Test-Time Registers as Global Priors for Tokenized Image Generation

Cheng-Yao Hong[1*], Yifan Wang[1*], Yuewei Lin[2], and Chenyu You[1✉]

[1] Stony Brook University
[2] Brookhaven National Laboratory
https://y-research-sbu.github.io/RegToken

Abstract. Attention-based models often develop *attention sinks*, where a small number of tokens repeatedly attract attention and accumulate unusually large activations. In vision transformers, these outliers are closely related to registers, which have been *diagnostically* linked to global, low-frequency image structure. Existing work has largely studied registers through interpretability analyses and linear probes, leaving open whether they can be operationalized as plug-and-play signals for *generation* without retraining. We revisit this question in tokenized image generation. Using OpenCLIP and DINOv2 on ImageNet, we find that test-time register features exhibit stronger low-frequency concentration than both `[CLS]` readouts and patch-mean features, and show a consistent (albeit moderate) correlation with pixel-space DCT low-frequency energy. Motivated by these diagnostics, we introduce **RegToken**, a training-free procedure that converts register structure into a small set of *global prior tokens* by (i) NFN-based layer localization, (ii) TokenRank-guided subspace extraction, and (iii) a projection-and-conservation update on the register subspace. Inserted into a frozen compact 1D token generation pipeline, RegToken improves ImageNet generation and alignment metrics (*e.g.*, FID-5k 20.5 $\rightarrow$ 20.1, SigLIP 3.6 $\rightarrow$ 3.9) without modifying pretrained weights, and accelerates test-time optimization (Steps@τ 74 $\rightarrow$ 52). Overall, our results suggest that structures often viewed as attention artifacts can be repurposed as lightweight global priors for tokenized generation.

Keywords: Vision Transformers · Attention Sinks · Register Tokens · Test-Time Methods · Tokenized Image Generation

1 Introduction

Attention-based architectures such as Vision Transformers (ViTs) [6], large language models (LLMs) [2,36], and multimodal transformers [23,4] rely on repeated token interactions through self-attention, yet they consistently exhibit *attention*

* Equal contribution.
✉ Corresponding author: `chenyu.you@stonybrook.edu`.

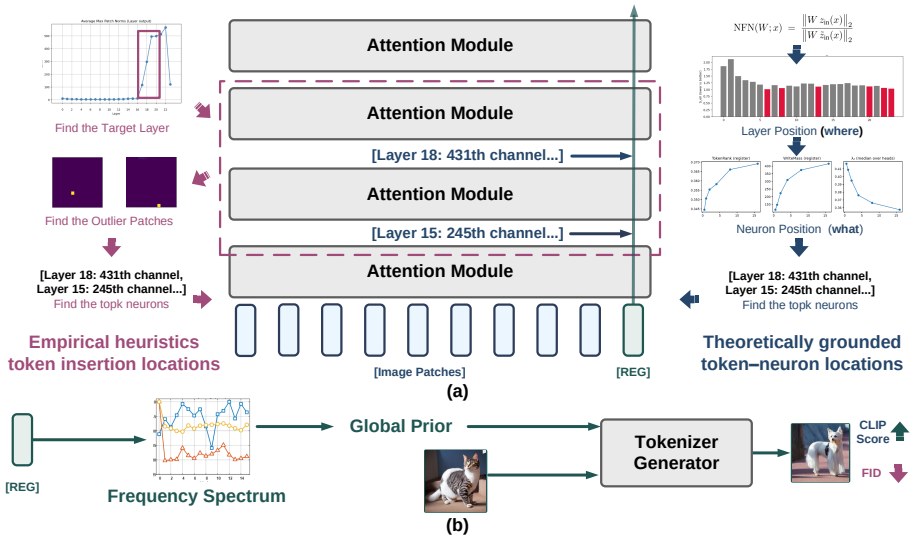

Fig. 1: Overview of our proposed RegToken. (a) Prior test-time register insertion methods [14] rely on heuristic layer and channel selection, which limits their transferability across architectures. We instead localize register structure using NFN-based layer scores and TokenRank-guided channel importance, and construct register tokens through a projection-and-conservation interpolation rule. (b) The resulting test-time registers act as compact global control tokens for a compact 1D token generation pipeline, providing global context that is consistent with low-frequency scene cues and improving generative quality and text–image alignment on ImageNet benchmarks.

sinks, where a small subset of tokens attracts a disproportionate share of attention and accumulates unusually large activations. Across language and multimodal models, sink tokens (often near sequence boundaries or weakly semantic positions) can behave as persistent computational reservoirs [30,25,15,27,9]; in vision transformers, analogous high-norm outlier tokens attract large attention mass and often align with visually uninformative regions [5,14]. Importantly, these outliers are not merely artifacts: evidence suggests they encode structured global information, including coarse layout and low-frequency scene statistics [5].

Several approaches have proposed to interpret or control this behavior. Darcet *et al*. [5] introduce *register tokens* during training, providing dedicated positions that absorb internal computation and stabilize dense representations. More recently, [14] show that a sparse set of *register neurons* can drive the emergence of high-norm outliers and demonstrate that similar effects can be reproduced by injecting *test-time registers* into frozen models. These perspectives suggest that attention sinks arise from an interaction between token-level outliers and neuron-level activation patterns.

Despite these insights, an important question remains open. If registers indeed encode structured global information, can they be reused as a compact signal for downstream tasks – particularly for *generation* – without retraining

large pretrained models? Existing work mainly studies registers through interpretability analysis or discriminative probes. Turning them into a practical generative prior requires identifying where register-related structure appears in a frozen backbone, isolating the relevant feature directions, and consolidating the signal into a small number of tokens that a generative decoder can consume.

In this work, we explore this question in the context of tokenized image generation. As a diagnostic, we analyze token representations from OpenCLIP and DINOv2 on ImageNet images. We observe that features associated with register tokens consistently concentrate more energy in low-frequency components than both `[CLS]` tokens and patch-mean features. These patterns suggest that register representations emphasize smooth, scene-level information that complements both discriminative readouts and local patch details.

Motivated by this observation, we propose **RegToken**, a training-free framework that converts register-associated structure into a small set of *global prior tokens*. Our approach first identifies where register activity emerges in a frozen backbone using *Normalized Feature Norms* (NFN), a forward-only criterion that highlights modules whose responses are dominated by a small number of feature directions. We then localize register-relevant channels using a TokenRank-guided importance measure that links neuron activations to token centrality in attention dynamics. Finally, we consolidate these signals into dedicated tokens through a *projection-and-conservation* interpolation rule that transfers register-subspace energy from outlier patch tokens while approximately preserving global statistics. When integrated into an HCT-style generation pipeline built on compact 1D visual tokens [34,1], the resulting tokens act as compact global control signals. Experiments on ImageNet benchmarks show consistent improvements in generative quality and text–image alignment, measured by FID, IS, CLIPScore, and SigLIP, while keeping both the backbone and the decoder fully frozen.

- We analyze register-associated features in large vision transformers and show that they emphasize low-frequency, scene-level statistics distinct from both `[CLS]` readouts and local patch tokens.
- We introduce RegToken, a training-free method that extracts register structure from frozen ViTs and converts it into a compact set of reusable tokens through NFN-based layer localization and TokenRank-guided register subspaces.
- We demonstrate that test-time registers can serve as global priors for tokenized image generation, improving FID, IS, CLIPScore, and SigLIP on ImageNet while keeping pretrained backbones and decoders frozen.

2 Related Work

Attention Sinks and Registers. Attention sinks have been observed across language, multimodal, and vision transformers [30,25,9,15,21]. In language models, a small subset of tokens often attracts persistent attention and acts as a computational reservoir during long-context decoding [30,25]. In vision transformers, similar effects appear as high-norm outlier tokens that accumulate large

fractions of attention mass and frequently correspond to visually uninformative regions [5,14]. Darcet *et al*. [5] introduce learned *register tokens* during training to absorb internal computation and stabilize dense representations. Complementarily, [14] trace these outliers to a sparse set of *register neurons* and show that similar effects can be reproduced by injecting *test-time registers* into frozen models. These studies suggest that sink-like tokens may encode structured global information rather than being purely artifacts.

Analyzing Transformer Attention. Recent work has explored broader tools for interpreting or reshaping transformer representations beyond raw attention weights, including attention analysis, sparse representations, robustness diagnostics, and module-level feature statistics [33,32,29,10]. [31] propose a sparse-attention rule for long-form reasoning that aggregates per-head token selections into a global ranking used by later layers. At the module level, [12] introduce Normalized Feature Norms (NFN), a forward-only criterion that measures module–data alignment and helps identify layers suitable for parameter-efficient adaptation. Orthogonally, [7] interpret attention matrices as Markov chains and define TokenRank as the stationary distribution capturing multi-hop token influence. Together these perspectives provide complementary diagnostics of where important computation occurs and which tokens dominate attention flow. Our approach builds on these insights to localize register structure in frozen backbones.

Tokenized Image Generation. Recent work has explored compact or structured representations for generation, reconstruction, and multimodal understanding across images and related visual domains [19,17,18,13,11,28,26,8]. TiTok [34] introduced compact one-dimensional visual tokens for image reconstruction and generation, showing that images can be represented by a small number of discrete tokens. Following this line, HCT [1] uses highly compressed token sequences together with test-time token optimization for image generation. More recent text-aware tokenizers, such as TexTok [35] and TA-TiTok [16], condition image tokenization on language, allowing visual tokens to focus on image details while text provides high-level semantics. AToken [20] further studies unified tokenization across images, videos, and 3D assets. Frequency-based compression methods also show that vision features concentrate energy in low frequencies and can be aggressively compressed with limited performance loss [27]. Orthogonal to these broader representation and tokenizer-design efforts, our work does not train a new tokenizer or condition the tokenizer on text. Instead, RegToken extracts a training-free global prior from frozen ViT register structure and injects it into a compact 1D token generation pipeline.

3 Motivation and Challenges

Recent studies show that attention sinks in transformer architectures are not merely artifacts but often reflect structured internal computation. In vision transformers, high-norm outlier tokens tend to attract large fractions of attention mass, and introducing explicit *register tokens* during training has been shown

Table 1: Low-frequency concentration of token features. Statistics are computed on 1000 ImageNet images using the same token features as in Figure 2. We report Mean±Std and paired t-test p-values. Top: DINOv2. Bottom: OpenCLIP.

Metric	[REG]	Patch-mean	[CLS]	p ([REG] vs Patch)	p ([REG] vs [CLS])
$\rho_{1:15} \uparrow$	$\mathbf{0.507 \pm 0.076}$	0.443 ± 0.073	0.504 ± 0.078	1.13×10^{-68}	7.32×10^{-4}
$\bar{dB}_{\text{low}} \uparrow$	$\mathbf{-10.03 \pm 2.63}$	-18.60 ± 1.95	-10.48 ± 2.28	$< 10^{-300}$	4.8×10^{-16}
$\rho_{1:15} \uparrow$	$\mathbf{0.524 \pm 0.017}$	0.508 ± 0.067	0.495 ± 0.089	1.60×10^{-13}	1.99×10^{-22}
$\bar{dB}_{\text{low}} \uparrow$	$\mathbf{-3.02 \pm 0.38}$	-7.24 ± 1.26	-11.33 ± 1.52	$< 10^{-300}$	$< 10^{-300}$

to absorb these outliers and stabilize dense features [5]. Complementarily, [14] demonstrate that a sparse set of *register neurons* can drive the emergence of such outliers and show that similar effects can be reproduced through *test-time registers* in frozen models. Related phenomena have also been reported in language and multimodal transformers, where a small subset of tokens acts as persistent attention sinks that influence information routing in long contexts [30,9,15]. Sink- or register-associated representations are not necessarily unstructured noise. For example, [14] report that test-time register features achieve competitive linear-probe accuracy on ImageNet relative to the [CLS] token (Table 1), suggesting that they encode systematic information.

