Steganography Beyond Pixels: Reimagining Image Steganography as Cross-Modal Linguistic Communication

Lijing Ren[1], **Denghui Zhang**[2,†]
[1]Institute of Artificial Intelligence, Guangdong Mechanical & Electrical Polytechnic, China
[2]Cyberspace Institute of Advanced Technology, Guangzhou University
Corresponding Author: denghui.zhang@gzhu.edu.cn

Abstract

The rising sophistication of digital surveillance poses hurdles for concealing sensitive data within innocuous communication channels. Conventional image steganography relies on detectable pixel-level perturbations. In this paper, we introduce a novel steganography framework that fundamentally reorients the steganographic containers from the visual domain to the linguistic domain. To seamlessly bridge the gap from raw pixels to discriminative logits, we leverage the reversible latent space of discrete diffusion models to compress high-resolution secret images into lightweight binary payloads. The semantic stability of textual data ensures the integrity of the hidden payload across diverse platforms. Extensive evaluations confirm that this cross-modal approach establishes a superior equilibrium between embedding capacity and statistical undetectability in comparison to existing paradigms.

1 Introduction

In the exponential proliferation era of multimedia content, digital images have become the cornerstone of modern information exchange, necessitating robust mechanisms for their secure transmission. The proliferation of high-resolution visual content has found critical applications including social media communication, autonomous driving and telemedicine (Wen et al., 2023; Deng et al., 2023). The ability to share visual data privately is not merely a technical convenience but a fundamental imperative for operational integrity. While conventional cryptography ensures the confidentiality of message content, it inherently fails to conceal the act of transmission itself. Encrypted traffic manifests as conspicuous, high-entropy statistical anomalies that are easily distinguishable from benign data flows, rendering such communications primary targets for deep learning-based traffic interceptors and governments' suspicion (Sarkar et al., 2020; UK Parliament, 2023). Encryption methods such as symmetric, asymmetric, or even homomorphic cryptography introduce prohibitive computational overhead and latency, which are often incompatible with real-time image processing workflows (Kartit, 2022; Zhang et al., 2024).

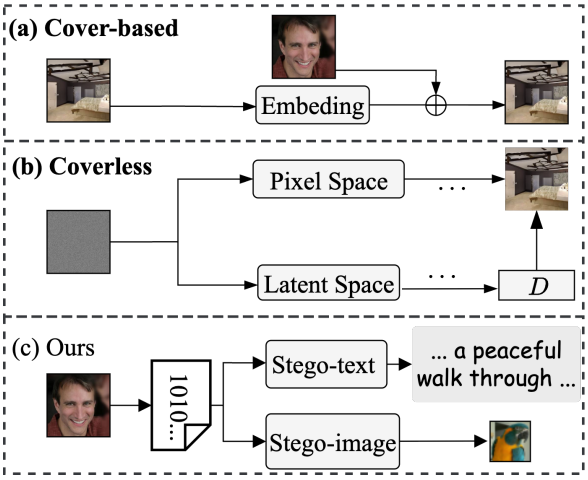

Figure 1: Comparison of steganographic paradigms. (a) Cover-based methods operate by modifying existing carriers, leaving detectable pixel-level artifacts. (b) Coverless methods synthesize stego-images directly from noise but remain confined within the visual domain. (c) Ours introduces a cross-modal paradigm shift: by quantizing high-resolution visual secrets into compact discrete binary sequences, our framework unifies embedding into diverse autoregressive modalities from tokenized text to visuals.

Steganography offers a discreet solution to this challenge by embedding secret data within innocuous cover media, thereby masking the very existence of communication (Zhu et al., 2018). By allowing encoded messages to traverse public networks without triggering suspicion, steganography provides a vital layer of protection against oppressive surveillance where the mere usage of encryption might flag a user. Steganography has relied on modifying the carrier's physical at-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22490–22501

tributes, such as Least Significant Bit (LSB) substitution or frequency-domain manipulation (Filler et al., 2011). These modification-based approaches leave detectable statistical artifacts and pixel-level residues that fall prey to modern deep steganalysis (Boroumand et al., 2019; You et al., 2021). To mitigate the reliance on fixed covers, Generative Image Steganography (GIS) utilizing GANs or VAEs was introduced. These methods frequently suffer from training instability and the disjoint architecture problem, where the lack of precise invertibility between encoder and decoder limits both stego-image quality and embedding capacity. While Invertible Neural Networks (INN) (Zhou et al., 2023; Zhai et al., 2024) attempt to unify these processes, they often prioritize mathematical bijectivity at the expense of generative flexibility, yielding outputs that subtly deviate from the natural image manifold and remain susceptible to advanced forensic detection (Aljarf et al., 2023).

The proliferation of large language models (LLMs) and multimodal AIGC systems has opened novel pathways for secure data concealment across modalities. An emerging frontier in steganography leverages word probabilities (logits) for text hiding (Kirchenbauer et al., 2023; Ding et al., 2023; Wang et al., 2025), which manipulates the sampling process of language models to guarantee provable security. Despite their theoretical elegance, these linguistic methods face a profound modality-capacity disparity when applied to visual data (Jiang et al., 2025). The high-entropy nature of high-resolution images fundamentally exceeds the sparse embedding bandwidth provided by standard text-based schemes, which are confined to short messages or low-resolution thumbnails. Bridging the gap between the information volume of the visual domain and the discrete, low-bandwidth nature of the linguistic domain remains an unsolved challenge.

The recent ascendancy of Diffusion Probabilistic Models (DPMs) provides a potential avenue to resolve this bottleneck by offering superior modeling of complex data distributions (Ho et al., 2020; Song et al., 2022; Liu et al., 2025; Chung et al., 2023; Tumanyan et al., 2023; Schusterbauer et al., 2025). Diffusion-based frameworks like CRoSS (Yu et al., 2023) have demonstrated success. However, they predominantly rely on Denoising Diffusion Implicit Models (DDIM) inversion for reconstruction (Song et al., 2022; Wallace et al., 2023; Wang et al., 2024). However, this inversion process is often inexact. Its fidelity is compromised by accumulating discretization errors and is fundamentally bottlenecked by the aggressive compression of VAEs in Latent Diffusion Models (Rombach et al., 2022). This compression often reduces the spatial dimensions from 512×512 to 64×64, narrowing the embedding space and creating an information bottleneck (Ohayon et al., 2025; Zhao et al., 2025). Techniques like Gaussian Shading (Yang et al., 2024) are limited to payloads of only a few hundred bits, far below the requirements for high-capacity visual transmission.

