Supplementary Information

Title: Overcoming the Semantic Bottleneck for Deterministic Structural Control in Text-to-Image Synthesis

Authors: Muhammad Bilal Khan*[1], Second Author[2]

Affiliations: [1]Robotics & Intelligent System Engineering (RISE) Lab, School of Mechanical and Manufacturing Engineering (SMME), National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan [2]King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia.

***Correspondence**: Correspondence and requests for materials should be addressed to Muhammad Bilal Khan (email: mkhan.rime23smme@nust.edu.pk).*

Overview of Supplementary Material

This document contains the following sections:

- **Supplementary Note 1:** Mathematical Derivations of Latent Steering.
- **Supplementary Note 2:** Pattern Dictionary Detail and Kernel Specifications.
- **Supplementary Note 3:** Full Ablation Studies on Temporal and Hierarchical Sensitivity.
- **Supplementary Note 4:** Failure Mode Analysis and Operational Boundaries.
- **Supplementary Note 5:** Domain-Specific Robustness and Generalization.

Note on Visuals: All Extended Data Figures 1–2 and Extended Data Tables 1–2 referenced in this document are located at the end of the Main Manuscript file, following the References section.

Supplementary Note 1: Mathematical Derivations of Latent Steering

The mathematical foundation of Procedural Latent Prompt Injection (PLPI) rests upon the reinterpretation of the reverse diffusion trajectory as a controllable dynamical system. To define the mechanism of structural steering, we first establish the standard probability flow of a latent diffusion model. In the continuous-time limit, the forward diffusion process is an SDE that transforms a data distribution p_0 into a noise distribution $p_T \approx \mathcal{N}(0, I)$. The corresponding reverse process, used for synthesis, is governed by the following SDE:

$$dz = [f(z,t) - g(t)^2 \nabla_z \log p_t(z)]dt + g(t)d\bar{w}$$

where $f(z,t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\nabla_z \log p_t(z)$ is the score function. In text-conditioned models, the score function is approximated by a neural network $\epsilon_\theta(z_t, t, c)$, where c is the linguistic embedding. The semantic bottleneck occurs because c influences the score function through a global cross-attention mechanism that lacks local geometric determinism.

The Steering perturbation

We introduce structural determinism by perturbing the latent trajectory during the reverse denoising process. We define a steering operator $\mathcal{S}$ that intervenes at discrete timesteps within the Plasticity Window. The steered latent state z_t' is formulated as a linear combination of the current noisy latent z_t and a procedural structural prior $\mathcal{P}$:

$$z_t' = (1 - \lambda)z_t + \lambda \mathcal{P}_{norm}$$

Here, $\mathcal{P}_{norm}$ is the procedurally synthesized kernel (e.g., symmetry or foveated focus) that has been normalized to match the first and second moments of the latent manifold at time t. By substituting this steered state back into the update rule, we effectively modify the effective score function:

$$\nabla_z \log p_t(z)_{steered} \approx \nabla_z \log p_t(z_t') + \nabla_z \mathcal{P}_{norm}$$

This intervention forces the SDE to evolve toward a geometric attractor. Because the kick occurs in the latent space before the U-Net predicts the noise residual, the model perceives the injected structure as a foundational feature of the noise field, which it then attempts to denoise into a coherent structural element.

Phase preservation: linear mixing vs. Slerp

A critical derivation in our framework is the proof that linear mixing outperforms Spherical Linear Interpolation (Slerp) for structural steering. Slerp is conventionally defined as:

$$\text{Slerp}(z, \mathcal{P}; \lambda) = \frac{\sin\big((1-\lambda)\theta\big)}{\sin\theta} z + \frac{\sin(\lambda\theta)}{\sin\theta} \mathcal{P}$$

where $\theta = \arccos\left(\frac{z \cdot \mathcal{P}}{|z||\mathcal{P}|}\right)$. While Slerp is geometrically optimal for interpolating between two vectors on a high-dimensional hypersphere (preserving the constant norm of Gaussian noise), it is mathematically detrimental to structural phase preservation in diffusion. In a high-dimensional latent space ($D \approx 10^4$), any random noise vector z and a structured procedural prior $\mathcal{P}$ are almost certainly orthogonal, meaning $\theta \approx \pi/2$. In this limit, Slerp simplifies to:

$$\text{Slerp}(z, \mathcal{P}; \lambda) \approx \cos\left(\lambda\frac{\pi}{2}\right) z + \sin\left(\lambda\frac{\pi}{2}\right) \mathcal{P}$$

