-presence threshold. In our experiments, we report BA separately for the video modality ($\text{BA}_v$) and audio modality ($\text{BA}_a$).

**Sample-level watermark detection.** For each modality, detection rate is the percentage of authentic samples whose recovered BA exceeds the frozen threshold $\tau_{acc}$. Joint watermark detection requires both modality BA gates to pass, before applying the binding test in Eq. (10). Thus extraction BA, watermark detection rate, and final authentication TPR are distinct quantities. Tables 1 and 3 reserve separate columns for detection rates; these require per-sample threshold decisions and cannot be inferred from mean BA. For non-mAVE baselines, use each method's fixed watermark-presence detector. A reported 100% detection rate would mean every evaluated authentic sample passed the relevant detector, not perfect bit recovery or a guarantee on unseen samples.

**Empirical false acceptance.** Let $a_-$ be accepted swapped pairs and $n_-$ all attempted swapped pairs for a given construction; the observed FPR is $100a_-/n_-$ percent. In Table 3, each method is evaluated on 1,000 independently sampled swaps. Each cross-prompt, same-prompt, and similarity-selected group in Table 5 also contains 1,000 independent swap trials, one per target video. mAVE accepts none of the swapped pairs in each evaluation, giving $a_-/n_- = 0/1{,}000$. For zero acceptances in $n_-$ independent trials, the one-sided upper confidence limit at level $1-\alpha$ is $1-\alpha^{1/n_-}$. With $\alpha = 0.05$ and $n_- = 1{,}000$, this is approximately $0.002991$, or $0.30\%$. The confidence limit applies separately to each test group. Authentic-pair counts separately determine TPR/FNR.

### F.3 Video Quality Metrics

We adopt the five video quality metrics from VBench (Huang et al., 2024). All scores are normalized to $[0, 1]$; higher is better. Further details are available in the original VBench paper.

**Subject Consistency.** Subject Consistency measures the temporal coherence of the main subject across frames by computing the DINO (Caron et al., 2021) feature similarity:

$$S_{\text{subject}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}\big(\langle d_1 \cdot d_t\rangle + \langle d_{t-1} \cdot d_t\rangle\big), \tag{F.3}$$

where $d_i$ is the DINO image feature of the $i$-th frame (normalized to unit length) and $\langle\cdot\rangle$ denotes the dot product (cosine similarity).

**Background Consistency.** Background Consistency evaluates the temporal consistency of background scenes by computing CLIP (Radford et al., 2021) feature similarity across frames:

$$S_{\text{background}} = \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{2}\big(\langle c_1 \cdot c_t\rangle + \langle c_{t-1} \cdot c_t\rangle\big), \tag{F.4}$$

where $c_i$ represents the CLIP image feature of the $i$-th frame, normalized to unit length.

**Motion Smoothness.** Motion Smoothness is evaluated by the frame-by-frame motion prior to video frame interpolation models (Li et al., 2023). Specifically, given a generated video consisting of frames $[f_0, f_1, f_2, f_3, f_4, \ldots, f_{2n-2}, f_{2n-1}, f_{2n}]$, the odd-numbered frames are manually dropped to obtain a lower-frame-rate video $[f_0, f_2, f_4, \ldots, f_{2n-2}, f_{2n}]$, and video frame interpolation is used to infer the dropped frames $[\hat{f}_1, \hat{f}_3, \ldots, \hat{f}_{2n-1}]$. The Mean Absolute Error (MAE) between the reconstructed frames and the original dropped frames is then computed and normalized into $[0, 1]$, with a larger number implying smoother motion.

**Dynamic Degree.** Dynamic Degree assesses the extent to which models tend to generate non-static videos. RAFT (Teed & Deng, 2020) is used to estimate optical flow strengths between consecutive frames of a generated video. The average of the largest 5% optical flows (capturing the movement of salient objects) is taken as the basis to determine whether the video is static. The final dynamic degree score is calculated by measuring the proportion of non-static videos generated by the model.

**Imaging Quality.** Imaging Quality is measured by the MUSIQ (Ke et al., 2021) image quality predictor trained on the SPAQ dataset, which is capable of handling variable-sized aspect ratios and resolutions. The frame-wise score is linearly normalized to $[0, 1]$ by dividing by 100, and the final score is then calculated by averaging the frame-wise scores across the entire video sequence.

# G LOCAL TEMPORAL PERTURBATIONS

This section evaluates local frame smoothing and adjacent-frame swaps using MarkDiffusion (Pan et al., 2025). The reported results concern these two operations; they do not establish robustness to arbitrary temporal transformations.

## G.1 ATTACK DEFINITIONS

Let $\{f_t\}_{t=1}^{T}$ denote the original video sequence of $T$ frames.

**FrameAverage.** A sliding-window smoothing attack that replaces each frame with the average of its $n$-neighbourhood ($n = 3$):

$$f'_t = \frac{1}{2n+1} \sum_{\tau=t-n}^{t+n} f_\tau, \tag{G.1}$$

where boundary frames are handled via zero-padding. The output has the same length $T$ but with blurred pixel values that suppress high-frequency watermark signals.