To address these limitations, we propose VLace (Vision-Language Alignment and Concealed Embedding), a training-free framework that fundamentally reimagines image steganography as cross-modal linguistic communication. While the information density of pixels generally far exceeds the length of AI-generated text, the inherent stochasticity and redundancy of the diffusion process provide a novel medium for steganography. As shown in Figure 1, rather than struggling to hide artifacts within pixel space, we leverage the inherent stochasticity of Stochastic Differential Equation (SDE)-based diffusion models to transcode visual secrets into the linguistic domain. By quantizing the stochastic trajectories of diffusion sampling noise into ultra-compact index sequences, VLace compresses high-resolution visual data into a format compatible with textual transmission. Rather than modifying existing content, these sequences function as control signals during LLM inference, mapping the quantized secret indices to specific tokens via a distribution-aligned sampling strategy to produce coherent cover text. This generated stego-text remains statistically indistinguishable from standard AI-generated content. The main contributions of this work are summarized as follows:

- We introduce a training-free cross-modal steganography paradigm that shifts the transmission medium from pixels to text, circumventing visual-domain steganalysis and censorship.
- We develop an ultra-compact image quantization pipeline based on SDE diffusion, bridging the representational gap between high-volume pixel data and low-bandwidth textual tokens.
- We design a distribution-aligned token embedding mechanism that guides LLM sampling with compressed indices, ensuring statistical undetectability while maintaining the precise

mathematical invertibility required for high-fidelity image reconstruction.

2 Related Work

Cover-based Steganography. Before deep learning, steganography relied on manipulating spatial or frequency domains (Filler et al., 2011). Steganography that embeds data inherently leaves modification traces in the cover image, especially at high capacities. Evolving from these manual constraints, deep Steganography including HiDDeN (Zhu et al., 2018) and SteganoGAN (Zhang et al., 2019), utilizes autoencoder-like structures to embed data, though often at the cost of perfect signal recovery. To mitigate information loss, subsequent research pivots toward Invertible Neural Networks (INNs) like HiNet (Jing et al., 2021) and invertible image rescaling (Xiao et al., 2023), which leverage bijective transformations to ensure exact secret reconstruction. Despite their theoretical elegance, INNs suffer from rigid architectural biases that frequently produce subtle, high-frequency artifacts detectable by deep forensic tools (Boroumand et al., 2019).

Diffusion-based Steganography. Diffusion models have revolutionized steganography by enabling the synthesis of cover images directly from secret-modulated noise, decoupling steganography from fixed covers. CRoSS (Yu et al., 2023) embeds secrets within the Gaussian latent space, while Gaussian Shading (Yang et al., 2024) advances this by ensuring strict adherence to standard distributions for provable security. Despite yielding high visual quality, these pixel-centric approaches are constrained by intrinsic architectural limitations. The reliance on DDIM inversion introduces cumulative discretization errors that degrade reconstruction fidelity (Wallace et al., 2023; Wang et al., 2024), while the finite dimensionality of the latent space severely caps payload capacity, precluding the transmission of full-resolution secret images. This yields embedding rates below 1 bits per pixel (bpp) and only a small number of bits, such as message and watermarking bits. VLace circumvents the latent bottleneck by quantizing the diffusion trajectory into discrete indices.

Linguistic Steganography. Textual steganography aims to conceal secrets within the generation process of natural language. Modern approaches focus on aligning secret bits with the probability distributions (logits) of LLMs (Li et al., 2024b). Chen et al. (Chen et al., 2022) utilize text-to-speech generative models for secure audio steganography. Techniques such as Discop (Ding et al., 2023) utilize distribution copies to maintain statistical consistency, while SparSamp (Wang et al., 2025) enhances embedding efficiency by making the sampling process sparse and pseudo-random. While these methods excel at hiding low-entropy textual data, their application to high-capacity visual data remains largely underexplored. Existing linguistic steganography lacks the mechanisms to handle the high redundancy and semantic complexity of images. The proposed VLace bridges this representational gap by mapping quantized visual features directly onto LLM token logits, facilitating the first training-free, image-to-text steganographic pipeline that achieves high-capacity transmission within innocuous text.

3 Methodology

3.1 Overview

The proposed VLace facilitates a reversible, cross-modal mapping between the high-dimensional visual space $\mathcal{I}$ and the discrete linguistic space $\mathcal{T}$. As illustrated in Figure 2, the pipeline bypasses traditional pixel-domain vulnerabilities by reformulating steganography as a trajectory-guided generative process. The workflow comprises three core modules:

1. Visual-to-Binary Quantization: Distills secret image features into a sequence of noise indices via a diffusion-driven codebook, compressing visual into discrete manifolds.

2. Linguistic Embedding: Aligns the payload with LLM token distributions to generate statistically imperceptible stego-text.

3. Deterministic Reconstruction: Recovers the visual secret by tracing the quantized sampling trajectory using the extracted indices.

3.2 Visual-to-Binary Quantization via Consistent Noise

To bridge the modality gap between high-dimensional images and discrete linguistic tokens, we propose a quantization mechanism that treats diffusion sampling as a sequence of deterministic choices. By reformulating the stochastic sampling process through the lens of consistency, we transform the visual secret into a compact index sequence.