This trigonometric scaling significantly distorts the additive relationship of the noise components. Conversely, linear mixing ($\mathcal{L}$) preserves the spatial phase of the structural signal. In the frequency domain, the latent tensor can be represented as a sum of Fourier components. Linear mixing acts as a superposition of these components:

$$\mathcal{F}(\mathcal{L}) = (1-\lambda)\mathcal{F}(z) + \lambda\mathcal{F}(\mathcal{P})$$

Because the U-Net's convolutional filters rely on the relative phase of these frequencies to detect edges and boundaries, linear mixing ensures that the structural gradients of $\mathcal{P}$ remain coherent. Slerp introduces non-linear phase shifts that decouple the injection from the U-Net's learned spatial weights, leading to blurred boundaries and structural decoherence. Our derivation demonstrates that linear mixing maintains the Gaussian properties required for the diffusion schedule while ensuring that the structural phase of the procedural prior dominates the denoising trajectory during the Plasticity Window.

Signal-to-noise ratio (SNR) and structural reification

The timing of the injection is governed by the evolution of the Signal-to-Noise Ratio (SNR). We define the reification of a structure as the point where the steering term $\lambda\mathcal{P}_{norm}$ exceeds the stochastic variance of the noise field. By analyzing the variance schedule β_t, we identify that between timesteps 10 and 20, the latent manifold is in a plastic state, coherent enough to hold a signal but chaotic enough to be steered. Injections made at SNR ≈ 0.1 allow the procedural prior to serve as a seed for the subsequent $z_{t\to 0}$ trajectory, effectively locking the global layout before the model commits to high-frequency pixel details.

Supplementary Note 2: Pattern Dictionary Detail and Kernel Specifications

The efficacy of Procedural Latent Prompt Injection (PLPI) depends on the precise mathematical construction of latent kernels. Unlike linguistic prompts, which are processed through high-dimensional

embedding lookups, these kernels are synthesized as spatial-frequency priors that directly modulate the latent manifold. This dictionary provides the technical specifications for the three primary categories of procedural patterns utilized in this study: focal, symmetric, and periodic kernels.

Gaussian focal kernels (Radial basis functions)

To enforce foveated attention and central subject placement, we utilize a Radial Basis Function (RBF) to generate a focal mask. The kernel is defined by the Euclidean distance d of each coordinate (i, j) in the latent tensor from the geometric center (c_i, c_j). For a latent resolution of $H \times W$, the distance is calculated as:

$$d(i,j) = \sqrt{(i - c_i)^2 + \left(j - c_j\right)^2}$$

The procedural prior $\mathcal{P}$ is then synthesized using the Gaussian exponential:

$$\mathcal{P}_{focal}(i,j) = \exp\left(-\frac{d(i,j)^2}{2\sigma^2}\right)$$

where σ acts as the focal breadth parameter. In our experiments, σ is typically set to 0.25 of the total latent dimension to ensure a soft fall-off that guides the U-Net's attention toward the center without creating hard edge artifacts. As seen in Extended Data Fig. 2, this creates a localized structural basin where feature emergence is prioritized.

Axial symmetry tensors

Bilateral and axial symmetry are enforced through a linear reflection operator. For vertical symmetry, which is essential for anatomical and architectural synthesis, we define a transformation that synchronizes the left and right hemispheres of the latent field. Given a latent tensor z, the symmetric prior is generated by:

$$\mathcal{P}_{sym} = \omega \cdot z + (1 - \omega) \cdot \text{Flip}_{horiz}(z)$$

where $\omega = 0.5$ represents a perfect structural average. To avoid a sharp seam at the axis of reflection, a 1D Gaussian smoothing kernel is applied along the central column. This operator ensures that the high-level semantic features, such as the lobes of a brain in an MRI scan or the towers of a cathedral are aligned with mathematical precision before the U-Net initiates the high-resolution decoding phase.