**FrameSwap.** A stochastic local permutation that, for each frame $f_t$ ($t \geq 2$), swaps it with its predecessor $f_{t-1}$ with probability $p = 0.25$:

$$(f'_{t-1},\, f'_t) = \begin{cases} (f_t,\, f_{t-1}), & \text{with prob. } p, \\ (f_{t-1},\, f_t), & \text{otherwise.} \end{cases} \tag{G.2}$$

The total frame count remains $T$, but the temporal ordering is locally disrupted.

## G.2 EXPERIMENTAL SETUP

We evaluate three configurations to isolate the effect of the watermarking method and the generation backbone:

1. **VideoShield (ModelScope)**: The original VideoShield algorithm (Hu et al., 2025a) running on ModelScope (Wang et al., 2023).
2. **VideoShield (LTX-2 T2AV)**: VideoShield ported to the same LTX-2 T2AV backbone used by mAVE.
3. **mAVE (LTX-2 T2AV)**: Our full method on LTX-2 T2AV.

Table G.1: **Temporal Attack Results (Video Bit Accuracy).** FrameAverage and FrameSwap evaluated on three configurations. Entries report video bit accuracy, not sample-level detection or pair-authentication rates.

| Configuration | FrameAvg | FrameSwap |
|---|---|---|
| VideoShield (ModelScope) | 0.995 | 0.732 |
| VideoShield (LTX-2 T2AV) | 0.948 | 0.699 |
| mAVE (LTX-2 T2AV) | 0.927 | 0.710 |

## G.3 ANALYSIS

Across the three configurations, FrameAverage retains higher BA than FrameSwap, consistent with the importance of temporal alignment for extraction.

The evaluated schemes recover payloads from indexed latent grids. For mAVE, $k_t = 1$ provides no temporal repetition, and randomization masks are assigned to coordinates in $\mathcal{J}_R$. Temporal disturbances can therefore misalign recovered signs with the expected masks. This explains the role of alignment without assuming that one decoded video frame corresponds to one latent time slice.

**FrameAverage.** The sliding-window average blurs pixel values but *does not alter the absolute frame index*: frame $t$ remains at position $t$ after the attack. Because the spatial repetition factors

$k_h = 4$, $k_w = 4$ provide strong redundancy, the per-coordinate numerical perturbation is easily corrected by majority voting and zero-thresholding. Hence, Bit Accuracy remains above 0.92.

**FrameSwap.** Swapping adjacent frames with probability $p = 0.25$ introduces local ordering changes. The measured video BA is 0.732, 0.699, and 0.710 for VideoShield (ModelScope), VideoShield (LTX-2), and mAVE (LTX-2), respectively. Adjacent swap decisions can interact, and joint generation/inversion need not preserve a one-to-one relation between video frames and latent time slices; a mixture based only on the nominal swap probability is therefore not a quantitative prediction of these rates.

# H Evaluation Prompt List

We curate 250 text prompts from VBench (Huang et al., 2024), evenly distributed across 8 semantic categories. Below we list representative examples from each category.

Table H.1: **Representative Evaluation Prompts by Category.** Selected examples from each of the 8 VBench categories.

| Category | Example Prompts |
|---|---|
| Animal | "A panda standing on a surfboard in the ocean in sunset"; "A jellyfish floating through the ocean, with bioluminescent tentacles" |
| Architecture | 'A 3D model of a 1800s victorian house"; "Asian garden and medieval castle" |
| Food | "Slow motion cropped closeup of roasted coffee beans falling into an empty bowl"; "A serving of pumpkin dish in a plate" |
| Human | "Vincent van Gogh is painting in the room"; "An astronaut feeding ducks on a sunny afternoon, reflection from the water" |
| Lifestyle | "This is how I do makeup in the morning"; "Home deco with lighted" |
| Plant | "Yellow flowers swing in the wind"; "An elephant spraying itself with water using its trunk to cool down" |
| Scenery | "Underwater coral reef"; "Train station platform" |
| Vehicles | "A car stuck in traffic during rush hour"; "A motorcycle cruising along a coastal highway" |

Each prompt is designed to involve both rich visual dynamics and plausible ambient or event-driven audio, ensuring meaningful evaluation of the audio-visual binding capability.

## H.1 Regeneration Attack Protocol

The regeneration row in Table 4 evaluates a content-preserving reconstruction attack, following the diffusion-regeneration setting studied for image watermarks (An et al., 2024). Apply an image diffusion model frame by frame: encode each watermarked frame into its latent space, add Gaussian noise at a low schedule level, reverse-denoise, and decode. Use an empty attack prompt, with no instruction to replace objects or change style. Retain frame order, frame rate, and duration. This is frame-wise purification, not joint video denoising. The noise level controls corruption strength; the number of reverse steps is a separate setting and is not assumed to be one.

Reconstruct the original audio through EnCodec (Défossez et al., 2023), preserving its duration and alignment, then remux the two streams. EnCodec is neural codec reconstruction rather than semantic audio editing. Run authentication on the reconstructed pair using the original generation prompt and session record, with the same frozen detection thresholds as the other attacks. Apply the reconstruction to both authentic and negative pairs when computing authentication rates.

