SPIN: SCALE-AWARE POSITION-UNBOUND IDENTITY INJECTION FOR CONSISTENT VISUAL AUTOREGRESSIVE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive (VAR) models generate images from coarse to fine over a schedule of scales, yet existing methods for subject-consistent generation improve identity consistency across prompts at the expense of pose and layout diversity. Methods that expose source activations to the target also expose their positional information. In VAR models, this locks the target's composition at the coarse scales where layout is determined, tying gains in identity consistency to losses in diversity. We present SPIN, a training-free method that relaxes this trade-off. Its premise is that the source–target *state gap* is a single object that a transformer block admits through exactly three interfaces, each capable of carrying identity while shedding positional information. State Read Transfer moves this gap into what a layer reads. RoPE-Rewound Subject Attention allows the target to attend to source keys whose rotary phase is rewound on the band that carries position. Pre-Render Write Transfer transfers the per-layer increment, in which everything accumulated before the layer cancels out. Each transfer is placed according to scale- and layer-dependent positional effects. On a 191-group benchmark, SPIN reduces DreamSim on Infinity-2B from 0.235 to 0.182 while retaining 89% of the backbone's pose diversity. It lies above the identity–diversity trend of twelve prior methods, and the same placement recipe improves identity across five tested next-scale backbones.

1 INTRODUCTION

Text-to-image (T2I) models generate diverse images from textual prompts. Applications such as storyboarding, visual narratives, and character-driven content additionally require a subject's identity—a character's face, an animal's markings, or an object's shape and texture—to remain consistent across scenes and prompts (Gal et al., 2023; Ruiz et al., 2023; Ye et al., 2023; Li et al., 2024; Tewel et al., 2024; Zhou et al., 2024; Liu et al., 2025; Li et al., 2025; Wang et al., 2025). Identity drifts across generations, yet constraints strong enough to prevent this drift also reproduce the source's pose and layout.

Existing diffusion-based methods address this trade-off through hand-designed schedules, attention operations, or correspondence cues (Tewel et al., 2024; Zhou et al., 2024; Liu et al., 2025; Li et al., 2025; Wang et al., 2025; Gao et al., 2026), choosing *when* and *where* to transfer identity heuristically. Many of these methods expose source activations to the target, carrying positional information along with identity and thus sacrificing layout diversity. Across prior methods, we observe a pronounced tension between identity consistency and diversity (Fig. 3).

We study this problem in scale-wise visual autoregressive (VAR) models (Tian et al., 2024; Han et al., 2025), a fast alternative to iterative denoising (Nichol et al., 2022; Rombach et al., 2022; Podell et al., 2024). VAR makes two useful structures explicit: spatial token grids provide direct token-to-region correspondence, and discrete coarse-to-fine scales place layout commitment at an architectural boundary rather than a tuned timestep. These properties provide the structure needed to decide when and where to transfer identity while limiting positional information, yet their use in subject-consistent generation remains underexplored. A prior VAR-native approach (Park et al., 2026) targets identity and style without addressing pose diversity.

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

We present **SPIN**, a training-free framework for subject-consistent VAR generation, built on the observation that the source–target *state gap* $\mathbf{\Delta} = \mathbf{h}^{\mathrm{src}} - \mathbf{h}^{\mathrm{tgt}}$ can be transferred as a single object through three interfaces of a transformer block, each capable of carrying identity while limiting positional carryover. Our contributions are:

- **Three transfers of one gap.** State Read Transfer (SRT) expresses the source–target gap at the attention input; RoPE-Rewound Subject Attention (R^2SA) exposes source keys and values with the positional rotary band rewound in the keys; and Pre-Render Write Transfer (PWT) transfers the source–target difference of per-layer increments $\mathrm{d}\mathbf{z}^l = \mathbf{h}^l - \mathbf{h}^{l-1}$ rather than the accumulated state, so that little of the state accumulated before the layer is carried over to the target.
- **Scale- and layer-aware placement.** Two measurements guide transfer placement: the source bank's attention share across scales and the positional content of layer writes across depth. R^2SA is withheld from the scales where layout is committed, PWT covers them instead, and no module acts at 1×1.
- **Analysis of the transfer mechanisms.** We provide evidence that the three transfers are complementary, that a layer increment $\mathrm{d}\mathbf{z}$ is less position-bound than the state $\mathbf{h}$ it comes from, and that positional effects concentrate in a rotary band $\mathcal{P}$ while identity does not.

SPIN reduces DreamSim on Infinity-2B from 0.235 to 0.182 while retaining 89% of the backbone's pose diversity, and the same recipe improves identity across all five tested next-scale backbones.