Periodic Nyquist kernels (high-frequency textures)

To steer the model toward specific textural densities, we employ periodic tensors designed at the Nyquist frequency of the latent space. A checkerboard or grid pattern is synthesized using the product of orthogonal sine waves:

$$\mathcal{P}_{freq}(i,j) = \sin(k_x \cdot i) \cdot \sin(k_y \cdot j)$$

where k represents the spatial frequency. Because the latent space is a $1/8$th compression of the pixel space, a kernel with a frequency of $k = \pi$ (one cycle every two pixels) corresponds to extremely high-frequency micro-textures in the final image. By adjusting k, we can deterministically control the granularity of the generated surface, such as the scale of cellular membranes in microscopy or the density of brickwork in architectural renders.

Moment matching and tensor normalization

A critical requirement for all kernels in the dictionary is moment matching. The U-Net's internal weights are calibrated to process tensors with zero mean and unit variance. Injecting a raw sine wave or Gaussian mask would shift the latent distribution, causing semantic collapse. Therefore, every procedural kernel $\mathcal{P}$ is subjected to the following normalization before injection:

$$\mathcal{P}_{norm} = \frac{\mathcal{P} - \mu_{\mathcal{P}}}{\sigma_{\mathcal{P}}} \cdot \text{std}(z_t) + \text{mean}(z_t)$$

This ensures that the procedural prompt is statistically indistinguishable from the surrounding noise z_t, allowing it to act as a steerable signal that the model treats as an inherent feature of the generative trajectory. This mathematical standardization allows the PLPI framework to remain domain-agnostic, functioning as a universal Lock-and-Key mechanism for structural control.

Supplementary Note 3: Full Ablation Studies on Temporal and Hierarchical Sensitivity

To validate the mechanistic underpinnings of Procedural Latent Prompt Injection (PLPI), we conducted an extensive suite of ablation studies focusing on two primary axes: temporal dynamics (the when) and architectural sensitivity (the where). These experiments isolate the contribution of the Plasticity Window and the hierarchical layers of the U-Net, providing empirical evidence for the Latent-First Hypothesis. The results summarized here correspond to the performance metrics detailed in Extended Data Table 1.

Temporal dynamics: characterizing the plasticity window

We evaluated the impact of latent injection across the entire 50-step diffusion trajectory, performing interventions at 5-step intervals. The objective was to determine the relationship between the Signal-to-Noise Ratio (SNR) and the reification of structural priors.

- The Stochastic Entropy Phase (Steps 0–10): Injections performed during the earliest stages of diffusion yielded negligible structural alignment. At $t > 40$ (in reverse diffusion timesteps), the latent manifold is dominated by high-variance Gaussian noise. The drift coefficient of the SDE is insufficiently defined to anchor the procedural signal. Consequently, the U-Net's score function effectively treats the injected tensor as random noise, leading to outputs that remain structurally stochastic.

- The Plasticity Window (Steps 10–20): A sharp phase transition in controllability was observed starting at step 10. Within this window, the latent representation begins to resolve into semi-coherent features. At $t \approx 15$, the SNR reaches a critical threshold where the injected geometric prior acts as a deterministic attractor. Quantitative analysis showed that interventions here result in the peak 19.6% improvement in CLIP-based structural alignment. This indicates that the manifold is plastic as it retains enough entropy to be steered but sufficient structure to propagate the signal into the final reconstruction.

- The Structural Rigidity Phase (Steps 20–50): Post-step 20, the macro-structure of the image is largely settled within the latent trajectory. Injections performed during this late phase failed to alter global composition (e.g., axial symmetry or subject placement). Instead, the steering kick manifested as superficial textural overlays or high-frequency artifacts. This confirms that structural determinism is a time-dependent phenomenon; once the SDE trajectory has committed to a specific structural basin, latent perturbations can no longer reorganize the global semantic layout.

Hierarchical sensitivity: encoder, mid-block, and decoder

The U-Net architecture processes information at varying scales, from high-resolution pixel-adjacent features in the outer layers to compressed semantic representations in the bottleneck. Our ablation study mapped the effectiveness of steering across these layers.