# I Extended Related Work

## I.1 From Single-Modality to Joint Audio-Visual Generation

**Diffusion Models and Flow Matching.** Generative AI has been revolutionized by Denoising Diffusion Probabilistic Models (Ho et al., 2020) and related image, audio, and video diffusion models (Rombach et al., 2022; Saharia et al., 2022; Huang et al., 2023; Blattmann et al., 2023). Related approaches include Rectified Flow (Liu et al., 2022), image-to-image diffusion bridges (Liu et al., 2023), and consistency models (Song et al., 2023), which provide distinct routes to transport modeling, conditional restoration, and fast generation.

**Joint Audio-Visual Generation.** Historically, audio-visual synthesis relied on cascaded pipelines—sequentially generating video-from-audio or audio-from-video. Such pipelines can model cross-modal dependencies through conditioning, but do not jointly denoise the two modalities. To address these limitations, the field is shifting towards Joint Audio-Visual Generation. State-of-the-art open-source frameworks such as LTX-2 (HaCohen et al., 2026) and MOVA (SII-OpenMOSS Team et al., 2026) process video and audio latents within a joint generative architecture. In these models, a unified denoising loop facilitates bidirectional information exchange via cross-attention, supporting temporal alignment. This shift from independent to joint generation provides the structural basis for our proposed entangled watermarking strategy.

**Rectified Flow and Invertibility.** These native bimodal models build upon Rectified Flow (Liu et al., 2022), which defines a straight interpolation $\mathbf{z}_t = t\mathbf{x}_1 + (1-t)\mathbf{z}_0$ between data $\mathbf{x}_1$ and noise $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As with deterministic DDIM sampling (Song et al., 2021), the associated ODE permits numerical inversion $\mathbf{z}_0 = \text{ODE-Solve}(\mathbf{x}_1, 1{\to}0, v_\theta)$. Learned trajectories need not be exactly straight, and numerical inversion remains approximate. Our solver experiments measure how this approximation affects watermark recovery.

**Deployment and verification protocols.** Gaussian Shading++ (Yang et al., 2025) encodes the random seed for pseudorandomization through a second channel based on pseudorandom error-correcting codes. Its soft-decision decoder addresses generation/inversion distortions, and public-key signatures support third-party verification of image watermarks. mAVE instead uses a public record index to retrieve session-specific secrets and verify the relation between two jointly generated modalities. An external requester supplies media and the original prompt or a request reference; the provider performs inversion and verification without releasing its session keys. Thus Gaussian Shading++ emphasizes image-level watermark deployment and public verification, while mAVE establishes a service-mediated test of shared AV generation history. These design choices address different verification workflows and protected objects.

## I.2 Video Watermarking

As generative capabilities expand, distinguishing synthetic content from authentic footage has become critical. Existing video watermarking schemes broadly fall into two categories: post-processing and in-processing.

**Post-processing Schemes.** These methods operate on the decoded video pixels, embedding imperceptible signals via additive perturbations or multiplicative masks (Luo et al., 2025; Fernandez et al., 2024). While effective for legacy copyright protection, they face the *invisibility-robustness trade-off*: increasing robustness against re-encoding often introduces visible artifacts. Furthermore, applying watermarks after generation does not fundamentally verify the provenance of the generative process itself.

**In-processing Schemes.** To integrate protection intrinsically, recent works leverage the generative process itself. The paradigm was established by Gaussian Shading (Yang et al., 2024), which proposed a provable watermarking scheme for image diffusion by coupling the generated output with a fixed initial noise state. Expanding into the temporal domain, methods like VideoShield (Hu et al., 2025a) and the recent VideoMark (Hu et al., 2025b) generalize these principles to video diffusion, embedding ownership information directly into spatiotemporal latents. The rapid evolution of these algorithms has also led to the development of integrated evaluation libraries such as MarkDiffusion (Pan et al., 2025), which benchmark the performance of various generative watermarking strategies. However, despite these advancements, current in-processing schemes treat the video la-

tent as an independent variable, disregarding the potential for cross-modal cryptographic binding in bimodal contexts. **Trainable video embedding and extraction.** LVMark (Jang et al., 2024) jointly optimizes a latent video decoder and a watermark decoder. Importance-based modulation selects decoder layers for embedding, while its recovery network combines low-frequency 3D-wavelet information with video color features to exploit temporal consistency. VIDSIG (Huang et al., 2025) partially fine-tunes the latent decoder: perturbation-aware suppression freezes perceptually sensitive layers, and a temporal alignment module guides coherent video generation during training. Both integrate learned watermark behavior into video generation. FlowMark (Asnani et al., 2026) instead learns object-mask guidance and region-aware media embedding, with noise-augmented training for robust recovery and temporally coherent watermark placement. Its adaptation takes place in trained watermarking components rather than a joint AV session-binding construction.