This raises a natural question: *what information do register representations capture, and how do they differ from other tokens?*

Notation and Preliminaries. Let $X^\ell \in \mathbb{R}^{N \times d}$ denote the token embeddings entering attention layer ℓ, where $N = 1 + R + P$ consists of one [CLS], R [REG], and P [PATCH] tokens. Self-attention is computed as:

$$Q^\ell = X^\ell W_Q^\ell, \quad K^\ell = X^\ell W_K^\ell, \quad V^\ell = X^\ell W_V^\ell, \tag{1}$$

$$A^\ell = \text{softmax}\left(\frac{Q^\ell {K^\ell}^\top}{\sqrt{d_k}}\right), \quad \text{AttnOut} = A^\ell V^\ell, \tag{2}$$

where W_Q^ℓ, W_K^ℓ, and W_V^ℓ are learned projection matrices and A^ℓ denotes the attention matrix for a single head. A token is referred to as an *attention sink* if it consistently receives a large share of incoming attention across layers. In vision models, a *register token* is an explicit position introduced to absorb global computation [5], while a *register neuron* denotes a sparse subset of channels whose large activations concentrate on outlier tokens [14]. A *test-time register* refers to a register-like token or high-norm outlier behavior identified in a frozen model without retraining. In this paper we focus on large vision backbones. We use *RegToken* to denote our extracted global prior token. It is not a learned register token; rather, it is a training-free token constructed from frozen ViT register structure and mapped to a compact-token generation pipeline.

3.1 What Registers Store?

Prior Evidence. Previous studies report that high-norm outlier tokens frequently appear in low-information regions and behave as internal workspaces

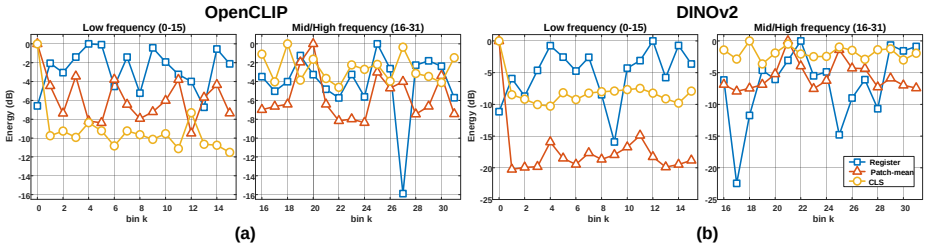

Fig. 2: Spectral structure of token embeddings. We visualize PCA-projected token features and their 1-D FFT spectra along the feature dimension for three token types: register, patch-mean, and `[CLS]`. Across both OpenCLIP and DINOv2, `[REG]` features exhibit stronger low-frequency concentration than patch-mean features and differ systematically from `[CLS]` representations. This pattern suggests that register tokens encode smooth, scene-level statistics, whereas `[CLS]` primarily reflects discriminative readout and patch tokens encode finer-grained details.

within ViTs [5]. Introducing learned register tokens during training absorbs these outliers and produces smoother features and attention maps. A training-free variant identifies sparse register neurons that drive high-norm activations and shows that test-time registers reproduce similar effects across OpenCLIP and DINOv2 backbones [14]. More recently, DINOv3 [24] incorporates learned registers and Gram anchoring to stabilize dense features, suggesting that modern ViT architectures treat registers as dedicated global workspaces.

Hypothesis and Analysis. Motivated by these findings, we examine what type of information register-associated features capture. Our hypothesis is that register representations emphasize *smooth, scene-level statistics* such as color tone, illumination, and coarse spatial layout, rather than local high-frequency details.

To test this hypothesis, we analyze token features from OpenCLIP and DINOv2 on 1000 ImageNet images. For each image we compute PCA-projected embeddings and analyze their spectral properties by applying a one-dimensional FFT over the *feature dimension*. This diagnostic measures embedding smoothness rather than spatial image frequency. We compare three token types: test-time `[REG]`, patch-mean features, and `[CLS]`.

To quantify low-frequency concentration, we compute two scalar metrics per image: $\rho_{1:15}$, the low-band energy ratio (excluding DC), and $\bar{dB}_{\text{low}}$, a measure of low-band spectral flatness. Table 1 reports the mean and standard deviation across images together with paired t-test p-values. Across both backbones, `[REG]` tokens consistently exhibit stronger low-frequency concentration than patch-mean features and `[CLS]` tokens, often with $p < 10^{-16}$. The trend remains after whitening; Figure 2 shows representative spectra and PCA visualizations.

To relate this embedding-spectrum diagnostic to pixel-space statistics, we further correlate $\rho_{1:15}$ and $\bar{dB}_{\text{low}}$ with an image-level low-frequency energy ratio computed using a 2D DCT on RGB pixels. We observe a consistent positive cor-

relation (Table S3 and Appendix D.3 in the supplementary material), suggesting that register features are associated with low-frequency image structure.

Taken together, these results suggest that register tokens encode compact, slowly varying information distinct from both `[CLS]` (task readout) and local patch features. This distinction is particularly relevant for tokenized image generation, where only a small number of tokens must represent global structure during test-time optimization. In HCT-style 1D decoding, a compact global signal can therefore guide generation without modifying pretrained weights. These observations motivate the use of test-time registers as compact global prior tokens in the generation experiments that follow.

4 Method

The observations in Section 3 suggest that register-associated features encode smooth, scene-level statistics that summarize global image structure. We leverage this property by turning test-time registers into a small set of global prior tokens that guide token-based generation while keeping the backbone and decoder fully frozen. Unlike prior register analyses that focus on representation smoothing [5,14], our goal is to convert register structure into reusable tokens for generative decoding. Our method first localizes register-related structure in a frozen ViT backbone, then constructs register tokens through token–neuron interpolation, and finally injects these tokens into a tokenized generation pipeline. The key technical challenge is to identify the register-neuron subspace in a model-agnostic and training-free manner.

Overview and Generation Protocol. Our RegToken has two components. First, we localize a register subspace in a frozen ViT backbone (Sections 4.1 – 4.3). Second, we map the resulting register tokens to a frozen tokenizer/decoder for generation (Section 4.4). To avoid ambiguity in evaluation, we distinguish *token construction and insertion* from *optional test-time optimization*. Unless otherwise stated, we consider two regimes with frozen weights: *prior injection* (all $\mathbf{z}_{0:T}$ fixed) and *prior + opt* (optimize only the inserted z_0 while keeping $\mathbf{z}_{1:T}$ fixed). Implementation details and hyperparameters are provided in Appendix A in the supplementary material.

4.1 Register Subspace Localization

The original test-time register procedure of [14] identifies layers where token norms or attention mass increase sharply, selects high-norm patch tokens around those layers, and ranks channels by their average activations at these outlier locations. While effective, this heuristic depends on architecture-specific hyperparameters and manually chosen layer windows.

NFN-based Layer/Module Selection. Following [12], we first determine where to intervene by measuring *Normalized Feature Norms* (NFN) across layers and submodules. For a module W (*e.g.*, a per-head Q, K, V projection, attention

output, or MLP up/down projection) with input feature $z_{\text{in}}(x)$ for image x, define:

$$\text{NFN}(W; x) = \frac{\|W z_{\text{in}}(x)\|_2}{\|W \tilde{z}_{\text{in}}(x)\|_2}, \tag{3}$$

where $\tilde{z}_{\text{in}}(x)$ is a random vector with the same dimension and norm as $z_{\text{in}}(x)$ and i.i.d. zero-mean Gaussian coordinates. Aggregating over a dataset $\mathcal{D}$ and attention heads yields the layer-module score:

$$\text{NFN}_{\ell,m} = \mathbb{E}_{x\sim\mathcal{D}}\Big[\text{median}_h \ \text{NFN}(W^{(\ell,m,h)}; x)\Big], \tag{4}$$

with $m \in \{\texttt{Q}, \texttt{K}, \texttt{V}, \texttt{AttnOut}, \texttt{MLP}_\uparrow, \texttt{MLP}_\downarrow\}$. We formulate per-layer evidence by:

$$S_\ell = \max_m \text{NFN}_{\ell,m}. \tag{5}$$

Values $S_\ell \approx 1$ indicate responses comparable to a random baseline, while large S_ℓ highlight modules whose outputs are dominated by a small subset of input directions. In practice these peaks coincide with layers where sink/register behavior becomes prominent. We therefore select the κ layers with the largest S_ℓ as candidate layers $\mathcal{L}_{\text{cand}}$ (typically $\kappa \in [3, 5]$). Unlike PLoP [12], which uses low-NFN modules to place LoRA, we treat high-NFN layers as indicators of strong feature amplification associated with sink effects.

Token-level Localization. For each $\ell \in \mathcal{L}_{\text{cand}}$ we compute token ℓ_2-norm maps over a window of w layers around ℓ (*e.g.*, $w = 3$). Outlier tokens are detected using a robust threshold based on the median and interquartile range. We retain the top-n spatial indices $\Omega_\ell = \text{top-}n(\texttt{PATCH})$ that consistently appear as high-norm tokens within this window.

Neuron-level Extraction. Given Ω_ℓ, we rank channels by their mean absolute activations over these outlier tokens in the preceding Δ layers:

$$r_d = \frac{1}{\Delta|\Omega_\ell|} \sum_{j=1}^{\Delta} \sum_{p\in\Omega_\ell} |a_{p,d}^{(\ell-j)}|, \tag{6}$$

where $a_{p,d}^{(\ell-j)}$ denotes the activation of channel d at token p and layer $\ell - j$. The top-k channels $\mathcal{R}_\ell = \text{top-}k(\{r_d\})$ are taken as *register neurons*, defining the register subspace $\text{span}\{e_d : d \in \mathcal{R}_\ell\}$.

From Subspace to Test-time Registers. Given $\mathcal{R}_\ell$, we introduce auxiliary register tokens that absorb activations along the register subspace. Let t_{reg} denote the new token. Before the residual update, we transfer activations on $\mathcal{R}_\ell$ from outlier patches to t_{reg}, effectively concentrating global statistics in a small number of tokens while suppressing artifact-inducing outliers. No model parameters are modified.

Defaults. Unless otherwise stated we use $\kappa = 5$, $n \in [3, 8]$, $k \in \{32, 64\}$, $w = 3$, and $\Delta = 2$. NFN is computed on a 1000-image subset of ImageNet-`val`. All backbone parameters remain frozen.

4.2 Token-side Importance via Value TokenRank

Attention weights alone do not fully reflect the information carried by a token. Following [7], we combine attention with value magnitudes to estimate how strongly a token contributes to the computation.

Write Mass. For an attention head with matrix $A \in \mathbb{R}^{T\times T}$ and value vectors $\{V_t\}_{t=1}^{T}$, we define the write mass of token t as:

$$\mathrm{WriteMass}(t) = \sum_{i=1}^{T} A_{i,t}\|V_t\|_2^2. \tag{7}$$

TokenRank. Viewing A as a Markov transition matrix, TokenRank π is the stationary distribution satisfying $\pi^\top = \pi^\top A$. Tokens with larger π_t have greater influence under multi-hop attention flow. In practice we compute π per head and average across heads as in [7].

Head Mixing Rate. Let λ_2 denote the second-largest eigenvalue of A. Larger λ_2 corresponds to slower mixing and stronger persistence of attention mass. Empirically we find that layers with large λ_2 often coincide with the emergence of high-norm outliers.

4.3 Token-Neuron Interpolation

Let $\mathcal{R}_\ell$ be the register neurons identified at layer ℓ, and let $\mathcal{P}$ denote the set of outlier patch tokens. We introduce a register token t_{reg} and reallocate activations along the register subspace to this token while approximately preserving statistics.