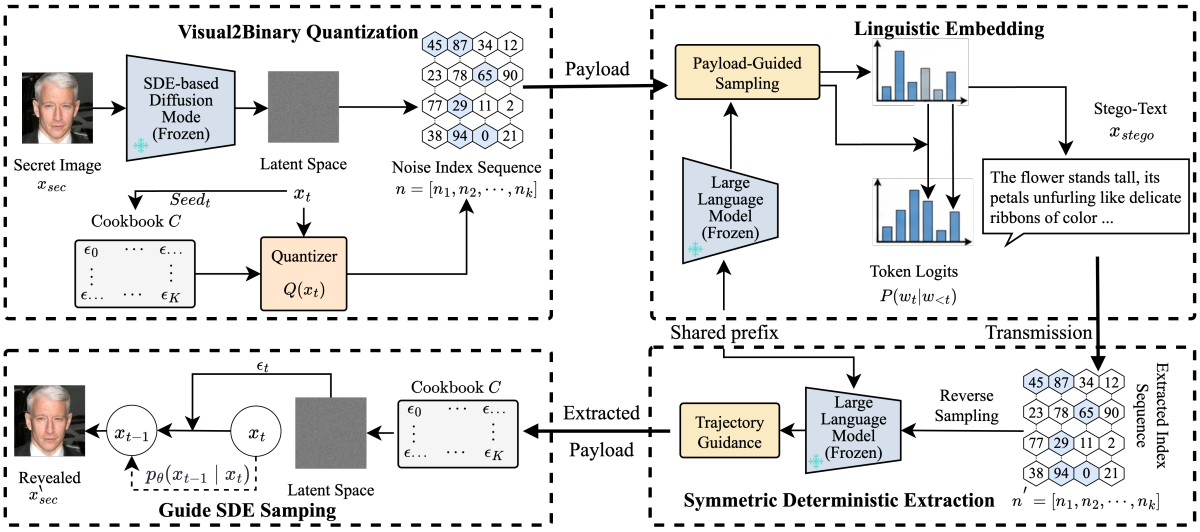

Figure 2: The framework of VLace, reimagining image steganography as cross-modal linguistic communication. Unlike traditional pixel-domain methods, VLace facilitates a Pixels-to-Logits (P2L) transition. The secret visual content is first quantized into a discrete innovation sequence through iterative residual alignment in the diffusion latent space. This sequence is then seamlessly woven into the generative distribution of an LLM. By shifting the carrier from pixels to linguistic tokens, the framework naturally bypasses visual forensics and channel distortions. The bottom-up reconstruction path leverages the generative priors of the diffusion model to ensure high-fidelity synthesis from extracted textual clues.

Denoising Diffusion Consistency. We establish a deterministic mapping between noisy states and the target image by leveraging the consistency property of diffusion models (Huberman-Spiegelglas et al., 2024). Consider a specific denoising trajectory where the stochasticity parameter σ_t is coupled with the noise schedule such that $\sigma_t = \sqrt{1 - \bar{\alpha}_{t-1}}$. Under this condition, the standard DDPM (Ho et al., 2020) sampling step simplifies to isolate the innovation term $\boldsymbol{\epsilon}_t$:

$$\boldsymbol{\epsilon}_t = \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0^t}{\sqrt{1 - \bar{\alpha}_t}} \tag{1}$$

where $\hat{\mathbf{x}}_0^t$ represents the model's estimate of the clean image at timestep t. Equation 1 reveals that the sequence $\{\mathbf{x}_T, \boldsymbol{\epsilon}_{T-1}, \ldots, \boldsymbol{\epsilon}_1\}$ forms a unique latent code for $\mathbf{x}_0$. Crucially, $\boldsymbol{\epsilon}_t$ provides the precise "direction" required to maintain self-consistency across the sampling trajectory, creating a deterministic link between any noisy state $\mathbf{x}_t$ and the visual secret $\mathbf{x}_0$.

Deterministic Codebook Generation. To discretize this continuous innovation space, we construct a time-dependent Gaussian codebook. For each timestep t, a codebook $\mathcal{C}_t^{S_t}$ is generated using a seed-based deterministic sampler, where the seed $S_t = \text{Hash}(S_0 \oplus t)$ is derived from a pre-shared cryptographic key S_0. Each codebook comprises $K = 2^k$ distinct patterns:

$$\mathcal{C}_t^{S_t} = \{\mathbf{c}_j^{(t)} \mid \mathbf{c}_j^{(t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\ j \in [0, 2^k - 1]\} \tag{2}$$

where $\mathbf{c}_j^{(t)} \in \mathbb{R}^d$ matches the latent dimensionality. This mapping ensures the receiver can replicate the search space without transmitting the codebooks, maximizing relative payload capacity.

Residual-Driven Noise Quantization. We propose an alignment-based quantization strategy prioritizing visual semantic preservation. At each encoding step t, the optimal codebook index n_t is selected to maximize the directional alignment with the reconstruction residual:

$$n_t = \arg \max_{n \in \{0, \ldots, K-1\}} \langle \mathbf{c}_n^{(t)}, \mathbf{x}_0 - \hat{\mathbf{x}}_0^t \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. This objective ensures the selected discrete noise term provides the most effective "correction" toward the target $\mathbf{x}_0$. The quantization proceeds iteratively: at each step t, we replace the continuous $\boldsymbol{\epsilon}_t$ with the selected codebook entry $\hat{\boldsymbol{\epsilon}}_t = \mathbf{c}_{n_t}^{(t)}$ to compute $\mathbf{x}_{t-1}$, converting the visual secret into a discrete sequence $\mathcal{N} = \{n_T, \ldots, n_1\}$. Early indices establish global structure, while late indices refine high-frequency textures.

Patch-wise Quantization for High-Resolution Scalability. Scaling the proposed quantization

mechanism to high-resolution images (e.g., 512×512) presents a critical computational bottleneck. The standard global quantization approach requires an exponentially large codebook (e.g., 2^{20}) to capture the diverse structural details of the entire latent space, which becomes computationally prohibitive. To reconcile the trade-off between reconstruction fidelity and computational efficiency, we introduce a patch-wise latent quantization strategy, inspired by the localized processing in Vision Transformers (ViT) (Dosovitskiy et al., 2020).

Instead of treating the latent representation $\mathbf{z}_t \in \mathbb{R}^{C\times H\times W}$ as a monolithic entity, we partition it into a grid of non-overlapping patches $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, where each patch $\mathbf{p}_i$ has dimensions $C \times h \times w$. The quantization process is then localized: for each patch $\mathbf{p}_i$ at timestep t, we generate a lightweight local codebook $\mathcal{C}_{t,i}$ of size 2^b (where $b \ll 20$, e.g., $b = 8$). The optimal noise index $n_{t,i}$ is determined by maximizing the cosine similarity between the local codebook entries and the patch-specific reconstruction residual:

$$n_{t,i} = \arg \max_{k\in\{0,\ldots,2^b-1\}} \langle \mathbf{c}_k^{(t,i)}, \mathbf{p}_i - \hat{\mathbf{p}}_{0,i}^t \rangle \quad (4)$$

This granular approach decomposes the high-dimensional global optimization problem into multiple tractable sub-problems. It allows the model to capture fine-grained texture distributions using smaller, efficient codebooks, ensuring high-fidelity reconstruction even with limited sampling steps, while reducing the computational complexity from exponential to linear with respect to the number of patches.