**Recent video-specific designs.** SIGMark (Zhu et al., 2026) combines global frame-wise pseudorandom coding keys with segment group-ordering to support scalable blind extraction and temporal recovery in modern video diffusion models. SKeDA (Yang et al., 2026) uses shuffle-key distribution-preserving sampling and differential attention to improve extraction under temporal distortions. Alongside VideoMark, these methods show that training-free embedding and distribution-preserving sampling are not unique to mAVE. Our distinction is the authenticated relation between audio and video from one native joint-generation session, rather than an inability of those methods to watermark modern video generators. These are conceptual comparisons, not additional empirical baselines.

**Indistinguishability and removal are separate properties.** Lee et al. (2025) study removal attacks exploiting boundary information in latent watermark schemes. Their results show why a distributional indistinguishability argument alone does not establish resistance to adversarial removal. In mAVE, public routing positions and the secret authentication payload have different roles: changing the routing signal may disrupt record recovery, while recovering an index does not establish a valid binding. We make no inference from public-index recovery or randomized-region prior preservation to resistance against boundary-guided or oracle-guided removal.

### I.3 Audio Watermarking and The Binding Gap

Parallel to the visual domain, audio watermarking is essential for safeguarding the auditory component of bimodal content.

**Deep Generative Audio Watermarking.** Learned audio watermarking methods include WavMark (Chen et al., 2023), which uses an invertible neural network for watermark embedding and extraction. A prominent example is AudioSeal (San Roman et al., 2024), which employs a localized detection mechanism capable of identifying AI-generated segments at the sample level of $1/16,000$s. This localization capability makes it highly effective against splicing and deepfake manipulation in long-form audio.

**The Binding Gap.** Despite the maturity of unimodal watermarking, a critical gap remains: *Binding Security*. Current approaches treat audio and video as separate entities, verifying $Video_{wm} \vee Audio_{wm}$ independently. This leaves joint audio-visual generation models vulnerable to Swap Attacks, where an adversary pairs a watermarked video with a mismatched or malicious audio track. mAVE addresses this gap by linking the modalities in their joint initialization space.

**Cross-Modal Consistency versus Provenance.** SyncNet (Chung & Zisserman, 2016) assesses lip–speech synchronization, ImageBind (Girdhar et al., 2023) supplies shared multimodal representations, and CLAP (Wu et al., 2023) measures audio–text compatibility. These tools provide useful proxy consistency checks, but their scores do not encode generation-session identity. Semantically compatible audio and video can still originate from different requests. We therefore evaluate consistency-based combinations as proxy baselines alongside unimodal watermarking, distinguishing perceptual compatibility from mAVE's same-session authentication objective.

**Relation to cross-modal media watermarking.** $V^2A$-Mark (Zhang et al., 2024) shares copyright information between video and audio and combines their decoded bits, while Kim et al. (2025) embed an audio spectrogram into visual-frame representations and recover it for feature-based tamper localization. These are learned media-domain embedding/recovery pipelines. Our architectural advantage concerns native integration with a joint DiT's initialization and inversion, rather than an

inability of those methods to post-process DiT outputs. mAVE complements these copyright and integrity protections with generation-session authentication, addressing the risk that separately valid media are misattributed as an original joint output.

### I.4 Fingerprints, Joint AV Signals, and Provenance Infrastructure

**Content-based fingerprinting.** Landmark-based audio identification (Wang, 2003) matches query recordings to a reference database using robust signal descriptors. Video hashes such as TMK+PDQF (Facebook, 2019) similarly support video similarity matching. Registering both tracks under one database entry can associate a known AV pair. The evidence comes from correspondence to reference content and the matching rule; mAVE instead embeds a keyed relation during joint generation and tests it against a session record. Fingerprinting is therefore useful for reference retrieval, while mAVE supplies a generation-time session signal. Both complement semantic compatibility with reference-based evidence.

**Joint AV evidence beyond watermarks.** Synchronization and shared-embedding methods quantify perceptual correspondence; AVFF (Oorloff et al., 2024) learns cross-modal features for classifying manipulated video. Such passive forensic evidence can flag suspicious audio-video combinations without generation-time embedding. mAVE supplies positive evidence of membership in a provider's generation session, including when both modalities are synthetic, complementing consistency and manipulation scores. A provenance application can present these signals together while distinguishing content resemblance, manipulation evidence, and authenticated session association.

**Cryptographic manifests.** C2PA (C2PA, 2026) carries signed provenance assertions and cryptographic asset bindings. Its soft-binding mechanism uses fingerprints or watermarks to associate content with manifests, including discovery after embedded metadata is lost. C2PA can describe joint AV assets and their ingredients; mAVE contributes an embedded same-session test.

**Proposed integration.** A provider could associate each mAVE session record with a signed manifest identifier. Given the original prompt or request context (Section 4.3), the verifier recovers $I$, checks AV binding, and resolves the issuer-scoped manifest reference. The client validates the manifest signature and applicable asset bindings separately. For transformed renditions, an updated manifest can reference the original as an ingredient and bind the current asset. The mAVE match supplies the generation-session association. Session secrets remain server-side. A standardized soft-binding integration would additionally define the algorithm identifier and resolution interface. This workflow combines media-carried session authentication with signed provenance history.