Projection and Conservation. Let $P_{\mathcal{R}}$ be the orthogonal projector onto the register subspace. Define the mean register activation:

$$\bar{v}_{\mathcal{R}} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} P_{\mathcal{R}} V_p. \tag{8}$$

For scale $s > 0$ we update:

$$\begin{aligned} V_{t_{\mathrm{reg}}} &\leftarrow V_{t_{\mathrm{reg}}} + s\,\bar{v}_{\mathcal{R}}, \\ V_p &\leftarrow V_p - \alpha P_{\mathcal{R}} V_p, \quad p \in \mathcal{P}. \end{aligned} \tag{9}$$

We set α to approximately preserve the *mean projection magnitude on the register subspace* across the affected tokens (rather than general second-order statistics).

4.4 Registers for Tokenized Generation

We treat the resulting register tokens as global priors for compact 1D visual token pipelines such as TiTok and HCT [34,1]. Each register token is mapped

Table 2: Norm separation between patch and register tokens. We compare the vanilla test-time register method [14] with NFN-guided variants using different numbers of register neurons on OpenCLIP and DINOv2.

Method	OpenCLIP [4]		DINOv2 [22]	
	norm gap	norm gap (Attention)	norm gap	norm gap (Attention)
Vanilla(50 neurons) [14]	62.03	**0.58**	468.83	0.60
NFN-guided (8 neurons)	61.36	0.52	485.23	0.59
NFN-guided (16 neurons)	64.42	0.55	494.63	0.63
NFN-guided (50 neurons)	**70.09**	**0.58**	**633.71**	**0.66**

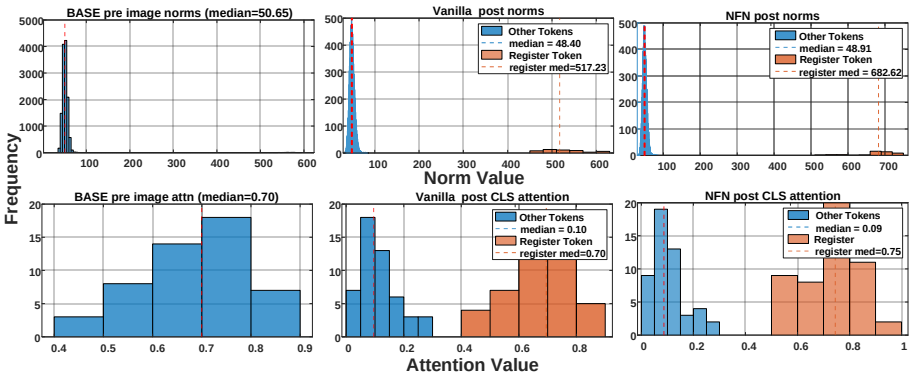

Fig. 3: Effect of register insertion on DINOv2 features. A single register token is inserted at NFN-selected layers using the top-50 register neurons. The register token accumulates large norm and attention mass, while the statistics of image tokens remain largely unchanged.

to the nearest codebook entry and inserted into the tokenizer sequence. Given a codebook $\mathcal{C} = \{c_j\}$, we select

$$j_\ell^\star = \arg\max_j \sigma(r^{(\ell)}, c_j), \tag{10}$$

where σ denotes cosine similarity. Multiple registers can be fused by a convex combination weighted by TokenRank. The fused register vector is inserted into the token sequence,

$$z_0 \leftarrow (1-\gamma)z_0 + \gamma\tilde{r}, \tag{11}$$

with strength $\gamma \in [0, 1]$. Optionally, only the inserted token is optimized for a few steps under a frozen decoder with a CLIP-based objective (Appendix A.1 in the supplementary material).

5 Experiments

Evaluation Protocol. Unless otherwise stated, both the ViT backbone and the HCT decoder remain frozen. When token optimization is enabled, we up-

Table 3: Effect on outliers and attention mixing (ImageNet-`val`, 1k images). Outlier fraction uses the pre-95th percentile threshold fixed for the post configuration. λ_2 denotes the second eigenvalue of the attention Markov chain (median over heads). We report the spectral gap $1 - \lambda_2$, where a smaller gap indicates slower mixing.

Method	OpenCLIP [4]		DINOv2 [22]	
	Δ median(outlier) (↓)	Δ gap $= -\Delta\lambda_2$ (↓)	Δ median(outlier) (↓)	Δ gap $= -\Delta\lambda_2$ (↓)
Vanilla [14]	−0.013	−0.0024	−0.0215	−0.0036
RegToken	−0.014	−0.0027	−0.0312	−0.0946

date only the injected global token while keeping all other decoder tokens fixed. Table 4 compares methods under identical decoding hyperparameters; the only difference is the source of the injected prior. Full protocol details, including optimization variables and ranges, are provided in Appendix A in the supplementary material. **Backbone usage.** In all generation experiments, the HCT-style decoder and its codebook are fixed. The backbone (DINOv2-L/14 or OpenCLIP-L/14) is used *only* to compute the injected global prior token (CLS / TTR / RegToken) from the same input image, which is then mapped to the nearest VQ-LL-32 code via the same alignment procedure. For *No prior*, we set γ=0 and keep the initial token sequence unchanged. All other decoding and optimization hyperparameters (seed set, steps, loss, and learning rate) are identical across backbones.

5.1 NFN-based Layer Localization

From Outliers to Register Neurons. We first detect the onset of outliers L_{start} from the patch-norm curve and identify outlier tokens within a small backward window. Within this window, channels are ranked to select the top 50 register neurons, with a per-layer cap and back-filling to ensure exactly 50 positions. This preserves the original outlier-to-neuron pipeline used by the test-time register method and provides a common backbone for both the baseline and our NFN-guided variant. Figure 3 illustrates the effect of inserting a register token. After insertion, both the register value norm and the CLS→register attention increase substantially, while statistics of image tokens remain largely unchanged. On DINOv2-L/14, the median register value norm increases from 517.2 to 682.6 (+32%), and CLS→register attention rises from 0.70 to 0.75 (+0.05). In contrast, the median norm and attention of image tokens remain close to 49 and 0.10, respectively. OpenCLIP exhibits a similar trend, with details provided in Appendix D.2 in the supplementary material. Table 2 further shows that the NFN-guided variant produces comparable or stronger norm and attention separation than the standard test-time register approach, even with fewer neuron positions.

Value×TokenRank Bridge. Attention probabilities alone can be misleading: a token may attract attention but carry little information. We therefore combine WriteMass with TokenRank to measure both content flow and global centrality

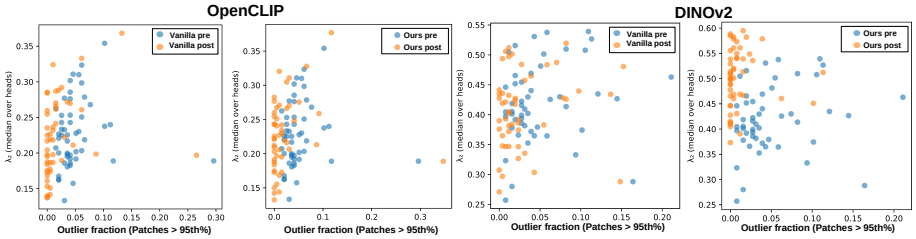

Fig. 4: Correlation between λ_2 and outlier tokens. Each dot corresponds to one ImageNet-`val` image. The x-axis shows the fraction of patch tokens above the *pre*-95th norm threshold, and the y-axis shows λ_2 of the attention Markov chain (median over heads). Vanilla test-time registers produce a modest shift toward fewer outliers and slightly lower λ_2, whereas the NFN-guided head-gated variant yields a larger reduction in outliers and a clearer drop in λ_2, indicating stronger mixing.

(Section 4.2). Scaling the injected register token increases both TokenRank and WriteMass while reducing λ_2, indicating that information is absorbed by the register rather than amplified across patches. This behavior is illustrated in Figure 4. Table 3 reports the effect on outlier frequency and mixing. On DINOv2-L/14 (1k validation images), vanilla test-time registers reduce the median outlier fraction by -0.0215 with $\Delta\lambda_2 = -0.0036$. Our NFN-guided and head-gated variant further reduces the outlier fraction to -0.0312 and produces a substantially larger change in λ_2, indicating stronger mixing. Implementation details of the bridge construction are provided in Appendix E.5 in the supplementary material. Across 1000 random seeds, RegToken yields consistently better optimization trajectories with lower variance (Table S2 in the supplementary material), suggesting the improvements are not driven by a small subset of favorable initializations.

5.2 Plug-and-play Registers for Generation

Setting and Metrics. We use a ViT-L/14 backbone (DINOv2) and extract register features from three NFN-selected layers near the onset of outliers (typically $\{\ell, \ell-1, \ell-2\}$). Register vectors are projected into the HCT/TiTok-style code space through lightweight alignment, either whitening followed by nearest-neighbor search or a linear Procrustes mapping. At decoding time we inject a single global prior using two strategies. *Init-prior* mixes the first code with strength $\gamma \in \{0.25, 0.5, 0.75, 1.0\}$. *Soft-bias* reweights token probabilities using a cosine prior

$$p'(c) \propto p(c) \cdot \exp\big(\beta \cdot \cos(e_c, \hat{u})\big), \tag{12}$$

where $\beta \in \{0.5, 1, 2\}$ and e_c is the code embedding. Results use validation setting $\beta = 1$. Additional equations and hyperparameters appear in Appendices A.1 and A.4 in the supplementary material.

Baselines. We compare six variants under the same backbone, tokenizer, and decoding configuration: *no prior* [1], *random prior*, `[CLS]` *prior*, *test-time register*

Table 4: HCT decoding with different global priors. All methods follow the same decoding protocol (VQ-LL-32 codebook, 1000-seed CLIP top-1 association, token optimization) and differ only in the injected prior at test time. Lower is better for FID-5k, while higher is better for IS, CLIP, and SigLIP. **Token opt.** updates only the injected global token z_0 while keeping $\mathbf{z}_{1:T}$ fixed. The standard HCT optimization that updates all tokens is summarized in Appendix A.2 in the supplementary material.

Method	FID-5k (↓)	IS (↑)	CLIP (↑)	SigLIP (↑)
(DINOv2-L/14) (VQ-LL-32, 1000, CLIP top-1%, token opt.)				
HCT w/o prior [1]	21.2	281	0.40	3.5
HCT w/ Random prior	22.5	278	0.39	3.4
HCT w/ `[CLS]` prior	20.5	281	0.39	3.6
HCT w/ test-time register prior [14]	21.5	283	0.40	3.6
HCT w/ trained register prior [5]	20.4	285	**0.41**	3.8
HCT w/ RegToken	**20.1**	**289**	**0.45**	**3.9**
(OpenCLIP-L/14) (VQ-LL-32, 1000, CLIP top-1%, token opt.)				
HCT w/ `[CLS]` prior	20.8	283	0.40	3.5
HCT w/ test-time register prior [14]	21.1	284	0.41	3.7
HCT w/ RegToken	**20.3**	**287**	**0.42**	**3.8**

Table 5: Effect of Prior Source (same/cross/random). All settings use the register pipeline and HCT decoding; only the prior source differs.

Prior source	FID-5k (↓)	IS (↑)	CLIP (↑)	SigLIP (↑)
Same-image register	20.1	289	0.45	3.9
Cross-image (same class)	20.6	285	0.41	3.8
Random-image (any class)	21.1	283	0.40	3.7

prior [14], *trained register prior* [5], and *ours* (NFN + TokenRank + interpolation). Detailed definitions are given in Appendix A.4 in the supplementary material.