2 Related Work

Text-to-image generation with visual autoregressive models. Diffusion models (Nichol et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Podell et al., 2024; Ramesh et al., 2022) have driven recent progress in T2I generation through iterative denoising, while flow-matching models (Black Forest Labs, 2024) provide an alternative continuous generative formulation. Discrete autoregressive models instead predict visual tokens (Ramesh et al., 2021; Ding et al., 2021; Yu et al., 2022; Sun et al., 2024); VAR (Tian et al., 2024) reformulates this as coarse-to-fine *next-scale* prediction over multi-scale token maps, extended to T2I by STAR-T2I (Ma et al., 2024), Switti (Voronov et al., 2025), HART (Tang et al., 2025), and Infinity (Han et al., 2025), the last with bitwise token prediction and an infinite-vocabulary classifier. Infinite Story (Park et al., 2026) adapts scale-wise models to training-free consistent generation but primarily targets identity and style consistency; the scale- and layer-wise roles of identity and pose in these models remain underexplored.

Consistent image generation. Subject-specific optimisation—Textual Inversion (Gal et al., 2023), DreamBooth (Ruiz et al., 2023), LoRA-based adaptation (Hu et al., 2022; Gu et al., 2023)—learns an embedding or adapts parameters per subject. Learned reference conditioning—IP-Adapter (Ye et al., 2023), PhotoMaker (Li et al., 2024), OminiControl (Tan et al., 2025), FLUX.1 Kontext (Black Forest Labs et al., 2025)—extracts the subject from reference images with pretrained conditioning modules. Training-free inference-time methods modify a pretrained model's generation: ConsiStory (Tewel et al., 2024) and StoryDiffusion (Zhou et al., 2024) share subject features or attention across images, 1Prompt1Story (Liu et al., 2025) and Zigzag Sampling (Li et al., 2025) control how subject information propagates, and Diptych (Shin et al., 2025) casts subject-driven generation as grid inpainting. Recent methods have explicitly addressed pose diversity: CharaConsist (Wang et al., 2025) preserves fine-grained identity under large motion with point-tracking attention, while CoDi (Gao et al., 2026) applies pose-aware identity transport early in denoising and identity refinement later. These pose-aware approaches remain diffusion-based; we study pose-diverse consistency in VAR generation, where scale and layer provide explicit structural axes for controlling when and where consistency is injected.

3 Method

3.1 Problem setting

A *prompt group* $\mathcal{G} = (\pi^0, \ldots, \pi^T)$ contains one source prompt and T target prompts describing the same subject with different actions and backgrounds, from which we generate I^0 and $I^1, \ldots, I^T$.