3.3 Linguistic Steganography via Distribution Alignment

To achieve stealthy transmission, the binary payload B derived from quantization is embedded into natural language. We employ a distribution-aligned sampling strategy to ensure the stego-text remains indistinguishable from native LLM output.

At each generation step t, the LLM provides a conditional probability distribution $P^{(t)} = P(w_t \mid w_{<t})$ over the vocabulary $\mathcal{V}$. To embed the payload bit $b \in B$, we partition the vocabulary into two disjoint subsets $\mathcal{T}_0$ and $\mathcal{T}_1$ based on a shared secret key (PRNG state). The partitioning is performed such that the cumulative probability mass of each subset matches the target distribution of bits in the payload (typically $P(b = 0) \approx P(b = 1) \approx 0.5$):

$$\sum_{w\in\mathcal{T}_0} P(w \mid w_{<t}) \approx \sum_{w\in\mathcal{T}_1} P(w \mid w_{<t}) \approx 0.5 \quad (5)$$

The next token $\hat{w}_t$ is then sampled exclusively from the subset $\mathcal{T}_b$ corresponding to the current payload bit, following the renormalized distribution:

$$P(w) = \begin{cases} \frac{P(w|w_{<t})}{\sum_{w'\in\mathcal{T}_b} P(w'|w_{<t})}, & \text{if } w \in \mathcal{T}_b \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

To further enhance capacity, we implement a Dynamic Huffman-Coding mechanism. By constructing a prefix tree over the logits at each step, we can embed multiple bits ($L \geq 1$) per token while maintaining $D_{KL}(P_{natural} \| P_{stego}) = 0$. This theoretical guarantee ensures that the stego-text is immune to linguistic steganalysis, as the sampling process remains perfectly consistent with the model's learned distribution.

Algorithm 1: VLace Encoding Pipeline

Input: Secret $\mathbf{x}$, Diffusion ϵ_θ, LLM P_ϕ, Key K, Steps T
Output: Stego-text $\mathbf{T}$

$\mathbf{z}_T \leftarrow \text{VAE.Encode}(\mathbf{x})$; $B \leftarrow \emptyset$
// Stage 1: Visual Quantization
for $t \leftarrow T$ **to** 1 **do**
  $\mathcal{C}_t \leftarrow \text{GenCodebook}(K, t)$;
  $\hat{\mathbf{z}}_0 \leftarrow \text{Predict}(\mathbf{z}_t, \epsilon_\theta)$
  $n_t \leftarrow \arg\max_i \langle \mathbf{c}_i \in \mathcal{C}_t, \mathbf{z}_t - \hat{\mathbf{z}}_0 \rangle$
  // Find index
  $\mathbf{z}_{t-1} \leftarrow \text{DPM\_Step}(\mathbf{z}_t, \mathbf{c}_{n_t})$
  $B.\text{append}(\text{Bin}(n_t))$
// Stage 2: Linguistic Embedding
$\mathbf{T} \leftarrow \emptyset$; $w_{<1} \leftarrow \text{InitContext}()$
while B *is not empty* **do**
  Pop bits b from B
  $w_t \sim P_\phi(\cdot|w_{<t})$ constrained by b
  $\mathbf{T}.\text{append}(w_t)$; Update $w_{<1}$
return $\mathbf{T}$

3.4 Extraction and Deterministic Visual Reconstruction

The extraction process inverts the linguistic embedding. The receiver tokenizes the stego-text and, using the shared PRNG state, reconstructs the partitions to recover the binary sequence B bit-by-bit. The final stage involves high-fidelity reconstruction of the secret image, detailed in Algorithm 2.

Algorithm 2: VLace Decoding Pipeline

Input: Stego-text $\mathbf{T}$, Models $\boldsymbol{\epsilon}_\theta, P_\phi$, Key K
Output: Reconstructed Secret $\hat{\mathbf{x}}$

$B \leftarrow \emptyset$; $w_{<1} \leftarrow \text{InitContext}()$
// Stage 1: Extract Bits from Text
foreach $w_t \in \mathbf{T}$ **do**
  Reconstruct Huffman mapping via $P_\phi(\cdot|w_{<t})$
  $b \leftarrow \text{Decode}(w_t)$; $B.\text{append}(b)$
  Update $w_{<1}$ with w_t
// Stage 2: Reconstruct Visual Latents
Parse $B \rightarrow$ Indices $\{n_T, \ldots, n_1\}$
$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ seeded by K
for $t \leftarrow T$ **to** 1 **do**
  $\mathcal{C}_t \leftarrow \text{GenCodebook}(K, t)$
  Retrieve noise $\hat{\boldsymbol{\epsilon}} \leftarrow \mathcal{C}_t[n_t]$
  $\mathbf{z}_{t-1} \leftarrow \text{ReverseStep}(\mathbf{z}_t, \hat{\boldsymbol{\epsilon}}, \boldsymbol{\epsilon}_\theta)$
return *VAE.Decode*$(\mathbf{z}_0)$

The extracted bitstream is de-serialized into indices $\{k_t\}$, which map back to quantized noise vectors $\{\hat{\varepsilon}_t\}$ via the shared codebook. We utilize a deterministic SDE Solver for the reverse diffusion process, where the trajectory is uniquely determined by the quantized noise:

$$\hat{\mathbf{z}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{z}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\hat{\mathbf{z}}_t, t)\right) + \sigma_t \mathbf{c}_{n_t} \quad (7)$$

By replacing random Gaussian noise with the extracted quantized noise $\mathbf{c}_{n_t}$, this approach mitigates discretization errors inherent in DDIM inversion, ensuring superior semantic consistency.