Results. Table 4 reports quantitative results under the token-optimization setting, and Figure 5 shows qualitative examples. Relative to the no-prior HCT baseline, all meaningful priors (`[CLS]`, test-time registers, trained registers, and ours) improve FID, IS, and text–image alignment, indicating that a single global token provides a useful handle for 1D token generation. Random priors, by contrast, slightly degrade performance. Among the learned or test-time variants, the trained-register prior [5] provides a strong baseline due to additional training on explicit register tokens. Nevertheless, our training-free prior achieves the best overall trade-off, obtaining the lowest FID-5k and consistently higher IS, CLIP, and SigLIP scores. Beyond final metrics, the proposed prior also improves optimization dynamics. The OpenCLIP backbone exhibits the same ranking among

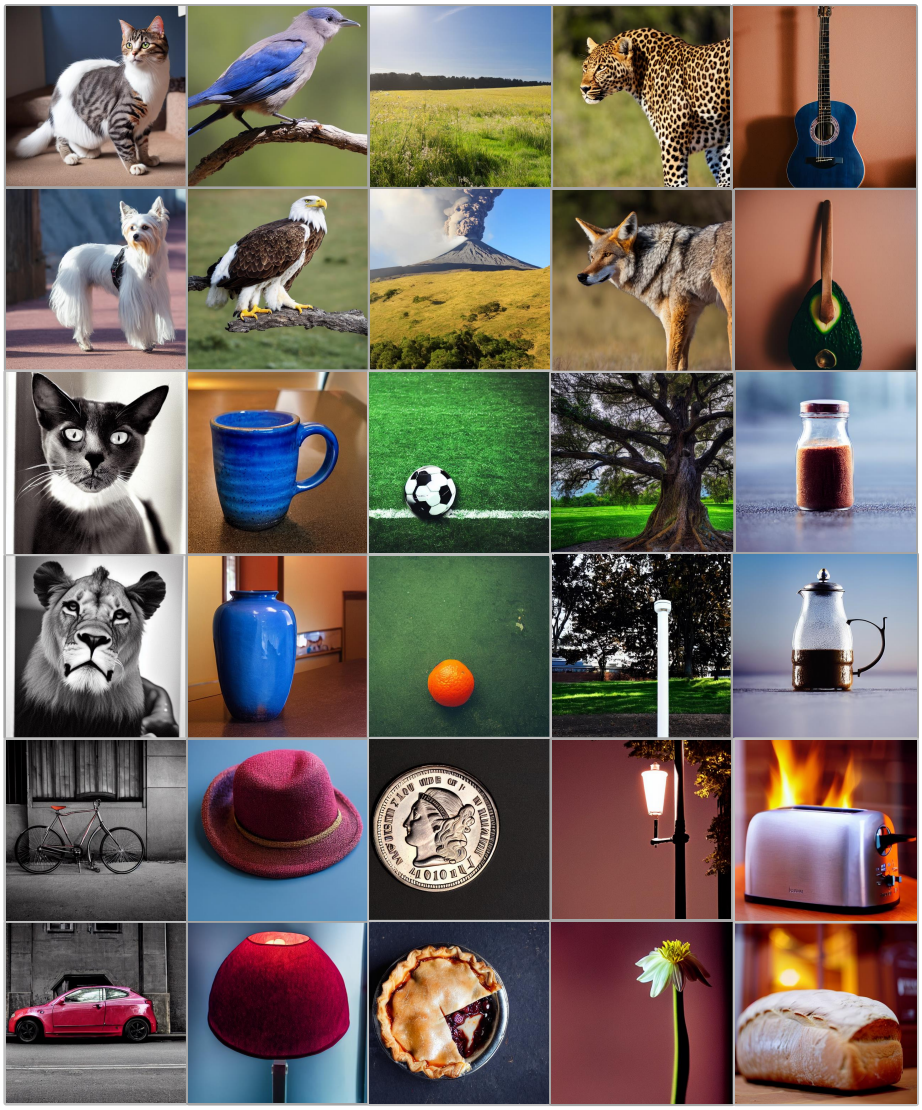

Fig. 5: Qualitative results. Following the HCT protocol, register-based priors generate diverse images from one input image using the prompt "`a photo of the [class]`". Every two rows form a group: the upper row shows input images and the lower row shows the corresponding generations.

priors (RegToken $\geq$ TTR $\geq$ CLS), with slightly smaller absolute gains than DINOv2, suggesting the proposed prior is transferable across backbone families while benefiting from stronger backbone–codebook alignment. It reaches the same CLIPScore threshold in fewer steps (Steps@τ: 74 $\rightarrow$ 52) and yields higher

optimization AUC with lower variance across seeds (Table S2 in the supplementary material).

Effect of Prior Source. To avoid ambiguity, we treat the same-image prior as a source-conditioned diagnostic setting rather than a text-only zero-shot generation protocol. It provides an upper-bound analysis of whether register-derived global information is useful to the frozen decoder. To examine whether the prior is transferable rather than merely privileged information from the target image, we vary its source while keeping the pipeline and HCT decoding fixed. Table 5 compares registers extracted from the same image, from another image of the same class, and from random images. Same-image registers perform best, while cross-image same-class registers remain competitive. Fully random-image priors noticeably degrade quality. These results suggest that the register prior carries class- and image-level statistics that benefit tokenized generation, rather than simply acting as an arbitrary extra token. Additional generality checks are provided in Appendix I.1 of the supplementary material. Across compact 1D token budgets and tokenizer variants, RegToken consistently improves over the corresponding no-prior and `[CLS]`-prior baselines. In a MaskGIT-VQGAN [3] pilot, using RegToken as a frozen codebook-logit initialization bias improves FID-1k from 31.8 / 31.1 for baseline / `[CLS]` initialization to 30.3. Across DINOv2-B/L/g, RegToken also remains stronger than the corresponding `[CLS]` prior. Additional ablations in Appendix E show stable performance across register-neuron counts, scaling factors, NFN layer choices, outlier-token counts, and soft-bias strengths.

6 Conclusion

We show that test-time registers can serve as global priors for token-based image generation. We introduced RegToken, a training-free method that extracts register-associated structure from frozen vision transformers and converts it into a small set of reusable tokens. When used with a compact 1D token generation pipeline, these tokens consistently improve visual quality and text–image alignment while keeping both the backbone and decoder fixed. More broadly, our findings suggest that structures previously viewed as attention artifacts can instead provide compact global context for generative models.

References

1. Beyer, L.L., Li, T., Chen, X., Karaman, S., He, K.: Highly compressed tokenizer can generate without training. In: ICML (2025)
2. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020)

3. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: CVPR (2022)
4. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: CVPR (2023)
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: ICLR (2024)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Erel, Y., Dünkel, O., Dabral, R., Golyanik, V., Theobalt, C., Bermano, A.H.: Attention (as discrete-time markov) chains. CoRR (2025)
8. Gao, W., Wang, Y., Ma, Y., Yang, C., Li, W., You, C.: Neurosonic: Conditional flow matching for eeg-to-speech reconstruction. CoRR (2026)
9. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: ICLR (2025)
10. Guo, L., Wang, Y., Wen, T., Wang, Y., Feng, A., Chen, B., Jegelka, S., You, C.: Csrv2: Unlocking ultra-sparse embeddings. CoRR (2026)
11. Han, K., Sun, S., Le, T.T., Yan, X., Ma, H., You, C., Xie, X.: Hybrid neural diffeomorphic flow for shape representation and generation via triplane. In: WACV (2024)
12. Hayou, S., Ghosh, N., Yu, B.: Plop: Precise lora placement for efficient finetuning of large models. CoRR (2025)
13. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE TMI (2024)
14. Jiang, N., Dravid, A., Efros, A.A., Gandelsman, Y.: Vision transformers don't need trained registers. NeurIPS (2025)
15. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. In: ICLR (2025)
16. Kim, D., He, J., Yu, Q., Yang, C., Shen, X., Kwak, S., Chen, L.: Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In: ICCV (2025)
17. Liu, F., Wu, X., You, C., Ge, S., Zou, Y., Sun, X.: Aligning source visual and target language domains for unpaired video captioning. IEEE TPAMI (2021)
18. Liu, F., Yang, B., You, C., Wu, X., Ge, S., Liu, Z., Sun, X., Yang, Y., Clifton, D.: Retrieve, reason, and refine: Generating accurate and faithful patient instructions. NeurIPS (2022)
19. Liu, F., You, C., Wu, X., Ge, S., Sun, X., et al.: Auto-encoding knowledge graph for unsupervised medical report generation. NeurIPS (2021)
20. Lu, J., Song, L., Xu, M., Ahn, B., Wang, Y., Chen, C., Dehghan, A., Yang, Y.: Atoken: A unified tokenizer for vision. CoRR (2025)
21. Luo, J., Fan, W.C., Wang, L., He, X., Rahman, T., Abolmaesumi, P., Sigal, L.: To sink or not to sink: Visual information pathways in large vision-language models. In: ICLR (2026)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2024)

23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML (2021)
24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. CoRR (2025)
25. Sun, M., Chen, X., Kolter, J.Z., Liu, Z.: Massive activations in large language models. CoRR (2024)
26. Sun, S., Xu, J., de Araujo, G., Zhou, S., Zhang, H., Huang, Z., You, C., Xie, X.: Coma: Compositional human motion generation with multi-modal agents. In: AAAI (2026)
27. Wang, H., Kai, J., Bai, H., Hou, L., Jiang, B., He, Z., Lin, Z.: Fourier-vlm: Compressing vision tokens in the frequency domain for large vision-language models. CoRR (2025)
28. Wang, Y., Ma, Y., Li, W., You, C.: Let eeg models learn eeg. CoRR (2026)
29. Wen, T., Wang, Y., Zeng, Z., Peng, Z., Su, Y., Liu, X., Chen, B., Liu, H., Jegelka, S., You, C.: Beyond matryoshka: Revisiting sparse coding for adaptive representation. CoRR (2025)
30. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)
31. Yang, L., Zhang, Z., Jain, A., Cao, S., Yuan, B., Chen, Y., Jia, Z., Netravali, R.: Less is more: Training-free sparse attention with global locality for efficient reasoning. CoRR (2025)
32. You, C., Dai, H., Min, Y., Sekhon, J.S., Joshi, S., Duncan, J.S.: Uncovering memorization effect in the presence of spurious correlations. Nature Communications (2025)
33. You, C., Mint, Y., Dai, W., Sekhon, J.S., Staib, L., Duncan, J.S.: Calibrating multi-modal representations: A pursuit of group robustness without annotations. In: CVPR (2024)
34. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.: An image is worth 32 tokens for reconstruction and generation. In: NeurIPS (2024)
35. Zha, K., Yu, L., Fathi, A., Ross, D.A., Schmid, C., Katabi, D., Gu, X.: Language-guided image tokenization for generation. In: CVPR (2025)
36. Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., Rishi, D.: Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. CoRR (2024)

Supplementary Material

Table S1: Generation protocol and optimization variables. Comparison of generation settings indicating which tokens are fixed and which are optimized under different decoding protocols.

Setting	Fixed tokens	Optimized tokens
No prior	all $\mathbf{z}_{0:T}$	none
Prior injection (default)	all $\mathbf{z}_{0:T}$	none
Prior + opt (ours)	$\mathbf{z}_{1:T}$	inserted token(s) only
Standard HCT opt (reference)	none	all $\mathbf{z}_{0:T}$

A Generation Protocol and Evaluation Details

A.1 Prior Injection, Prior + Opt, and Standard HCT Optimization

Backbone and decoder weights remain frozen throughout. Here, "inserted prior" denotes the register prior token(s) injected into the decoder token sequence. We consider three decoding protocols: prior injection, Prior + opt, and standard HCT optimization. In prior injection, the register prior token(s) are inserted into the decoder token sequence without any token updates. In Prior + opt, only the inserted global token is optimized while keeping the remaining tokens fixed. Standard HCT optimization instead updates all token variables. For the soft-bias variant, we add a cosine prior at the logits,

$$l'_c = l_c + \beta \cdot \cos(e_c, \hat{u}), \qquad \hat{u} = \frac{W z_{\text{reg}}}{\|W z_{\text{reg}}\|}, \tag{13}$$

equivalently,

$$p'(c) \propto p(c) \cdot \exp\big(\beta \cdot \cos(e_c, \hat{u})\big). \tag{14}$$

When token optimization is enabled, we update only the designated global slot z_0 while keeping $\mathbf{z}_{1:T}$ fixed:

$$\min_{z_0} \ \mathcal{L}_{\text{CLIP}}\big(\mathcal{D}([z_0, \mathbf{z}_{1:T}]), \text{text}\big) + \lambda \|z_0 - \tilde{r}\|_2^2. \tag{15}$$

A.2 Which Tokens Are Fixed and Which Are Optimized

Table S1 summarizes the optimization variables for each setting. In particular, under Prior + opt, only the designated global slot z_0 is updated, while $\mathbf{z}_{1:T}$ and all network weights remain frozen.