### I.5 Training and Verification Objectives

Table I.1 separates where a watermark is introduced from what its verifier establishes. Learned video methods develop robust video signatures or recoverable messages; cross-modal hiding schemes additionally recover media content or localize manipulation. mAVE uses fixed generator weights and a server-held session reference to authenticate an audio-video pairing. These objectives are complementary: a copyright message, a tamper map, and a shared-generation-session decision provide different information about the same media.

Table I.1: **Architectural and Verification-Objective Comparison.** Training describes watermark-specific optimization, beyond the pretrained media generator.

| Method | Embedding and training | Verification objective |
|---|---|---|
| LVMark (Jang et al., 2024) | Learned latent-decoder modulation and watermark decoder | Video watermark-message recovery for attribution |
| VIDSIG (Huang et al., 2025) | Partial latent-decoder fine-tuning with temporal alignment | Video signature detection and recovery |
| FlowMark (Asnani et al., 2026) | Trainable mask-guided media embedding | Video message recovery and integrity protection |
| $V^2$A-Mark (Zhang et al., 2024) | Learned video/audio watermarking modules | Copyright recovery and manipulation localization |
| Kim et al. (2025) | Trainable invertible cross-modal hiding | Audio recovery and audiovisual tamper localization |
| mAVE | Joint latent initialization; fixed generator weights; no watermark-specific training | Same-session AV pair authentication via joint inversion and a server record |

# J Session-Binding Controls and Swap Construction

This section specifies the additional evaluations in Tables 3 and 5.

## J.1 Shared Session-ID Control

For each generation session, sample a random tag $s \in \{0,1\}^N$ with $N = 128$ in the default comparison and store it with the session record for ideal-grid reconstruction. Embed the same $s$ into both modalities using the same latent carriers, repetition budget, inverse sampling, and modality-specific randomization as mAVE. The remaining payload allocation and backbone are held fixed. The tag is session-specific and is not the vendor-wide ownership message or the public record index $I$. The control has no functional dependency of its audio tag on the video digest.

The verifier performs the same prompt-conditioned inversion and session lookup, recovers $\hat{s}_v, \hat{s}_a$, and computes $S_{ID} = N^{-1} \sum_j \mathbb{1}[\hat{s}_v[j] = \hat{s}_a[j]]$. It accepts when both modality-validity BA gates pass and $S_{ID} > \tau_{ID}$. The agreement threshold is selected on the clean validation split and then frozen. The original aggregate evaluation populates the Shared Session-ID row in Table 3; the separately constructed hard-swap evaluation populates Table 5.

## J.2 Mechanism Ablations

**No binding gate.** Reuse the full mAVE watermarked samples and recovered scores. Change only the final verdict to $(\mathrm{BA}_v > \tau_{acc}) \wedge (\mathrm{BA}_a > \tau_{acc})$. This isolates the contribution of the explicit binding-score gate without changing the embedding or generated content.

**No hash dependency or binding gate.** Generate a new matched sample set. Replace the digest-dependent audio binding bits with pseudorandom bits derived from the audio session sub-key, with no dependence on $B_v$ or its digest. Retain session-key derivation, carriers, payload length, repetition, prompts, seeds, backbone, and solver settings. Verify the recovered modality bits against their corresponding ideal grids using the two BA gates. This tests whether session-specific watermark validity alone accounts for swap rejection. Unlike removing only the gate, it also removes the hash dependency at embedding time. It is a combined ablation; SR-Rand below isolates the binding-string construction while retaining the gate.

## J.3 Matched Server-Reference Controls

The decoded-tag agreement control compares two noisy estimates, whereas mAVE compares recovered binding bits with a server-reconstructed target. The following controls distinguish the reference choice from the construction of the binding string. Both use the same record-retrieval protocol, public-index layout, total payload, repetition factors, backbone, prompt/seed assignments, and joint-

inversion settings as mAVE. Tables 3 and 5 report their measured rates separately from decoded-tag agreement.

**SR-Rand: random binding with the same gate.** For each session, independently sample and store a uniform $N$-bit string $r$ unrelated to the video grid. Keep the mAVE video grid unchanged. At the same $N$ audio binding positions, replace the selected video-digest bits by $r$, retaining all other audio bits, randomization, and repetition. Regenerate the output pair with this modified embedding. Let $\hat{r}_a$ be the recovered audio binding string. The verifier uses

$$S_{rand} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\hat{r}_a[j] = r[j]],$$

and accepts only if both original BA gates and $S_{rand} > \tau_{rand}$ pass, where $\tau_{rand}$ is the frozen calibration threshold for this score. Ideal audio grids are rebuilt using the stored $r$. This replaces the binding target's origin while retaining its length, positions, noise-free reference, and dedicated gate. Unlike the earlier combined ablation, it does not remove the gate.