- Encoder and Input-Level Injection: Interventions at the input level (z_t) primarily influenced high-frequency components. While these successfully enforced local patterns, such as periodic textures or fine-grained edges, showed poor leveraged control over the scene's composition. Because the encoder progressively compresses this signal, much of the structural prior is filtered out before it reaches the semantic bottleneck.

- The Mid-Block Bottleneck (Optimal Injection Point): The U-Net's mid-block represents the highest semantic density and lowest spatial resolution (8×8). Ablation results confirm that this is the most effective site for structural steering. Perturbations at this bottleneck are magnified during the decoding process, as every subsequent upsampling layer is conditioned on the steered semantic blueprint. This allows for a global compositional shift that maintains symmetry and focus with minimal injection strength, achieving the highest variance regulation.

- Decoder and Skip-Connection Sensitivity: Injecting into the decoder path in isolation proved counter-productive. Because the decoder is constantly receiving cleaner structural information from the encoder via skip connections, decoder-only injections create a semantic conflict. This mismatch between the encoder's original stochastic structure and the decoder's steered structure results in ghosting artifacts and edge incoherence. This finding reinforces the necessity of Latent-First intervention, where the steering must occur early enough in the hierarchy to guide the entire reconstruction path.

Parametric sensitivity: the strength coefficient (λ)

Finally, we ablated the injection strength parameter λ to find the optimal operating range. We observed a non-linear relationship between structural fidelity and semantic realism. At $\lambda < 0.3$, the steering effect is often too subtle to overcome the model's learned priors. At the identified optimum of $\lambda = 0.4$ to 0.6, the procedural signal and the model's generative capacity exist in a state of productive interference, yielding high-fidelity, deterministic results. However, as λ approaches 1.0, the injection induces manifold departure, the latent state is pushed so far from the learned distribution that the U-Net fails to decode it into a realistic image, producing over-saturated, repetitive patterns known as the blow-over effect.

These ablation studies provide a rigorous map of the operational boundaries of PLPI. They establish that deterministic control is not a product of brute-force noise replacement but a precise mathematical intervention timed to the latent manifold's phase transition and positioned at its semantic core. All metrics from these studies, including the cross-domain consistency of the 10-20 step window, are detailed in the Full Ablation Matrix in Extended Data Table 1.

Supplementary Note 4: Failure Mode Analysis and Operational Boundaries

The deterministic nature of Procedural Latent Prompt Injection (PLPI) necessitates a rigorous characterization of its failure modes to establish safe operational boundaries. While the framework provides a robust Lock-and-Key mechanism for structural control, excessive or poorly timed mathematical interventions can force the latent trajectory into regions of the manifold that are

incompatible with the pre-trained U-Net's learned distribution. These failure states, characterized below, are visually catalogued in Extended Data Fig. 1 – Failure Modes Grid.

The blow-over effect and contrast saturation

The most common failure mode occurs when the injection strength coefficient (λ) exceeds the stability threshold of the latent manifold (typically $\lambda > 0.7$). In this regime, the procedural prior $\mathcal{P}_{norm}$ overpowers the stochastic noise z_t to such a degree that the denoising trajectory becomes hyper-deterministic. This results in the blow-over effect, where the model generates unnatural, repetitive patterns and high-frequency noise artifacts. Mechanistically, this occurs because the U-Net attempts to denoise a signal that lacks the necessary stochastic variance to trigger its learned semantic priors. The resulting output often exhibits extreme contrast saturation and halo effects, as the predicted noise residuals ϵ_θ become over-scaled in an attempt to reconcile the rigid procedural signal with the expected Gaussian distribution.

Semantic collapse and manifold departure

Semantic collapse represents a more severe failure state where the latent injection causes a manifold departure. This typically occurs when the geometric prior is fundamentally alien to the model's architectural biases, for instance, injecting extremely dense, high-frequency checkerboard patterns into a model trained primarily on natural, low-frequency datasets. In such cases, the U-Net cannot map the steered latent state to any coherent feature set, resulting in an output of unrecognizable geometric shrapnel or single-color blocks. This failure highlights that structural steering is a collaborative process between the external kick and the internal score function; if the steering force pushes the latent state into a zero-probability region of the manifold, the generative process fails.