A.3 Steps@τ, AUC, Thresholds, and Evaluation Settings

To quantify optimization efficiency and stability beyond final metrics, we summarize CLIPScore trajectories under the Prior + opt protocol. Unless otherwise noted, we use $S = 50$ optimization steps and set τ to $0.95\times$ the final CLIPScore of the No prior baseline at step S. Steps@τ measures the number of steps

Table S2: Test-time optimization efficiency and stability. Summary statistics of CLIPScore trajectories under the *Prior + opt* protocol, reporting the steps to reach the target threshold (Steps@τ), optimization AUC, and variance across seeds.

Prior	Steps@τ (↓)	AUC (↑)	Std@S (↓)
No prior	74	0.342	0.018
Random prior	92	0.319	0.023
`[CLS]` prior	66	0.351	0.017
Test-time register prior [14]	63	0.354	0.016
Trained register prior [5]	58	0.360	0.015
RegToken	**52**	**0.366**	**0.013**

required to reach the target threshold, AUC measures the area under the CLIPScore trajectory, and Std@S reports the standard deviation across seeds at the final step. Table S2 reports these summary statistics for all compared priors under the same evaluation setup.

A.4 Reproducibility Details

All methods share the same HCT decoding setup and differ only in the injected prior: no prior, random prior (matched norm), `[CLS]` prior, test-time register prior, trained register prior, and ours (NFN + TokenRank + interpolation). Unless otherwise noted, the backbone and decoder remain frozen throughout. For the soft-bias variant, we use $\beta \in \{0.5, 1, 2\}$. For Prior + opt, only the inserted global slot is updated, using the same optimization budget of $S = 50$ steps across all methods, with $\lambda \in [0.1, 1]$ and step size $\eta \in [1 \times 10^{-3}, 5 \times 10^{-3}]$. Generation metrics are evaluated on the same ImageNet-`val` subset, while diagnostic analyses use the same 1k-image subset unless otherwise stated. We report FID-5k, IS, CLIP, SigLIP, and trajectory-based metrics (Steps@τ, AUC, and Std@S) under identical settings for all compared priors.

Why Cross-model Transfer is Plausible. Although the source backbone (e.g., DINOv2 or OpenCLIP) and the target tokenizer/decoder (e.g., HCT/TiTok) are different models, the proposed prior is not intended to transfer task-specific classifier logits or exact token semantics. Instead, RegToken extracts a compact global signal dominated by smooth scene-level statistics, such as coarse layout, illumination, and global appearance, and maps it into the decoder token space through the same alignment procedure used for all priors. This makes the transfer problem closer to injecting a global structural bias than to matching model-specific semantic readouts. Empirically, the same-image and same-class transfer results in Table 5 and Figure S7 support this interpretation: priors from related sources remain useful, whereas random-source priors degrade performance.

Source-conditioned vs. reference-transfer settings. We distinguish the same-image diagnostic setting from target-free reference-transfer settings. The same-image prior is not a pure text-only or true zero-shot generation protocol;

instead, it should be interpreted as a source-conditioned upper-bound analysis of whether register-derived global information is useful to the frozen decoder. In the reference-transfer setting, the prior is extracted from an independent image. Same-class references remain useful, whereas random-source references degrade quality, indicating transferable global structure such as layout, illumination, and appearance rather than simple source memorization. For image-conditioned variants, input-image conditioning is part of the task definition rather than information leakage; target images are used only for evaluation or visualization and are never provided when constructing the prior.

B Additional Qualitative Comparisons

B.1 Same-image Prior Transfer

We first visualize the case where the inserted prior is extracted from the same source image. This setting isolates whether the prior can provide useful global structure to the frozen decoder without introducing cross-image mismatch. Compared with no prior and alternative global summaries, RegToken produces more coherent global layout and more stable target-category rendering across examples. Figure S1 shows representative cases.

B.2 Same-class / Cross-image Prior Transfer

We next test whether the proposed prior remains useful when transferred across images. We consider priors extracted from different source images within the same semantic class, as well as random cross-class source images, while keeping the target prompt fixed within each row. These examples help distinguish genuinely transferable global structure from trivial memorization of a specific source image. Figure S2 summarizes representative cases.

B.3 Comparison with Different Priors

We further compare RegToken against alternative global summaries and prior choices under the same-image prior setting. Specifically, we compare matched-norm random priors, patch-mean priors, `[CLS]` priors, test-time registers, and trained registers. These comparisons clarify that the gain does not come from merely injecting an additional token, but from the specific structure extracted by the proposed NFN + TokenRank pipeline. Figure S3 provides side-by-side examples.

B.4 Additional Decoding Examples Across Prompts/Classes

Finally, we provide a small gallery of additional decoding examples across diverse prompts and semantic classes beyond the representative cases shown in the main paper and preceding appendix figures. Since Figures S1 to S3 already compare

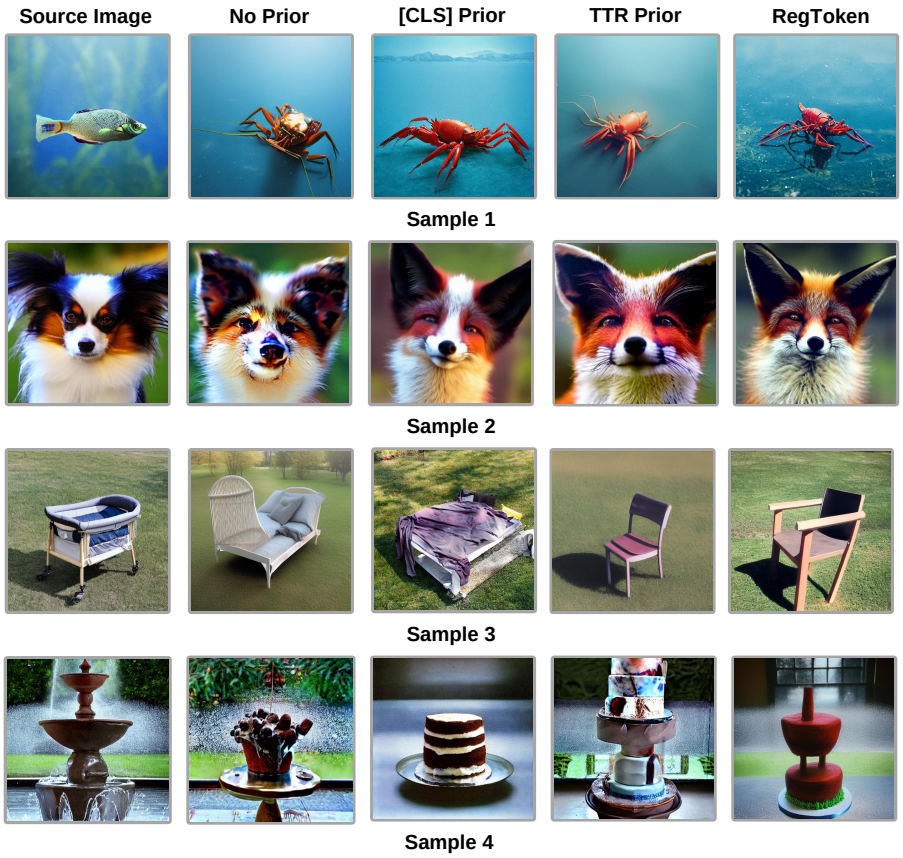

Fig. S1: Same-image Prior Transfer. Qualitative comparison when the inserted prior is extracted from the same source image. Each row shows one source image and the decoded result for the same target prompt under different prior choices. Compared with no prior and alternative global summaries, RegToken produces a more coherent global layout and a more stable target-category rendering, while still preserving useful image-specific global characteristics from the source.

prior-source relations and prior families in detail, this subsection focuses on qualitative breadth. Figure S4 shows additional source–target examples illustrating that RegToken generally maintains coherent global structure and target-category consistency under the same frozen-decoder setting.

C Causal Validation of the Proposed Prior

To complement the descriptive and correlational analyses in the main paper and Section D, we provide additional controlled comparisons to test whether the gain of RegToken comes from the specific structure of the extracted prior, rather than

Fig. S2: Same-class and Cross-image Prior Transfer. Each row compares different prior-source relations under the same target prompt. From left to right, we show the original prior source image, the no-prior result, a different source image from the same semantic class, the corresponding same-class transfer result, a random cross-class source image, and the corresponding random-source transfer result. Same-class transfer generally preserves more useful source-derived global structure than random-source transfer, while still producing the target category, indicating that the proposed prior captures genuinely transferable global information rather than merely memorizing a specific source image.

from simply injecting an extra token or adding arbitrary global bias. We focus on three aspects: frequency decomposition, shuffled/matched controls, and source sensitivity under prior transfer.

C.1 Low-frequency vs. High-frequency Components

To test whether the gain is associated with specific frequency components of the proposed prior, we decompose the inserted prior in the embedding domain using a 1D DCT over the feature dimension. The first k coefficients are retained as the low-frequency component, while the remaining coefficients form the high-frequency component. Each component is transformed back to the original embedding space and evaluated under the same decoding setup. This comparison isolates whether the observed improvement is mainly driven by the smoother global structure of the prior or by high-frequency residual information. We report both generation-quality and optimization-efficiency metrics under the same protocol described in Section A. Consistent trends are observed in both generation quality and optimization efficiency, indicating that the smoother component

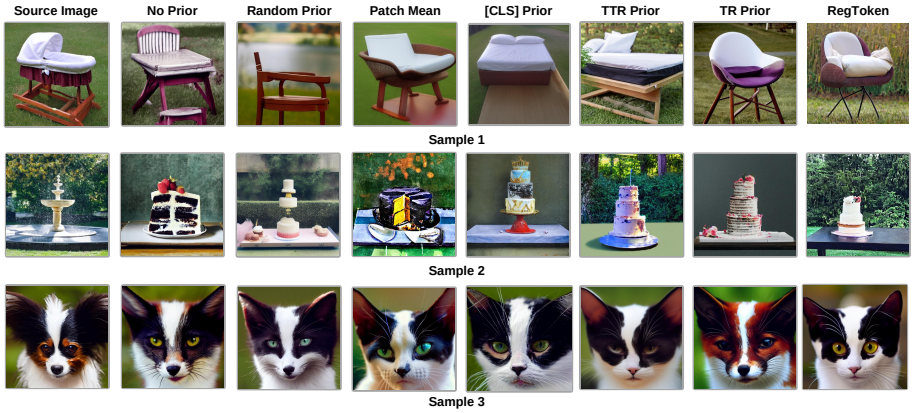

Fig. S3: Comparison with Different Priors. Qualitative comparison under the same-image prior setting using different global summaries and prior choices. From left to right, we show the source image, no-prior decoding, matched-norm random prior, patch-mean prior, `[CLS]` prior, test-time register prior, trained register prior, and RegToken. Compared with simpler global summaries and alternative prior constructions, RegToken produces more coherent target-category rendering while better preserving useful source-derived global structure, indicating that the gain comes from the specific prior extracted by the proposed NFN + TokenRank pipeline rather than from generic token injection.