**SR-ID: shared tag with noise-free references.** Use the existing Shared Session-ID embedding: a uniform $N$-bit session tag $s$ occupies dedicated slots in each modality, replacing the same number of payload bits rather than adding capacity. Keep the mAVE total per-modality bit/repetition budget. Query the video-selected session record and compare both recovered tags $\hat{s}_v, \hat{s}_a$ to that record's stored $s$, defining

$$S_{ID,v} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\hat{s}_v[j] = s[j]], \qquad S_{ID,a} = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\hat{s}_a[j] = s[j]].$$

Retain the two modality-validity BA gates and additionally require $\min(S_{ID,v}, S_{ID,a}) > \tau_{SR}$, where $\tau_{SR}$ is selected by the common calibration rule. The audio is not checked against its own donor-session record. On an identical Shared Session-ID sample set, this comparison can reuse the recovered tags to isolate the reference change; newly generated samples must not be substituted silently. SR-ID is a stronger session-tag baseline, while SR-Rand is the direct control for mAVE's binding-string construction.

**Calibration and evaluation.** Select thresholds using the same held-out clean-validation procedure and a common target recall specified before testing, then freeze them for all swap groups. Report the actual test TPR alongside FPR; do not adjust thresholds to force equal test recall. Retain the modality BA calibration and report its pass rate separately. Use identical session-pair identities and candidate pools across methods, invert each spliced output pair anew with the target video's original prompt, and include every attempted query in the denominator. Report binding-score distributions for authentic pairs and each swap group. Where the clean-recall target differs from the original operating point, recompute all methods at the common target rather than transplanting earlier rates.

### J.4 Swap Groups

**Cross-prompt random.** For a target video session $A$, randomly choose audio from another session $B$ with a different prompt.

**Same-prompt, different-session.** Pair video and audio from independent requests sharing the same prompt but using different session secrets and seeds. The four-seed-per-prompt setup supplies candidate pairs; the matched pair $A = B$ is excluded.

**Similarity-selected.** For each target video, form a fixed pool of audio from other sessions and select the candidate with the highest ImageBind audio–video similarity. Use the same candidate-session identities across methods and do not select by mAVE's secret binding score. Candidate-pool size: 20.

All groups splice the actual output media and run the full detector on the resulting pair; they do not combine watermark bits extracted from the original matched pairs. The video session's original prompt is provided for the joint inversion. The same original authentic session set supplies the TPR comparison. The controls use the shared evaluation setup, with thresholds frozen before the swap tests.

# K NOTATION GLOSSARY

Subscripts $v$ and $a$ denote video and audio. A hat denotes a decoded estimate; a tilde denotes a value recovered by inversion or an attacked sample, as specified locally. Superscript "ideal" denotes a noise-free reconstruction from a session record. Table K.1 collects the method, detector, proof, and evaluation notation. Some appendix formulas use local conventions; these are distinguished explicitly below.

Table K.1: **Symbols and Definitions.**

| Symbol | Meaning |
|---|---|
| **Sessions, keys, and operators** | |
| $A, B$ | Distinct generation sessions, each corresponding to one inference request. |
| $I, \hat{I}$ | Public session-record index and its decoded estimate; repeated at fixed, unencrypted video positions. |
| $m, K_{priv}$ | Server-held session secret, with $K_{priv} = m$; never embedded. |
| $L_I, L_m$ | Bit lengths of the public index and server secret. |
| $\mathcal{J}_I, \mathcal{J}_I(b), r_I$ | Fixed public index coordinates, copies of index bit $b$, and copies per index bit. |
| $\mathcal{J}_R$ | Complementary randomized coordinates; excludes the public index. |
| $d_v, d_a, \eta_v, \eta_{AV}$ | Video/audio scalar latent counts and public-coordinate fractions within video and joint latents. |
| $a_-, n_-, \alpha$ | Accepted and attempted swapped-pair counts, and tail probability in the independent-trial confidence statement. |
| $n_I, R_I$ | Number of attempted samples and exact full-index recovery rate. |
| $\Pr(\cdot), P(\cdot)$ | Probability of an event; distinct from the prompt string $P$. |
| $P, E_P$ | Original generation prompt and its normalized string $E_P = \mathrm{Norm}(P)$. |
| Norm | Lowercase conversion and whitespace trimming. |
| $K_{sess}$ | Session key derived from the secret and normalized prompt. |
| $K_{grid}, K_{mask}, K_{perm}$ | Purpose-separated subkeys for grid construction, per-copy masking, and block permutation. |
| $\mathrm{KDF}, \mathrm{label}_x$ | Purpose-separated key derivation and distinct application labels. |
| $\lambda_{key}, \mathrm{Prefix}_{\lambda_{key}}$ | Fixed secret-hash prefix length and its truncation operator, independent of binding length $N$. |
| SHA-256 | Cryptographic hash used for key derivation and video-grid digests. |
| HMAC | Hash-based message authentication code, used to derive modality-specific pseudorandom base bits. |
| $\|, \oplus$ | Bit-string concatenation and bitwise exclusive OR. |
| $\mathrm{seed}_{shared}$ | Shared 64-bit seed extracted from the first eight bytes of the secret hash. |
| $S_{shared}$ | Shared pseudorandom base-bit stream seeded from the secret. |
| $\mathcal{K}, \mathcal{S}$ | Masking bit keystream and its ChaCha20 byte-stream representation. |
| **Latent spaces, bit grids, and embedding** | |
| $\mathcal{Z}_v, \mathcal{Z}_a, \mathcal{Z}$ | Initial video/audio latent spaces and their joint product space. |
| $\mathcal{M}$ | Set of initial latent pairs satisfying the session-binding predicate. |
| $\mathbf{z}_v, \mathbf{z}_a$ | Video and audio latent tensors; $\mathbf{z}$ denotes a generic or joint latent. |
| $C, T, H, W$ | Latent tensor sizes along channels, time, height, and width. In temporal-attack formulas, $T$ instead counts video frames. |
| $B_v, B_a, B, L$ | Video/audio bit grids, a generic grid, and grid length. |
| $B_v^{ideal}, H_{ideal}$ | Noise-free video grid reconstructed from the session key, and its cryptographic digest used as the binding reference. |
| $h_v$ | Digest $\mathrm{SHA\text{-}256}(B_v)$ inserted into the audio grid at generation time. |
| $\mathbf{t}, c_0$ | Synchronization template and its designated video channel. Bold $\mathbf{t}$ is distinct from scalar flow time $t$. |
| $\mathcal{I}_{inject}$ | Operator inserting template and index bits into designated video-grid slots. |
| $\mathcal{I}_{time}, \mathcal{I}_{bind}$ | Sets of synchronization-template and audio-binding positions. |
| $\phi(i), N$ | Digest-bit position assigned to binding slot $i$, and number $N = \lvert\mathcal{I}_{bind}\rvert$ of independent comparisons. |