4 Experiments

4.1 Experimental Setup

Datasets and Models. To evaluate the scalability and generalization of our framework, we conduct experiments across two distinct resolution settings. For standard benchmarks (256×256), we utilize CelebA-HQ, AFHQ, and LSUN-Bedroom, employing pre-trained DDPMs as the generative backbone. To assess performance in high-capacity scenarios (512×512), we incorporate the Kodak and MS-COCO datasets utilizing Stable Diffusion (SD) v1.5. For each dataset, 1,000 images are randomly sampled to ensure statistical reliability. On the linguistic side, we deploy Llama-2-7B and GPT-2 to synthesize the steganographic cover text. See Appendix A.1 for more details.

Evaluation Metrics. We assess performance across three critical dimensions. Reconstruction fidelity is quantified using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) for pixel-level accuracy, complemented by LPIPS and DISTS for perceptual similarity (Ding et al., 2024). Steganalysis resistance is evaluated using deep steganalysis tools, including SRNet (Boroumand et al., 2019) for visual and BiLSTM-Dense (Yang et al., 2020) for linguistic anomalies, where a detection accuracy closer to 0.5 signifies optimal indistinguishability. We also evaluate the naturalness of the generated text using Perplexity (PPL) and KL divergence to ensure the stego-text remains statistically consistent with natural language distributions (Detlefsen et al., 2022).

4.2 Main Results

Steganographic Quality. Quantitative comparisons across five benchmark datasets, detailed in Table 1, highlight the superior performance of VLace in reconstructing high-fidelity secret images. While INN baselines like HiNet (Jing et al., 2021) achieve the highest PSNR and SSIM scores owing to their mathematically bijective architecture, this pixel-perfect alignment often comes at the cost of generative flexibility and requires strictly aligned distribution assumptions. In contrast, our proposed framework excels in perceptual metrics, achieving the lowest LPIPS and DISTS scores across most datasets. This indicates that although our VAE-compressed reconstruction may incur minor pixel-level variance, the diffusion-based decoding synthesizes textures and semantic details that align better with human visual perception than traditional methods. Furthermore, compared to the diffusion-based competitor CRoSS (Yu et al., 2023), VLace demonstrates a substantial improvement in stability. This validates that our discrete index-guided sampling mitigates the error accumulation inherent in DDIM-inversion-based approaches, offering a robust balance between reconstruction accuracy and perceptual realism. See Appendix A.2 for qualitative demonstration.

Steganalysis of stego-text. Table 2 reveals that the near-identical behavior between payload-carrying and payload-free outputs. The steganographic embedding is achieved through distributional alignment, not pattern disruption or stego modification. This statistical indistinguishability, maintained consistently across different network architectures and payload conditions, indicates that security is inher-

Table 1: Quantitative comparison of steganographic reconstruction quality across five benchmark datasets. Metrics with ↑ indicate higher values are better; ↓ indicate lower values are better. Best results are highlighted in bold.

Method	*CelebA-HQ*				*AFHQ*				*LSUN-Bedroom*			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
4bit-LSB	30.17	0.95	0.127	0.163	31.87	0.91	0.090	0.190	29.95	0.96	0.031	0.065
Baluja (Baluja, 2020)	29.71	0.93	0.177	0.110	30.25	0.92	0.160	0.130	29.30	0.93	0.091	0.139
HiDDeN (Zhu et al., 2018)	29.68	0.92	0.286	0.184	33.95	0.92	0.250	0.160	28.94	0.91	0.249	0.152
HiNet (Jing et al., 2021)	**39.42**	**0.98**	0.051	0.045	**39.15**	**0.98**	0.048	0.039	**40.21**	**0.99**	0.035	0.030
PUSNet (Li et al., 2024a)	27.12	0.88	0.135	0.112	26.79	0.87	0.142	0.150	27.22	0.89	0.125	0.135
CRoSS (Yu et al., 2023)	25.57	0.87	0.078	0.062	36.52	0.94	0.090	0.130	24.32	0.84	0.177	0.114
Ours	32.89	0.92	**0.044**	**0.039**	32.88	0.92	**0.030**	**0.028**	33.15	0.96	**0.028**	**0.027**

Method	*Kodak*				*MS-COCO*			
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
4bit-LSB	29.10	0.92	0.057	0.055	30.81	0.90	0.032	0.058
Baluja	30.58	0.95	0.110	0.092	29.13	0.93	0.127	0.148
HiDDeN	27.93	0.91	0.176	0.140	28.33	0.86	0.118	0.106
HiNet	**38.85**	**0.97**	0.055	0.042	**39.72**	**0.99**	0.028	0.025
PUSNet	27.56	0.86	0.115	0.130	26.98	0.88	0.105	0.120
CRoSS	25.31	0.80	0.084	0.080	23.79	0.80	0.083	0.079
Ours	28.06	0.89	**0.027**	**0.025**	28.12	0.90	**0.018**	**0.020**

Table 2: Text steganography analysis demonstrates plug-and-play compatibility and distribution-preserving stealth. Our method maintains near-natural metrics.

Metric	*GPT-2*			*Llama-2-7B*			Diff
	Natural	w/o	w/	Natural	w/o	w/	Δ
PPL ↓	25.3	26.1	26.8	18.7	19.5	20.3	+0.7
bpc ↑	0.00	0.00	0.34	0.00	0.00	0.32	+0.34
P_E ↓	0.500	0.507	0.523	0.500	0.503	0.518	+0.016

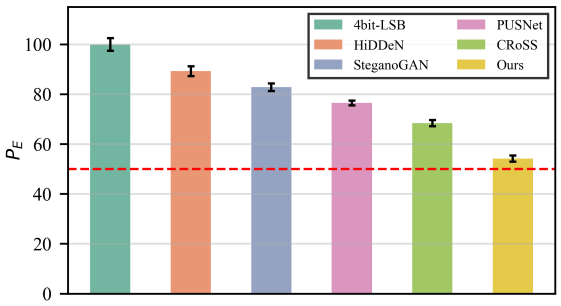

Figure 3: Comparative evaluation of anti-steganalysis based on P_E.

ent to the embedding paradigm itself rather than dependent on any specific model implementation. VLace demonstrates strong platform and model agnosticism, enabling reliable and undetectable text steganography across varied deployment scenarios without requiring architectural adjustments or sacrificing detection resistance. Unlike traditional methods that embed images within images, our approach uses cross-modal transformation to conceal the secret within innocuous text streams, bypassing statistical steganalysis tools designed for intra-modal detection.