Fig. S4: Additional Decoding Examples Across Prompts and Classes. A small gallery of additional examples beyond the representative cases shown in the main paper and earlier appendix figures. The top row shows source images, and the bottom row shows the corresponding RegToken outputs under different target prompts. These examples illustrate the breadth of the proposed prior across diverse semantic classes under the same frozen-decoder setting.

alone already explains a substantial portion of the improvement, although the

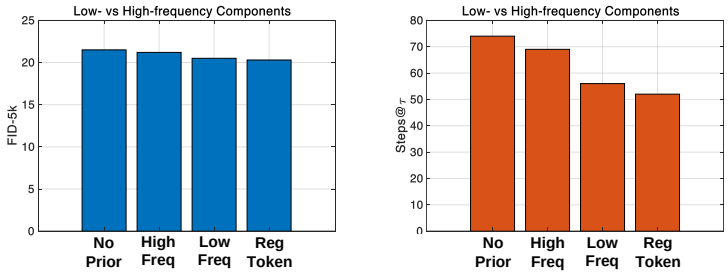

Fig. S5: Low- vs. High-Frequency Components of the Proposed Prior. We decompose the inserted prior in the embedding domain into low- and high-frequency components using a 1D DCT over the feature dimension, and evaluate each component under the same decoding protocol. Left: generation quality measured by FID-5k. Right: optimization efficiency measured by Steps@τ. The low-frequency component recovers most of the gain of the full RegToken prior, whereas the high-frequency component yields only limited improvement over the no-prior baseline. These results suggest that a substantial part of the benefit comes from the smoother global structure of the proposed prior, while the full prior remains the strongest overall. We set $k = 16$ in all experiments.

complete prior still performs best. The quantitative comparison is summarized in Figure S5.

C.2 Shuffled / Matched-norm / Random Controls

We next compare the proposed prior with several controlled alternatives to test whether the gain is merely due to injecting an additional token with similar magnitude. In particular, we consider a matched-norm random prior and shuffled versions of the extracted prior, and compare them against the full RegToken prior under the same decoding protocol. These controls help distinguish structured prior content from trivial token injection effects. Here, the shuffled prior is constructed by randomly permuting the feature dimensions of the extracted prior while preserving its ℓ_2 norm. We report both generation-quality and optimization-efficiency metrics under the same protocol described in Section A. The controlled comparison is summarized in Figure S6.

C.3 Source Sensitivity Analysis

We study how sensitive the proposed prior is to the choice of source image. Specifically, we compare same-image transfer, same-class cross-image transfer, and random-source transfer under the same decoding setup. These results help characterize how much of the benefit comes from transferable global structure versus source-specific alignment. The quantitative results are summarized in Figure S7 and correspond to the same source conditions reported in the main paper (Table 5).

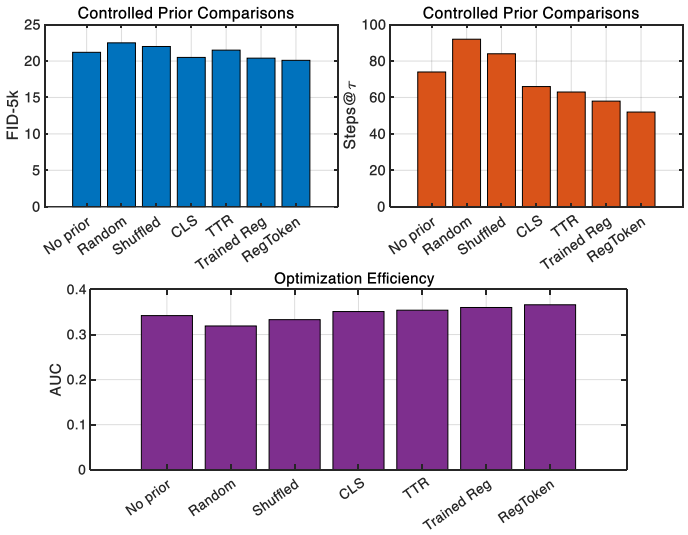

Fig. S6: Controlled Comparisons with Shuffled and Alternative Priors. We compare RegToken with several controlled alternatives under the same decoding protocol, including a matched-norm random prior, a shuffled prior obtained by permuting the feature dimensions of the extracted prior, a `[CLS]` prior, a test-time register prior, and a trained register prior. Left: generation quality measured by FID-5k. Right: optimization efficiency measured by Steps@τ. RegToken consistently outperforms the shuffled and random controls, indicating that the gain does not come from simply injecting an additional token of similar magnitude, but from the structured content of the proposed prior.

D Additional Analysis of Register Structure

D.1 NFN-based Layer/Module Localization

We compute NFN scores for each module and then aggregate them by $S_\ell = \max_m \mathrm{NFN}_{\ell,m}$ over a 1k-image random subset of ImageNet-`val`, following the procedure described in Section 4.1. The $\kappa = 5$ layers with the largest S_ℓ are selected as candidate insertion locations. Figure S8 visualizes the resulting layer×module heatmap for DINOv2 and the identified candidate layers.

D.2 NFN-based Layer Localization on OpenCLIP

OpenCLIP exhibits behavior similar to DINOv2. As shown in Figure S9, the median register value norm increases from 84.91 to 92.95 (+10.5%) on the same 1k-image ImageNet-`val` subset. In contrast, the median CLS→register attention and the corresponding statistics for image tokens remain largely unchanged.

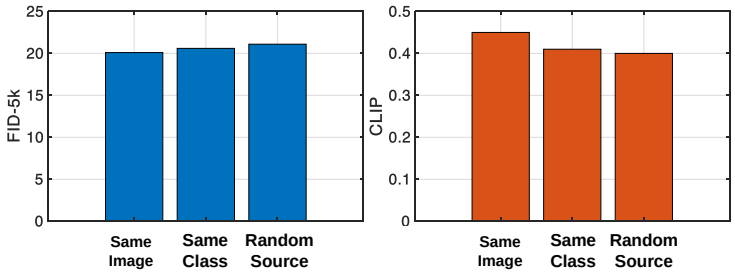

Fig. S7: Source Sensitivity of the Proposed Prior. We compare three source conditions under the same decoding protocol: a prior extracted from the same image, a prior transferred from a different image of the same class, and a prior transferred from a random image. Left: generation quality measured by FID-5k. Right: text-image alignment measured by CLIP. Same-image transfer performs best, while same-class cross-image transfer still retains a clear advantage over random-source transfer, suggesting that the proposed prior contains partially transferable global structure beyond source-specific alignment.

Table S3: Bridge to Pixel-space Frequency. Spearman correlation between the image-level low-frequency energy ratio ($\mathrm{LF}_{\mathrm{DCT}}$) computed from pixel-space DCT and token-spectrum metrics derived from register features, evaluated on 1k ImageNet images.

Backbone / Metric	Spearman ρ (↑)	p-value
DINOv2, $\rho_{1:15}$	0.22	$< 10^{-10}$
DINOv2, $\bar{d}B_{\mathrm{low}}$	0.19	$< 10^{-8}$
OpenCLIP, $\rho_{1:15}$	0.14	$< 10^{-5}$
OpenCLIP, $\bar{d}B_{\mathrm{low}}$	0.11	$< 10^{-3}$

D.3 Bridge: Token-spectrum vs. Pixel DCT Low-frequency

To relate the embedding-spectrum diagnostic to pixel-space statistics, we correlate it with an image-level low-frequency energy ratio (Table S3). Specifically, we compute $\mathrm{LF}_{\mathrm{DCT}}$ as the energy ratio of a fixed $k \times k$ low-frequency block ($k = 8$) in the 2D DCT of RGB pixels and report Spearman correlations with two-sided p-values over 1000 ImageNet images.

D.4 Optimization Dynamics Summary

We further interpret the trajectory-based metrics introduced in Section A. Under the *Prior + opt* protocol, only the injected token is optimized while $\mathbf{z}_{1:T}$ remains fixed. Compared with the alternative priors, RegToken reaches the target threshold in fewer steps, achieves the largest optimization AUC, and exhibits the smallest final variance across seeds. These trends indicate that the proposed prior improves both optimization efficiency and stability.

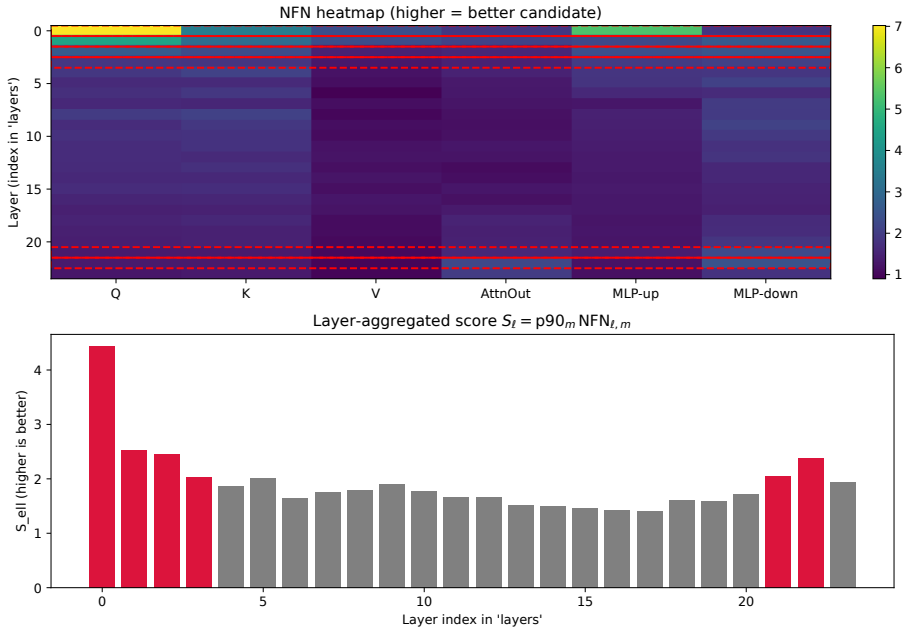

Fig. S8: NFN-based layer localization in DINOv2. The heatmap shows NFN scores across layers and modules, computed on a 1k-image subset of ImageNet-`val`. Layers with the largest scores are selected as candidate intervention points, which consistently align with the emergence of sink/register behavior.

E Additional Ablations

The full configuration (bottom row of Table S4) provides the best overall trade-off. Starting from the TTR baseline, adding NFN-based layer selection and TokenRank head gating progressively improves FID, IS, and SigLIP, while the final interpolation step yields the largest gain in FID-5k and text–image alignment. Performance remains stable across the hyperparameter settings in Tables S5 and S6. Varying the number of register neurons, selected channels, NFN-selected layers, outlier tokens, and scaling strengths changes FID only mildly, indicating that the method is relatively stable with respect to these choices. Removing NFN-based layer selection or TokenRank head gating leads to a small but consistent degradation, showing that both components contribute to the final performance.

E.1 Number of Selected Register Neurons

We first study the effect of the number of selected register neurons. As summarized in Table S5, varying the number of selected neurons from $k = 16$ to $k = 64$ leads to only minor changes in FID-5k, with the default setting $k = 32$ achieving

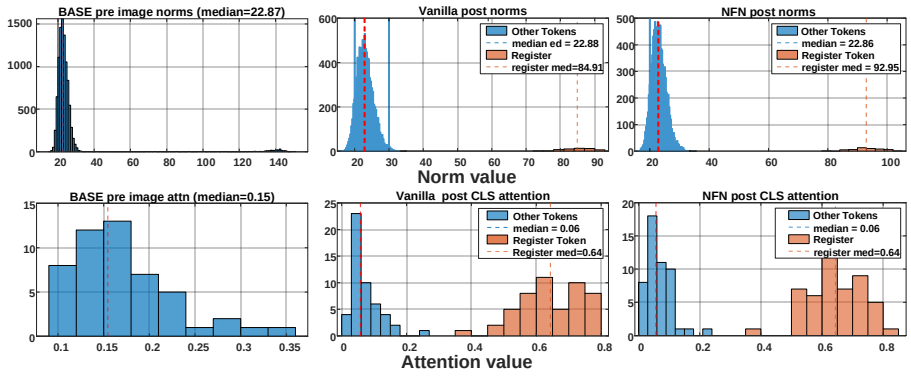

Fig. S9: Effect of register insertion on OpenCLIP features. We insert a single register token at NFN-selected layers using the top-50 register neurons. The register token accumulates larger feature norms and attention mass, while the distributions of image tokens remain largely unchanged.