Temporal artifacts outside the plasticity window

The timing of the injection is as critical as its strength. As detailed in Extended Data Fig. 1, interventions made after the Plasticity Window (steps > 20) frequently result in ghosting or structural schizophrenia. At this late stage, the U-Net has already committed to a structural trajectory. A late-step injection attempts to force a new geometry onto a solidified layout, leading to a conflict where the model generates two overlapping, incoherent structures. Conversely, injections made too early (steps < 10) are often simply ignored due to high-entropy diffusion, resulting in a total loss of deterministic control.

These failure modes define the operational limits of PLPI: for optimal fidelity, λ must be maintained between 0.4 and 0.6, and the intervention must be strictly synchronized with the Plasticity Window to avoid the semantic and structural decoherence observed at the boundaries of the diffusion trajectory.

Supplementary Note 5: Domain-Specific Robustness and Generalization

The robustness of Procedural Latent Prompt Injection (PLPI) was evaluated across three disparate image distributions: clinical medical imaging, high-resolution scientific microscopy, and complex artistic compositions. Unlike learned-control methods that require domain-specific fine-tuning or specialized textual datasets, PLPI maintains operational stability through its latent-first mathematical approach. By manipulating the manifold drift rather than linguistic embeddings, the framework ensures that geometric priors are enforced regardless of the underlying semantic category. The performance across these domains is quantified in Extended Data Table 2 – Per-Domain Metrics.

Precision-critical medical synthesis (MRI)

In the synthesis of medical imaging, such as magnetic resonance imaging (MRI) of the brain, structural symmetry and anatomical registration are prerequisites for diagnostic utility. Standard diffusion processes often suffer from structural drift, where stochastic noise resolves into anatomically impossible asymmetries. We applied vertical symmetry tensors at the U-Net mid-block bottleneck to enforce axial alignment.

The results demonstrate that PLPI produces brain scans with 19.6% higher structural registration scores compared to stochastic baselines. This deterministic steering ensures that bilateral structures such as the ventricles and cortical hemispheres, are mirrored with mathematical precision. Such reproducibility is vital for generating synthetic datasets for diagnostic training, where anatomical hallucinations can lead to catastrophic failure in downstream machine-learning models.

Structural biology and cellular microscopy

For cellular microscopy, the challenge lies in capturing high-frequency micro-structures like cell boundaries, filament networks, and organelles. These features are often lost to the smoothing effects of text-conditioned encoders. We utilized high-frequency Nyquist kernels at the input-level injection point to guide the emergence of periodic tissue patterns.

Ablation on these biological datasets revealed that while mid-block injections define the global layout of the slide, input-level refinements are necessary to preserve the fidelity of micro-scale structural details. PLPI achieved an 18.4% improvement in edge-consistency metrics within this domain. By providing a spatial-frequency blueprint directly to the latent space, the framework allows for the synthesis of complex biological textures that are impossible to describe with equivalent linguistic precision.

Artistic composition and procedural layouts

The framework's generalization was further stress-tested on artistic datasets requiring foveated focus and specific subject placement. In standard text-to-image synthesis, the primary subject is often misplaced or partially obscured by peripheral noise or background clutter. By injecting Gaussian foveated focus masks, we established a deterministic compositional basin that guides the model's attention toward the center of the frame.

In the generation of complex subjects, such as astronauts or gothic cathedral facades, this central-focus injection reduced peripheral variance by 12.3%, ensuring that the primary subject remains compositionally dominant. In the Cathedral case study, the combination of axial symmetry and focal injection allowed for the deterministic placement of the rose window and arches, demonstrating that PLPI can coordinate multi-scale geometric constraints without textual prompting.

Summary of generalization metrics

As detailed in Extended Data Table 2, PLPI demonstrates a universal steering effect. Across all domains, the improvement in structural alignment remained within a tight 1.5% variance band, proving that the Plasticity Window (timesteps 10–20) and the Lock-and-Key mechanism are fundamental properties of the diffusion manifold rather than artifacts of specific training data. This cross-domain robustness establishes PLPI as a reliable framework for industrial and scientific applications where the unpredictable nature of text-based prompts has historically limited the adoption of generative models.