Steganalysis results. Figure 3 confirms our distribution-aligned steganography stays resistant to steganalysis, with a red dashed line indicating the algorithm cannot distinguish cover from stego. Traditional methods become detectable as the collected pair data increases. This resilience of our method stems from the zero distributional divergence, which prevents classifiers from learning discriminative features regardless of data size.

Large-in-Small steganography. To probe the capacity boundaries of VLace, we conduct a large-in-small experiment utilizing ImageGPT (32×32) as the steganographic carrier for high-resolution (256×256) secret images in Table 3. By treating pixels as discrete tokens within a 512-cluster vocabulary, our framework achieves an unprecedented embedding rate of approximately 8.7 bpp, approaching the theoretical entropy limit of $\log_2 512 = 9$ bits and surpassing the conventional 4 bpp ceiling of LSB methods. Despite the extreme spatial compression ratio, where the cover con-

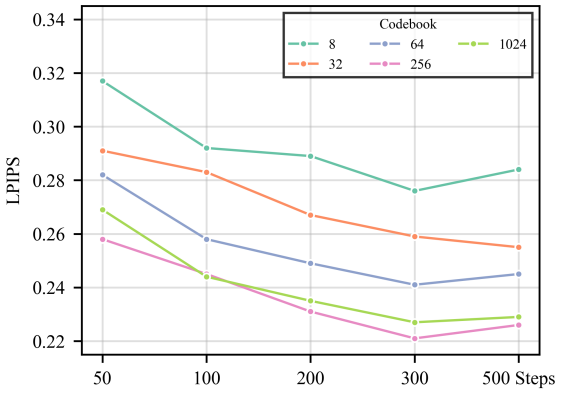

Figure 4: Impact of sampling steps and codebook sizes on reconstruction quality.

Table 3: Performance of Large-in-Small steganography: hiding 256×256 secret images into 32×32 ImageGPT covers. VLace achieves ≈ 8.7 bpp with high fidelity.

Dataset	Method	Bits	bpp	PSNR	SSIM	LPIPS
CelebA	4bit-LSB	4096	4.0	18.45	0.42	0.610
	Ours	8850	8.6	31.52	0.93	0.045
Bedroom	4bit-LSB	4096	4.0	17.92	0.38	0.635
	Ours	8920	8.7	32.14	0.91	0.048

tains only $1/64$th the pixels of the secret. VLace maintains robust reconstruction fidelity. This capability validates that our discrete quantization pipeline acts as an efficient semantic compressor, while the receiver's diffusion backbone leverages generative priors to complete high-frequency details from the compact index payload, enabling the covert transmission of heavy visual assets through extremely bandwidth-constrained channels. See Appendix A.3 for additional qualitative results.

Robustness Against Channel Distortions. Table 5 demonstrates that VLace achieves perfect robustness against visual degradation. This advantage derives from leveraging the semantic invariance of text: while image containers suffer from re-encoding and compression artifacts on platforms like WeChat or Twitter, textual data is transmitted via lossless standards. Thus, in the absence of manual editing, our text-based payload remains bit-wise identical after cross-platform relay. This empirically proves that migrating to the linguistic domain solves the fundamental channel distortion challenge plaguing pixel-based methods.

4.3 Ablation study

Sampling steps and sizes of codebook. Figure 4 reveals that optimal latent compression requires balanced parameters. A moderate number of steps aligns quantized noise with image semantics without accumulating redundancy, while a mid-sized codebook captures nuanced noise features efficiently, avoiding the detail loss of small codebooks and the redundancy of large ones. This configuration maximizes perceptual fidelity, ensuring high-quality steganographic reconstruction for secure image-to-text transmission.

Table 4: Ablation Study of Top-k Sampling Strategy on Text Generation Quality and Steganographic Stealth.

Metric	20	50	100	200	300
PPL (↓)	5.17	4.89	4.78	4.71	4.68
P_E (↓)	0.021	0.042	0.068	0.095	0.123

Table 5: Robustness evaluation against real-world channel distortions.

Distortion	HiDDeN	SteganoGAN	CRoSS	Ours
Clean	29.68	27.12	25.57	28.71
JPG90	26.45	22.10	18.32	28.71
JPG50	19.82	14.55	10.15	28.71
Social Media	21.50	16.80	15.20	28.71

Top-k sampling. Table 4 evaluates the Top-k sampling in balancing text quality and steganographic security within our framework. Increasing Top-k enhances linguistic fluency by diversifying token selection, thereby lowering PPL and improving text naturalness. However, this expansion also distorts the statistical properties of the generated text, such as token frequency and syntactic consistency, making it more detectable to steganalyzers.

5 Conclusion

In this work, we introduce VLace, a training-free framework that bridges the challenge of the modality-capacity disparity by reformulating image steganography as a cross-modal, generative process. By quantizing the deterministic trajectories of diffusion models, we demonstrated that high-resolution visual secrets can be transformed into ultra-compact binary payloads. These payloads are then seamlessly embedded into AI-generated text via a distribution-aligned sampling mechanism, ensuring the stego-text is statistically indistinguishable from natural LLM outputs while enabling high-fidelity reconstruction. VLace is opening new avenues for protecting sensitive visual data in an increasingly monitored digital landscape.

Limitations

While VLace exhibits exceptional resilience to technical channel distortions, it remains sensitive to semantic textual alterations, such as insertion, deletion, or adversarial rewriting or paraphrasing. This limitation stems from the fundamental trade-off between embedding capacity and robustness: our framework prioritizes the high bandwidth required to conceal full-resolution visual secrets, necessitating precise token-level alignment to decode the dense index sequence. In contrast, resilience against semantic editing is typically the domain of watermarking schemes, which operate under significantly constrained capacity regimes to maximize redundancy. Extending high-capacity steganography to withstand linguistic perturbations including advanced error-correction coding or semantic-invariant mapping remains a challenging frontier for future exploration.

Despite our patch-wise quantization achieving high compression ratios, the intrinsic information density gap between images and text remains a physical bottleneck. To conceal a high-resolution image (e.g., 512×512) with high fidelity, the system may need to generate a relatively long textual passage like one thousand bits ($\sim$ two hundred tokens/words). In scenarios with strict character limits such as SMS, transmitting such extensive text may appear conspicuous or be operationally infeasible. Future work could explore more aggressive semantic compression techniques to further reduce the required carrier length.