Table S4: Component Ablation for HCT Decoding (DINOv2-L/14 + ImageNet-`val` 5k). Starting from the test-time register (TTR) baseline [14], we progressively add NFN-based layer selection, TokenRank head gating, and the token–neuron interpolation rule. All settings follow the same HCT decoding protocol as in the main paper.

Method	FID-5k (↓)	IS (↑)	CLIP (↑)	SigLIP (↑)
TTR baseline [14]	21.5	283	0.40	3.6
TTR + NFN layer selection	21.3	284	0.40	3.7
TTR + NFN + TokenRank head gating	20.8	286	0.41	3.8
TTR + NFN + TokenRank + interpolation (RegToken, HCT)	**20.3**	**287**	**0.41**	**3.9**

the best result. This suggests that the proposed prior is relatively stable with respect to the exact number of selected register dimensions within a moderate range.

E.2 Effect of NFN-based Layer Selection

We next examine the role of NFN-based layer selection. Removing NFN-based layer selection leads to a consistent degradation in generation quality, as shown in Table S5. A similar trend is observed in the progressive component ablation in Table S4, where adding NFN-based layer selection to the TTR baseline improves FID-5k from 21.5 to 21.3. These results indicate that selecting intervention layers based on NFN provides a small but consistent benefit.

E.3 Effect of TokenRank Head Gating

We further evaluate the effect of TokenRank-based head gating. As shown in Table S5, removing TokenRank gating degrades FID-5k from 20.3 to 20.9. The

Table S5: Ablation of Register Hyperparameters and Components (DINOv2-L/14 + HCT). FID-5k on ImageNet-`val` (5k images) for different numbers of register neurons k, register scales s/γ, and settings with or without NFN-based layer selection and TokenRank head gating.

Setting	k	s / γ	FID-5k (↓)
RegToken (default)	32	2.0	**20.3**
Fewer neurons	16	2.0	20.7
More neurons	64	2.0	20.7
Smaller scale	32	1.0	20.9
Larger scale	32	4.0	21.0
w/o NFN layer selection	32	2.0	21.1
w/o TokenRank gating	32	2.0	20.9

progressive ablation in Table S4 shows the same trend: adding TokenRank head gating on top of NFN-based layer selection improves FID-5k from 21.3 to 20.8. This confirms that TokenRank helps identify heads that more effectively carry the proposed prior.

E.4 Effect of Interpolation / Conservation Update

Finally, we assess the contribution of the interpolation / conservation update. In the progressive component ablation of Table S4, adding the interpolation rule on top of NFN-based layer selection and TokenRank head gating yields the best overall performance, improving FID-5k from 20.8 to 20.3 while also slightly improving IS and SigLIP. This suggests that preserving the structured prior during insertion is important for achieving the full gain of RegToken.

E.5 Details for the Value-to-TokenRank Bridge

This subsection describes the implementation details of the diagnostics used in Section 4.1. We couple a per-token write measure with a global centrality score to avoid interpreting raw attention in isolation. For a head with attention $A \in \mathbb{R}^{T\times T}$ (row-stochastic) and values $\{v_t\}$, we define the *write mass* of token t as:

$$\mathrm{WriteMass}(t) = \sum_{i=1}^{T} A_{i,t}\,\|v_t\|_2^2,$$

which better reflects how much content is written *into* t than attention alone. Treating A as a Markov chain, *TokenRank* is the stationary distribution $\pi^\top = \pi^\top A$, computed per head and then median-aggregated within a layer. We quantify head mixing by the second eigenvalue λ_2 of A; a smaller λ_2 corresponds to a larger spectral gap and faster mixing.

Table S6: Hyperparameter sensitivity. FID-5k under different RegToken hyperparameter settings. The method remains stable across moderate ranges of channel count, NFN-selected layers, outlier-token count, initialization strength, and soft-bias strength.

Hyperparameter	Values	FID-5k (↓)
# channels k	64 / 128 / 256 / 512	20.5 / 20.3 / **20.1** / 20.4
NFN top-κ layers	1 / 3 / 5	20.4 / **20.1** / 20.2
# outlier tokens q	1 / 2 / 4 / 8	20.3 / **20.1** / 20.4 / 20.6
Init strength γ	0.25 / 0.5 / 0.75 / 1.0	20.6 / 20.3 / **20.1** / 20.2
Soft-bias β	0.5 / 1.0 / 2.0	20.4 / **20.1** / 20.3

In practice, we use post-softmax attention matrices, power iteration for π, and report per-layer medians across heads for $\pi(\texttt{REG})$, WriteMass($\texttt{REG}$), and λ_2. To assay absorption, we scale the register coordinates on their subspace by $s \in \{0.5, 1, 2, 4, 8, 16\}$ and track TokenRank($\texttt{REG}$), WriteMass($\texttt{REG}$), and the median λ_2. Monotone increases in the first two together with a decrease in λ_2 indicate that the added token behaves as a controlled absorber rather than merely amplifying patch outliers.

We also use TokenRank to gate heads: at a candidate layer, we keep the top-h heads by $\pi(\texttt{REG})$ (typically $h = 1$–3) and apply the test-time register only on those heads. This strengthens absorption and reduces image-token outliers, yielding larger shifts in the λ_2–outlier joint plot than the vanilla baseline. In practice, we find that the resulting absorption curves are qualitatively similar across input images and classes, suggesting that the diagnostics capture a largely backbone-specific rather than image-specific effect.

E.6 Hyperparameter Sensitivity

We further evaluate the sensitivity of RegToken to key hyperparameters. As shown in Table S6, performance remains stable over moderate ranges of register-channel count, NFN-selected layers, outlier-token count, initialization strength, and soft-bias strength. The best setting is not isolated to a single brittle hyperparameter choice.

E.7 Fusion with `[CLS]`

We also evaluate whether RegToken and `[CLS]` provide complementary global signals. We form a fused prior $u = \mathrm{norm}(\alpha u_{\mathrm{reg}} + (1-\alpha) u_{\mathrm{cls}})$ with $\alpha \in \{0.25, 0.5, 0.75\}$. As shown in Table S7, RegToken-heavy fusion slightly improves over RegToken alone, while CLS-heavy fusion is less stable. This indicates that RegToken supplies the main global prior and `[CLS]` can add light semantic information.

Table S7: Fusion with `[CLS]`. Results are reported as DINOv2 / OpenCLIP. RegToken-heavy fusion performs best, while CLS-heavy fusion is less stable.

Prior	FID-5k (↓)	IS (↑)	CLIP (↑)	SigLIP (↑)
`[CLS]`	20.5 / 20.8	281 / 283	0.39 / 0.40	3.6 / 3.5
RegToken	20.1 / 20.3	289 / 287	0.45 / 0.42	3.9 / 3.8
RegToken + `[CLS]` ($\alpha = 0.25$)	20.4 / 20.7	282 / 283	0.41 / 0.40	3.5 / 3.6
RegToken + `[CLS]` ($\alpha = 0.50$)	20.2 / 20.5	286 / 285	0.43 / 0.41	3.7 / 3.7
RegToken + `[CLS]` ($\alpha = 0.75$)	**20.0** / **20.2**	**289** / **288**	**0.45** / **0.43**	**4.0** / **3.8**

F Component Summary and Practical Interpretation

To complement the ablation results in Section E, we summarize the practical role of each component in RegToken. Rather than introducing additional experiments, this section distills the ablation trends into an interpretable view of what each component contributes, which failure mode it addresses, and what minimal configuration already provides a strong improvement over the baseline.

F.1 What Each Component Contributes

A concise practical summary of these component-wise contributions is provided in Table S8. The ablations show that the gain of RegToken is not tied to a single component. Instead, each design choice contributes incrementally: NFN-based layer selection improves intervention placement, TokenRank head gating improves head selection, and the interpolation step yields the final performance gain by stabilizing the inserted prior representation.

F.2 Which Failure Mode Each Component Addresses

The corresponding mapping from each design choice to the failure mode it mitigates is summarized in Table S9. The ablation results suggest that RegToken works best when the prior is inserted at the right layer, routed through informative heads, and transferred in a way that preserves its structured global information.

F.3 Minimal Working Configuration

A practical takeaway from Tables S4 and S5 is that the full RegToken configuration provides the best overall performance, but a simpler configuration already yields a meaningful gain over the baseline. In particular, adding NFN-based layer selection and TokenRank head gating on top of the TTR baseline reduces FID-5k from 21.5 to 20.8, even before applying the final interpolation step. This suggests that a minimal working configuration can be formed by combining NFN-based layer selection with TokenRank head gating, while the interpolation / conservation update is the final ingredient needed to recover the full gain of RegToken.

Table S8: Practical Summary of the Main Components. A concise interpretation of the role of each component based on the ablations in Tables S4 and S5.

Component	Main Contribution	Evidence
NFN-based layer selection	Selects intervention layers where sink/register behavior is strongest, improving where the prior is inserted.	Adding NFN-based layer selection improves FID-5k from 21.5 to 21.3 in Table S4; removing it degrades FID-5k to 21.1 in Table S5.
TokenRank head gating	Identifies heads that more effectively carry the injected prior and improves prior routing.	Adding TokenRank head gating further improves FID-5k from 21.3 to 20.8 in Table S4; removing it degrades FID-5k to 20.9 in Table S5.
Interpolation / conservation update	Stabilizes the inserted prior representation and preserves useful global structure during transfer to the decoder token space.	Adding the interpolation step improves FID-5k from 20.8 to 20.3 and also slightly improves IS and SigLIP in Table S4.
Register neuron selection (k)	Provides a compact subspace for constructing the proposed prior while remaining stable over a moderate range of selected dimensions.	Varying k from 16 to 64 changes FID-5k only mildly (20.7 to 20.3) in Table S5.
Scale / strength (s or γ)	Controls the strength of prior insertion without requiring fine hyperparameter tuning.	Changing the scale from 1.0 to 4.0 only changes FID-5k from 20.9 to 21.0 in Table S5.

G Linear Probe on ImageNet

Consistent with prior work on trained and test-time registers, register-based features achieve competitive but slightly lower ImageNet linear-probe accuracy than the `[CLS]` token, while clearly outperforming simple patch averaging. This pattern suggests that registers encode compact scene-level statistics rather than serving as the strongest discriminative readout. Accordingly, the main paper focuses on repurposing these features as plug-and-play global priors for generation, where global context such as style, lighting, and layout is more directly useful. The quantitative comparison is summarized in Table S10. These results support the view that register-based summaries preserve meaningful global information even when they are not the best classification readout, which is consistent with their use as transferable priors for frozen tokenized generation.

H Failure Cases and Limitations

H.1 Failure Examples

We observe several failure cases in which the proposed prior does not fully resolve the limitations of the frozen decoding pipeline. Typical examples include

Table S9: Failure Modes Addressed by Each Component. Interpretation of the main failure mode mitigated by each design choice.

Component	Failure Mode Addressed	Practical Effect
NFN-based layer selection	Inserting the prior at suboptimal layers where sink/register behavior is weak.	Improves the consistency of intervention points across images.
TokenRank head gating	Routing the prior through heads that do not effectively absorb or propagate the inserted global token.	Improves the usefulness of the inserted prior and reduces ineffective head usage.
Interpolation / conservation update	Naively inserting the prior in a way that disrupts the decoder token space or fails to preserve structured information.	Makes prior transfer more stable and yields better final generation quality.
Register neuron selection	Using an over-complete or under-complete set of feature dimensions when forming the prior.	Maintains a compact but robust prior representation.
Scale / strength	Over-weak or over-strong prior injection that either has little effect or perturbs decoding too aggressively.	Keeps the method stable across a moderate range of insertion strengths.

strong source-induced semantic drift, target-category confusion, incomplete object structure, and failures under uncommon compositional or geometric demands. In such cases, the global prior may still shape coarse layout or appearance, but it does not reliably recover the exact target semantics or the full local part arrangement. Figure S10 shows representative examples.