**Table K.1 (continued)**

| Symbol | Meaning |
|---|---|
| $\mathrm{Base}_a$ | Audio sub-key-derived base grid at positions not reserved for template or binding bits. |
| $B^{diff}, M_{rand}$ | Repeated bit tensor and sign-bit map: XOR-masked on $\mathcal{J}_R$, unmasked on $\mathcal{J}_I$. |
| $k_c, k_t, k_h, k_w$ | Repetition factors along the four latent axes; $k_{all} = k_c k_t k_h k_w$ is the copy count per bit. |
| $\Pi$ | Randomized inverse-CDF map from bit grids to Gaussian latents. |
| $\Phi, \Phi^{-1}, \mathrm{ppf}$ | Standard normal CDF and its inverse (quantile function). |
| $u_i, u_i^{\delta}, z_i$ | Auxiliary uniform variate, its half-interval clamp, and the resulting latent coordinate; Appendix E specifies the numerical rule. |
| $\mathcal{I}_w, y_i, \pi$ | Watermark-coordinate set, assigned binary value, and coordinate mapping induced by the secret block permutation in Appendix E. In Appendix F, $\pi(j, l)$ locates copy $l$ of payload bit $j$. |
| $\delta$ (sampling) | Uniform-input clipping margin, set to $0.001$ in Appendix E; distinct from inversion cutoff $\delta_t$. |
| **Inversion, authentication, and evaluation** | |
| $\mathbf{x}_v^{(A)}, \mathbf{x}_a^{(B)}$ | Video from session $A$ and audio from session $B$; their splice is $\tilde{\mathbf{x}}$. |
| $\mathbf{x}_{query}$ | Queried pair in the generator's joint data-latent representation, used as the inversion endpoint. |
| $t, \mathbf{z}_t$ | Flow time and the joint latent state at that time. |
| $v_\theta, \theta, c$ | Flow velocity field, generator parameters, and condition from the original prompt. |
| $\delta_t, \tilde{\mathbf{z}}_0$ | Minimum inversion timestep and recovered approximation to the initial latent. |
| $\mathcal{D}, \mathrm{Verify}$ | Final authentication detector and binding predicate; the detector combines modality-validity and binding gates. |
| $S_{bind}, \tau_{bind}$ | Fraction of decoded audio-binding bits matching the ideal digest, and its acceptance threshold. |
| $\mathrm{BA}_v, \mathrm{BA}_a, \tau_{acc}$ | Video/audio bit accuracies against ideal grids, and the modality-validity threshold. |
| $H_0, H_1$ | Same-session pair and mismatched-session pair hypotheses. |
| $\mathrm{TPR}, \mathrm{FNR}$ | True-positive and false-negative rates on authentic pairs. |
| $\mathrm{TNR}, \mathrm{FPR}$ | True-negative and false-positive rates on swapped pairs in the authentication protocol. |
| $\mathrm{Acc}$ | Balanced accuracy $(\mathrm{TPR} + \mathrm{TNR})/2$ in Table 3. |
| $n_s, \mathcal{C}, M$ | Session-pool size, ordered cross-session candidate family, and its size $M = n_s(n_s - 1)$. |
| $Q_{ab}, Q_0, Q_\delta$ | Ordered swap query, fixed baseline swap, and locally perturbed donor-audio query. |
| $\kappa, r_\tau, d_H$ | Minimum matching-bit count, maximum allowed mismatch count, and Hamming distance. |
| $\varepsilon_N(b)$ | Uniform Hamming-ball probability at radius $\min(N, r_\tau + b)$; $b = 0$ gives the exact baseline binding tail. |
| $b, \beta, \beta_g$ | Fixed decoded-bit change budget, per-pair stability-failure probability bound, and pool-wide failure bound. |
| $U, U_0, U_\delta, h$ | Recovered binding vector, baseline/perturbed versions, and selected target digest bits. |
| $\mathcal{G}_b, \mathcal{G}_b^{pool}$ | Events that all allowed perturbations preserve routing and obey the bit-change budget, for one pair or a pool. |
| $K, V, \mathcal{D}_K, q$ | Fixed verification secrets, adversary side information, stateless binary verifier, and verification-query budget. |
| $\mathcal{A}_t, \mathrm{Forge}$ | Adversary next-query rule and event that an admissible mismatched candidate is accepted. |
| $\varepsilon_{single}, Q_t^{\perp}$ | Best allowed no-verification-feedback success bound and candidate $t$ on the all-reject path. |
| $\Delta_{\epsilon_{atk}}, \epsilon_{atk}$ | Allowed perturbation set and total input budget relative to the original donor audio. |