Acknowledgements

This work is supported in part by the Youth Program of Humanities and Social Sciences of the MoE (23YJCZH291) and the Key Research Platforms and Projects of Universities in Guangdong Province (Science and Technology) (2023KTSCX253, 2024KTSCX153).

References

Ahd Aljarf, Haneen Zamzami, and Adnan Abdul-Aziz Gutub. 2023. Is blind image steganalysis practical using feature-based classification? *Multimedia Tools and Applications*, 83:4579 – 4612.

Shumeet Baluja. 2020. Hiding Images within Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1685–1697.

Mehdi Boroumand, Mo Chen, and Jessica Fridrich. 2019. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193.

Kejiang Chen, Hang Zhou, Hanqing Zhao, Dongdong Chen, Weiming Zhang, and Nenghai Yu. 2022. Distribution-Preserving Steganography Based on Text-to-Speech Generative Models. *IEEE Transactions on Dependable and Secure Computing*, 19(5):3343–3356.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. 2023. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*.

Binyue Deng, Denghui Zhang, Fashan Dong, Junjian Zhang, Muhammad Shafiq, and Zhaoquan Gu. 2023. Rust-Style Patch: A Physical and Naturalistic Camouflage Attacks on Object Detector for Remote Sensing Images. *Remote Sensing*, 15(4):885.

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. TorchMetrics - Measuring Reproducibility in PyTorch. *Journal of Open Source Software*, 7(70):4101.

Jinyang Ding, Kejiang Chen, Yaofei Wang, Na Zhao, Weiming Zhang, and Nenghai Yu. 2023. Discop: Provably Secure Steganography in Practice Based on "Distribution Copies". In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2238–2255.

Keyan Ding, Rijin Zhong, Zhihua Wang, Yang Yu, and Yuming Fang. 2024. Adaptive structure and texture similarity metric for image quality assessment and optimization. *IEEE Transactions on Multimedia*, 26:5398–5409.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *2021 International Conference on Learning Representations (ICLR)*.

Tomáš Filler, Jan Judas, and Jessica Fridrich. 2011. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3):920–935.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *2020 Advances in Neural Information Processing Systems (NeurIPS)*.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2024. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478.

Jun Jiang, Zijin Yang, Weiming Zhang, Nenghai Yu, and Kejiang Chen. 2025. StegoZip: Enhancing linguistic steganography payload in practice with large language models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. 2021. HiNet: Deep Image Hiding by Invertible Network. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4713–4722.

Ali Kartit. 2022. New approach based on homomorphic encryption to secure medical images in cloud computing. *Trends in Sciences*, 19(9):3970–3970.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*.

Guobiao Li, Sheng Li, Zicong Luo, Zhenxing Qian, and Xinpeng Zhang. 2024a. Purified and Unified Steganographic Network. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27559–27568.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024b. Autoregressive Image Generation without Vector Quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yiming Liu, Kezhao Liu, Yao Xiao, ZiYi Dong, Xiaogang Xu, Pengxu Wei, and Liang Lin. 2025. Towards Understanding the Robustness of Diffusion-Based Purification: A Stochastic Perspective. *International Conference on Representation Learning*, 2025:96416–96442.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2025. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *Machine Intelligence Research*, 22(4):730–751.

Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. 2025. Compressed Image Generation with Denoising Diffusion Codebook Models. In *International Conference on Machine Learning (ICML) 2025*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Debmalya Sarkar, P. Vinod, and Suleiman Y. Yerima. 2020. Detection of tor traffic using deep learning. *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.

Johannes Schusterbauer, Ming Gui, Frank Fundel, and Björn Ommer. 2025. Diff2Flow: Training flow matching models via diffusion model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. In *2021 International Conference on Learning Representations (ICLR)*.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930.

UK Parliament. 2023. Online Safety Act 2023 - Parliamentary Bills - UK Parliament.

Bram Wallace, Akash Gokul, and Nikhil Naik. 2023. EDICT: Exact Diffusion Inversion via Coupled Transformations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22532–22541.

Fangyikang Wang, Hubery Yin, Yue-Jiang Dong, Huminhao Zhu, Hanbin Zhao, Hui Qian, Chen Li, and 1 others. 2024. Belm: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. *Advances in Neural Information Processing Systems*, 37:46118–46159.

Yaofei Wang, Gang Pei, Kejiang Chen, Jinyang Ding, Chao Pan, Weilong Pang, Donghui Hu, and Weiming Zhang. 2025. SparSamp: Efficient Provably Secure Steganography Based on Sparse Sampling. In *34th USENIX Security Symposium (USENIX Security '25)*.

Wenying Wen, Jiacong Fan, Yushu Zhang, and Yuming Fang. 2023. APCAS: Autonomous Privacy Control and Authentication Sharing in Social Networks. *IEEE Transactions on Computational Social Systems*, 10(6):3169–3180.

Mingqing Xiao, Shuxin Zheng, Chang Liu, Zhouchen Lin, and Tie-Yan Liu. 2023. Invertible Rescaling Network and Its Extensions. *International Journal of Computer Vision*, 131(1):134–159.

Hao Yang, YongJian Bao, Zhongliang Yang, Sheng Liu, Yongfeng Huang, and Saimei Jiao. 2020. Linguistic steganalysis via densely connected lstm with feature pyramid. In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, page 5–10, New York, NY, USA.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. 2024. Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12162–12171.

Weike You, Hong Zhang, and Xianfeng Zhao. 2021. A siamese CNN for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 16:291–306.

Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. 2023. CRoSS: Diffusion Model Makes Controllable, Robust and Secure Image Steganography. In *Advances in Neural Information Processing Systems (NeurIPS) 2023*.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. 2024. Normalizing Flows are Capable Generative Models. *Preprint*, arXiv:2412.06329.

Denghui Zhang, Muhammad Shafiq, Gautam Srivastava, Thippa Reddy Gadekallu, Le Wang, and Zhaoquan Gu. 2024. STBCIoT: Securing the Transmission of Biometric Images in Customer IoT. *IEEE Internet of Things Journal*, 11(9):16279–16288.

Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. 2019. SteganoGAN: High Capacity Image Steganography with GANs. *CoRR*.