H.2 Cases Where the Prior Helps Less

The proposed prior is most useful when global structure plays an important role in the decoded image, such as scene layout, illumination, or coarse semantic context. Its benefit is less pronounced when the target generation depends primarily on fine local texture, highly specific object instances, or precise geometric arrangement. This is consistent with our analysis that register-based priors mainly capture compact global statistics rather than serving as a complete replacement for token-level optimization or richer generative modeling.

H.3 Scope and Generalization Limits

This work focuses primarily on large vision transformers. Similar attention-sink phenomena have been reported in large language models, multimodal models, and retrieval architectures, but our method has not yet been evaluated in those settings. In addition, while the proposed approach improves token-based generative pipelines, more complex architectures with attention modules (*e.g.*, DiT)

Table S10: Linear-Probe Accuracy on ImageNet-1k (DINOv2-L/14). A linear classifier is trained on frozen features from different global summaries. Results for trained registers and test-time registers are consistent with prior reports [5,14]. The `CLS` token remains the strongest discriminative readout, while register-based features retain non-trivial semantic content.

Feature	Top-1 Accuracy (%) (↑)
[CLS] token	85.4
Patch mean	84.4
Trained register token [5]	83.1
Test-time register token [14]	84.5
RegToken (NFN + TokenRank register)	84.8

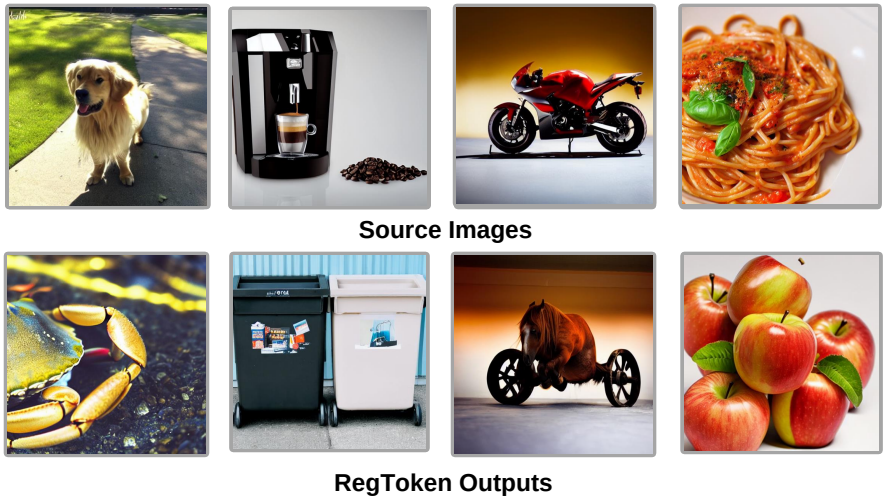

Fig. S10: Representative Failure Cases. Top row: source images. Bottom row: corresponding RegToken outputs. The examples illustrate several characteristic failure modes, including incomplete target-object structure, source-to-target semantic drift, and category confusion under challenging source-target transfers. These cases are consistent with our analysis that RegToken mainly provides compact global structure rather than a full replacement for fine token-level modeling and precise geometric composition.

may exhibit different sink dynamics and may require different intervention strategies. Understanding how register-like mechanisms interact with such models remains an important direction for future work. At the same time, we believe the current scope is still practically meaningful: RegToken operates in a fully frozen setting, requires no retraining, and under the Prior + opt protocol updates only a single injected global token. Moreover, its benefit appears not only in final generation metrics but also in optimization efficiency and stability, as reflected

by Steps@τ, AUC, and Std@S in Table S2. We therefore view the present setting as a useful testbed for lightweight plug-and-play priors in tokenized generation, while leaving broader validation on other generative architectures to future work.

I Algorithms and Additional Generality Checks

We summarize the three main procedures used in the paper:

- Algorithm 1: NFN-based layer and register-neuron localization with TokenRank diagnostics.
- Algorithm 2: Register-guided decoding for 1D tokenizers such as HCT.
- Algorithm 3: The Value-to-TokenRank bridge used to quantify register absorption and select heads for gating.

Algorithm 1 NFN-based Localization and TokenRank.

Require: ViT, dataset $\mathcal{D}$, image batch $\mathcal{B}$, $\kappa, n, k, w, \Delta, s$
Compute $\mathrm{NFN}_{\ell,m}$ over $\mathcal{D}$ for each layer ℓ and module m
Aggregate $S_\ell = \max_m \mathrm{NFN}_{\ell,m}$ and pick $\mathcal{L}_{\text{cand}} = \text{top-}\kappa$ layers with largest S_ℓ
for $\ell \in \mathcal{L}_{\text{cand}}$ **do** ▷ Register-neuron extraction
    Detect outlier tokens $\mathcal{P}$ via ℓ_2-norm maps across a window of w layers
    Rank channels by mean $|a_{p,d}|$ over $p \in \mathcal{P}$ and previous Δ layers
    Take $\mathcal{R}_\ell = \text{top-}k$ channels as register neurons
    Form t_{reg} and update values by the projection-and-conservation update in Section 4.3 with scale s
end for
for each head **do** ▷ Value×TokenRank diagnostics
    Compute WriteMass, TokenRank π, and λ_2 (second eigenvalue)
    Log curves of TokenRank(REG), WriteMass(REG), and λ_2 as s varies
end for
return Modified forward pass with test-time registers and per-head diagnostics

Algorithm 2 Register-guided Decoding for 1D tokenizers.

Require: Register tokens $\{r^{(\ell)}\}$, codebook $\mathcal{C}$, base tokens $\mathbf{z}$, weights $\{\alpha_\ell\}$, strength γ
for each $\ell \in \mathcal{L}_{\text{use}}$ **do**
    $j_\ell^\star \leftarrow \arg\max_j \sigma(r^{(\ell)}, c_j), \quad \tilde{r}^{(\ell)} \leftarrow c_{j_\ell^\star}$
end for
$\tilde{r} \leftarrow \sum_\ell \alpha_\ell \tilde{r}^{(\ell)}$ ▷ Fuse layer-wise registers
$z_0 \leftarrow (1-\gamma) z_0 + \gamma \tilde{r}$ ▷ Insert as global prior token
return decoded image $\mathcal{D}([z_0, \mathbf{z}_{1:T}])$

Algorithm 3 Value-to-TokenRank Bridge for Absorption.

Require: Attention matrices $\{A^{(\ell,h)}\}$, value tensors $\{V^{(\ell,h)}\}$, register index t_{reg}, scales $\mathcal{S} = \{s_1, \ldots, s_M\}$, number of gated heads h_{gate}

for each layer ℓ and head h **do**
  for each scale $s \in \mathcal{S}$ **do**
    Scale register coordinates in $V^{(\ell,h)}$ by s on the register subspace
    Recompute attention $A^{(\ell,h)}$ if needed and obtain updated values $V^{(\ell,h)}$
    Compute WriteMass(t_{reg}) via WriteMass(t) $= \sum_i A_{i,t}^{(\ell,h)} \, \|V_t^{(\ell,h)}\|_2^2$
    Treat $A^{(\ell,h)}$ as a Markov chain; compute TokenRank $\pi^{(\ell,h)}$ (stationary dist.)
    Estimate the second eigenvalue $\lambda_2^{(\ell,h)}$ of $A^{(\ell,h)}$
    Log TokenRank(t_{reg}), WriteMass(t_{reg}), and $\lambda_2^{(\ell,h)}$ for scale s
  end for
end for
For each layer ℓ, aggregate TokenRank(t_{reg}) over heads and select the top-h_{gate} heads as gated heads
return Absorption curves (TokenRank/WriteMass/λ_2 vs. s) and head-gating sets

Table S11: Generality across compact 1D token budgets. FID-5k results for different compact tokenizer variants. The VQ-LL/VQ-BB/VQ-BL names denote the tokenizer variants used in this diagnostic check, and the suffix indicates the token budget. Each cell reports DINOv2 / OpenCLIP. RegToken consistently improves over no-prior and `[CLS]`-prior baselines across token budgets.

Tokenizer variant	No prior	`[CLS]` prior	RegToken	ΔFID over no prior
VQ-LL-32	21.2 / 21.7	20.5 / 20.8	**20.1** / **20.3**	-1.1 / -1.4
VQ-BB-64	21.8 / 22.3	21.3 / 21.7	**20.9** / **21.1**	-0.9 / -1.2
VQ-BL-64	22.5 / 23.0	22.0 / 22.4	**21.7** / **21.9**	-0.8 / -1.1
VQ-BL-128	23.5 / 24.0	23.0 / 23.6	**22.6** / **23.3**	-0.9 / -0.7

Table S12: MaskGIT-VQGAN pilot. RegToken is used as a frozen codebook-logit initialization bias. It improves FID, CLIP, and SigLIP over MaskGIT, random-init, and `[CLS]`-init controls.

Prior on MaskGIT	FID-1k (↓)	CLIP (↑)	SigLIP (↑)	(Δ) over no prior
MaskGIT baseline	31.8	0.392	2.45	–
Random init bias	31.6	0.393	2.46	-0.2 / +0.001 / +0.01
`[CLS]` init bias	31.1	0.398	2.55	-0.7 / +0.006 / +0.10
RegToken init bias	**30.3**	**0.406**	**2.72**	-1.5 / +0.014 / +0.27

I.1 Generality Beyond the Main HCT-style Setting

To complement the main HCT-style experiments, we evaluate whether the proposed prior generalizes beyond the primary decoder and backbone setting. We consider three axes: compact 1D token budgets based on TiTok/HCT-style tokenizers [34,1], an alternative MaskGIT-VQGAN decoding interface [3], and DINOv2 backbone scales. The VQ-LL/VQ-BB/VQ-BL names denote the compact tokenizer variants used in this diagnostic check, and the suffix indicates the token budget.

Compact 1D token budgets. We first test multiple compact-token variants with different token budgets. Each cell reports results for DINOv2 / OpenCLIP priors under the same frozen decoding protocol. Across 32-, 64-, and 128-token variants, RegToken consistently improves over the corresponding no-prior and `[CLS]`-prior baselines.

MaskGIT-VQGAN pilot. We also test an alternative MaskGIT-VQGAN [3] decoding interface. Since this pipeline does not use the same inserted HCT-style global slot, we map the register-derived prior to a frozen codebook-logit initialization bias. As shown in Table S12, RegToken improves over the MaskGIT baseline as well as random and `[CLS]` initialization controls.

Backbone scales. Finally, we evaluate RegToken across DINOv2 backbone scales. RegToken remains stronger than the corresponding `[CLS]` prior across DINOv2-B/L/g, with smaller gains on ViT-B and comparable gains on ViT-g.

Table S13: Generality across DINOv2 backbone scales. FID-5k comparison between [CLS] and RegToken priors across DINOv2 backbone sizes.

Backbone on FID-5k (↓)	[CLS]	**RegToken**	**Gain over** [CLS]
DINOv2-L/14	20.5	**20.1**	+0.4
DINOv2-B/14	21.0	**20.9**	+0.1
DINOv2-g/14	20.3	**19.9**	+0.4

Together, these results suggest that the benefit of RegToken is not limited to the specific HCT-style decoder configuration used in the main experiments. Instead, the proposed register-derived prior provides a reusable global signal across token budgets, decoder interfaces, and backbone scales. Full diffusion or autoregressive integration may require different injection interfaces and is left for future work.