**Table K.1 (continued)**

| Symbol | Meaning |
|---|---|
| $\mathcal{F}, L_{\mathcal{F}}, \eta$ | Complete numerical inversion map, its certified local norm bound, and resulting latent perturbation bound. |
| $G_i, s_j, \gamma_i, e_i, b_{cert}$ | Binding-bit copy group, derandomized sign, vote margin, potentially changing copy count, and certified logical-bit change count. |
| $\nu$ | Adversary-success distinguishing gap supplied by a real-to-ideal reduction, when available. |
| $s, \hat{s}_v, \hat{s}_a$ | Shared Session-ID tag and its decoded video/audio estimates (Appendix J). |
| $r, \hat{r}_a, S_{rand}, \tau_{rand}$ | Stored random binding target, recovered audio target, matching score, and threshold in SR-Rand (Appendix J). |
| $S_{ID,v}, S_{ID,a}, \tau_{SR}$ | Video/audio matches to the stored tag and common score threshold in SR-ID. |
| $S_{ID}, \tau_{ID}$ | Cross-modal tag agreement and its frozen threshold in the Shared Session-ID control. |
| T2AV, I2AV, TI2AV | Text-to-AV, image-to-AV, and text-and-image-to-AV generation. |
| **Theory and appendix-local conventions** | |
| $\mathcal{A}, \mathcal{B}$ | Efficient distinguishing algorithms in the randomization reduction. |
| $I$ in $\mathcal{N}(0, I)$ | Identity covariance matrix, not the public session index. |
| $f, \mathbb{1}[\cdot]$ | Bounded output statistic in Appendix A.1, and indicator function. |
| $l, M, U, Z, s_l$ | Bits per coordinate, sign/bin vector, independent uniform variates, sampled vector, and bin map in Appendix A.1; $M$ there is local, not the swap-pool size. |
| $\Gamma, M_R, R, Y^s, Y^{ref}$ | Public context, real randomized authentication signs, uniform replacement signs, and real/matched numerical pipeline outputs. |
| $\epsilon_{rand}, \epsilon_{sep}, \epsilon_{stream}$ | Joint randomization advantage and its purpose-separation/stream-generator hybrid contributions. |
| $\mathcal{S}_{num}, \mathcal{P}_{num}, \omega$ | Numerical sampler, sampler optionally composed with the generator, and independent implementation coins. |
| $\varphi_d, \gamma_d, \mathrm{TV}, D_{KL}, H$ | Standard Gaussian density/law, total variation, relative entropy, and entropy in natural-log units. |
| $D, d_I, d_R, C(I)$ | Full scalar latent count, public/randomized region counts, and repeated index code. |
| $\ell_\delta, a_\delta, H_\delta, T_\delta$ | Magnitude floor/cap, magnitude clamp, and signed clamp. |
| $\chi_\delta, v_\delta, W_2$ | Clamp mean-square displacement, clamped second moment, and quadratic Wasserstein distance. |
| $P_\delta^{ref}, P_{num}^{ref}, \zeta_{num}$ | Continuous clamped reference, numerical matched reference, and their normalized coupling discrepancy. |
| $F, L_F, X_G$ | Expected output statistic as a function of latent input, its normalized Lipschitz constant, and an output from standard Gaussian initialization. |
| $N, t_i, \epsilon$ (Appendix B) | Solver step count, discrete times, and final time $\epsilon = \delta_t$. This local $N$ is not the binding length. |
| $\alpha_t, \epsilon_\theta$ | DDIM signal-schedule coefficient and predicted noise in Appendix B. |
| $\mu_j, \mathrm{Block}_j$ | Median of repeated latent values for bit $j$, and its repetition block (Appendix C). |
| $m_j, \hat{m}_j, \tilde{b}_{j,l}$ | Generic payload bit, decoded bit, and recovered copy in Appendix F; these $m_j$ do not denote the server secret. |
| $d_i, c_i$ | Normalized DINO and CLIP frame features used for subject/background consistency; $c_i$ is not prompt condition $c$. |
| $f_t, f_t'$ | Original and attacked video frames in Appendix G. |
| $n, p$ | Temporal averaging radius and adjacent-frame swap probability. |