Zengrui Zhao, Celimuge Wu, Yangfei Lin, Lei Zhong, Yusheng Ji, Tomoaki Ohtsuki, and Mehdi Bennis. 2025. LRGD: Low-Rank Guided Diffusion for Robust Image Transmission in Semantic Communication. *IEEE Transactions on Cognitive Communications and Networking*, pages 1–1.

Zhili Zhou, Yuecheng Su, Jin Li, Keping Yu, Q. M. Jonathan Wu, Zhangjie Fu, and Yunqing Shi. 2023. Secret-to-Image Reversible Transformation for Generative Steganography. *IEEE Transactions on Dependable and Secure Computing*, 20(5):4118–4134.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data With Deep Networks. In *Computer Vision– ECCV 2018*, pages 682–697, Cham.

A Appendix

A.1 Implementation Details

Generative Configurations. All experiments are conducted on a single NVIDIA RTX 4090 GPU (24GB). It is important to note that it is unnecessary to transmit the model itself because many large models are readily available to the public on platforms like Hugging Face and GitHub. Therefore, both parties only need to agree on which model to use and then load it separately. We tailor the sampling strategies to the specific model architectures:

- **Pixel-Space (DDPM):** We employ a standard linear noise schedule with $T = 1,000$ sampling steps. The visual-to-binary quantization is calibrated to yield a payload capacity of approximately 8,000 bits per image.
- **Latent-Space (SD v1.5):** To reconcile sampling efficiency with reconstruction fidelity, we adopt the DPM-Solver++ scheduler (Lu et al., 2025). This reduces the inference trajectory to only 20 steps (compared to 50–100 steps in standard DDIM) while maintaining high visual quality. The corresponding payload capacity is calibrated to approximately 10,000 bits.

Null-Text Reconstruction. A distinct advantage of our framework is its independence from auxiliary text prompts. Unlike prompt-dependent paradigms such as CRoSS (Yu et al., 2023) or Gaussian Shading (Yang et al., 2024), which require precise text descriptions to guide the restoration process, our quantized index sequence inherently encapsulates the semantic and structural information of the secret image. Consequently, we perform reconstruction using a null-text condition, eliminating the bandwidth overhead of transmitting prompts.

Robustness Evaluation Setup. We report the Secret Recovery PSNR (dB) of the extracted secret image. Social Media Transfer simulates compression algorithms used by platforms like WeChat. Note that pixel-based attacks (JPEG, Noise, Resizing) are applied to the image carriers of baseline methods, while our text carrier undergoes standard ASCII transmission (Copy-Paste).

A.2 Textual Stealth and Visual Fidelity

To empirically validate the imperceptibility of our framework, we present a detailed case study in Table 6 for 1024 bits. The experiment utilizes the context *I remember this film, it was the first film I had watched at the cinema* to trigger generation. VLace embeds the quantized indices of four distinct CelebA-HQ images into the generated narrative. The resulting text maintains the syntactic fluency and semantic logic of the Normal Sampling baseline, avoiding the repetition or linguistic anomalies often associated with high-capacity steganography. the Revealed row demonstrates that this linguistic obfuscation does not compromise visual integrity. The reconstructed images retain the fine-grained facial features and textures of the Orig inputs, confirming that our method achieves a seamless unification of linguistic naturalness and high-fidelity visual reconstruction.

Table 6: Qualitative demonstration of text stealthiness and visual fidelity. Despite embedding the four secret images (bottom), the generated Stego-text remains linguistically indistinguishable from the baseline Normal Sampling. The Revealed images exhibit negligible perceptual deviation from the Orig, confirming effective high-capacity transmission via natural language generation.

Context:
I remember this film, it was the first film I had watched at the cinema.

Normal Sampling:
I remember this film, it was the first film I had watched at the cinema. And it was one of the first films I saw where you saw a woman doing something really interesting. And I remember thinking, "That's very interesting, because you can see the woman's personality and you can see the strength of her character - but it's very subtle." It was the first time I had watched a film where I was like, "Wow, she's very strong, I guess she's a strong woman." And I can remember thinking, "Well, that's what's going to make me think of her more." I remember she was in the middle, I think, of a divorce and she was working as a waitress and she was looking for a job, she saw this job that she was very good at but she could never quite get the pay - but she was going through an emotional breakdown. She was very depressed. And she was doing a lot of drugs. And she worked at the diner, she did the dishes and she'd go over and she'd say, "Oh, by the way, I'm going to work at this diner," and she'd leave the job. I mean, she's just one of those people. And she was doing this, she was trying to get back into the community

Steto-text:
I remember this film, it was the first film I had watched at the cinema. The film was called 'The Last Days of the Great Depression', and it was about a group of people who were trying to get through the Depression. I remember it being an interesting film. It was shot in New York, in the 1940s. It was one of the first films I saw; I remember it. I remember sitting on the couch in the dark, and it was like a movie that was about a lot of people and you could feel the atmosphere. It was very scary. And it was a very depressing film. (laughs) I remember being very scared, but then I remember going home and having a very good time watching the film. It was a very good film. So what was the biggest challenge of shooting 'The Last Days of the Great Depression'? Well, the biggest challenge was the weather. I mean, it's kind of a weird story, but this was a film that was shot in New York, and it was a winter movie that was shot in New York. So I was trying to make the film in New York, and then I had to shoot it in a city that didn't have the weather that I wanted.

Orig Revealed

Figure 5: Qualitative visualization of stego-images generated by ImageGPT across different prefix lengths. The columns correspond to prefix token counts of 100, 200, 400, and 500, respectively, followed by the original image (Orig). A larger prefix provides stronger semantic guidance, resulting in higher visual resemblance to the ground truth.

A.3 Visualizing the Large-in-Small Steganography

We provide a qualitative comparison of stego-images synthesised by ImageGPT under varying degrees of autoregressive guidance. Figure 5 illustrates shorter prefixes (e.g., 100 tokens) grant the model greater generative freedom, which facilitates higher embedding capacity but may result in semantic deviation from the original subject. Conversely, increasing the prefix length to 500 tokens provides a robust semantic anchor, forcing the generative process to strictly adhere to the original visual manifold. This visualization empirically validates our Large-in-Small strategy, demonstrating that a balanced prefix ratio can maintain perceptual fidelity while securing substantial embedding space.