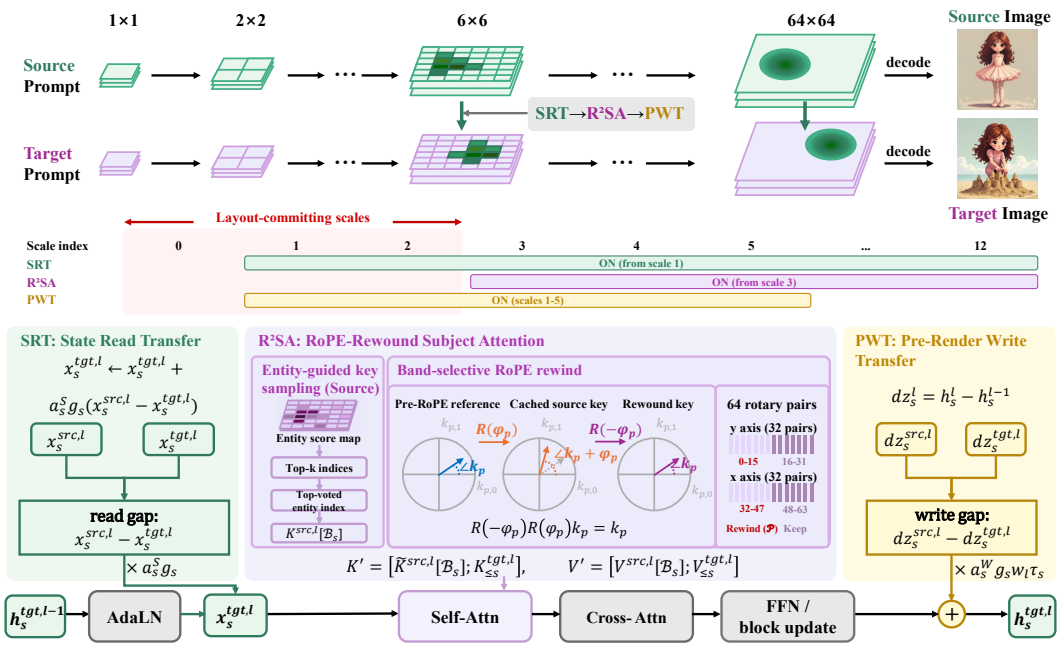

Figure 1: **Overview.** *Top:* the source and target prompts are generated by the same next-scale backbone; the source pass is unmodified and only records activations, which the target pass reads through SRT, R^{2}SA and PWT. *Middle:* the scales at which each module is active; R^{2}SA is withheld from the layout-committing scales, which PWT covers instead (Sec. 3.7). *Bottom:* the three modules attached to the three interfaces of one transformer block — the attention input (SRT, Eq. 5), the self-attention keys and values (R^{2}SA, Eq. 6), and the layer increment (PWT, Eq. 9). Site restriction to $\mathcal{I}_{s,l}$ is omitted from the panels for brevity.

Each target image I^t should preserve the subject identity in I^0 while following its corresponding prompt π^t. Across $\{I^0, \ldots, I^T\}$, we measure subject **identity** by mean pairwise subject similarity and **diversity** by mean pairwise distances in pose and centroid location, alongside the **prompt fidelity** of I^t to π^t. The unmodified backbone provides a reference for identity and diversity. Copying the source also copies its layout, creating a trade-off observed across prior methods (Fig. 3). We aim to improve this trade-off **without training**: no model parameters are added or updated, no gradients are computed, and all modules modify activations only at inference.

Notation. Scalars are italic (a, g, s, l), vectors and matrices bold ($\mathbf{h}, \mathbf{x}, \mathbf{K}$), and index sets calligraphic ($\mathcal{I}, \mathcal{B}, \mathcal{L}, \mathcal{A}$). Superscripts src/tgt denote the two passes; l indexes the L layers, s the $S+1$ scales, and d is the hidden width. $\mathbf{x}^l$ is indexed by the layer that *consumes* it and $\mathbf{h}^l$ by the layer that *produces* it, so $\mathbf{x}^l$ is computed from $\mathbf{h}^{l-1}$.

3.2 Backbone and Transfer Interfaces

The backbone is a next-scale VAR transformer over a fixed grid schedule, with n_s tokens at scale s and $N_s = \sum_{j \le s} n_j$ cumulative tokens. At each scale, it processes the upsampled accumulated code map through L blocks, attends to scales $0, \ldots, s$ through cached keys and values, and adds a sampled residual code block to the map. The first few scales, termed *layout-committing scales*, largely establish placement, size, and coarse pose from a few tokens. Position is encoded by 2D axial RoPE, with $\theta_j = 10{,}000^{-j/32}$ for the j-th rotary pair on each axis. These rotations make attention logits depend on query–key grid offsets, exposing positional components that can be selectively rewound.

Let $\mathbf{h}_s^l \in \mathbb{R}^{n_s \times d}$ denote the residual stream after block l. Given AdaLN modulation (γ, σ, β) from the conditioning embedding, each block computes

$$\mathbf{x}_s^l = \mathrm{LN}(\mathbf{h}_s^{l-1}) \odot (1+\sigma_1) + \beta_1, \qquad \mathbf{h}_s^l = \mathbf{h}_s^{l-1} + \gamma_1 \odot \mathrm{SA}(\mathbf{x}_s^l) + \gamma_{\mathrm{ca}} \odot \mathrm{CA}(\cdot) + \gamma_2 \odot \mathrm{FFN}(\cdot), \quad (1)$$

with LN layer normalisation, SA self-attention, CA text cross-attention and FFN the feed-forward network (the latter two on the normalised running state). Each block exposes three transfer interfaces, used by SRT, R^{2}SA, and PWT, respectively (Fig. 1, bottom):

$$\underbrace{\mathbf{x}_s^l}_{\text{what the layer } \textbf{reads}} \longrightarrow \underbrace{\{\mathbf{K}, \mathbf{V}\} \text{ of SA}}_{\text{what it may read } \textbf{from}} \longrightarrow \underbrace{\mathbf{h}_s^l - \mathbf{h}_s^{l-1}}_{\text{what it } \textbf{writes}} .$$

3.3 Core idea: one gap, three transfer interfaces

We define the *state gap* between the two passes at matching coordinates as

$$\boldsymbol{\Delta}_s^l \;=\; \mathbf{h}_s^{\mathrm{src},l} - \mathbf{h}_s^{\mathrm{tgt},l}. \tag{2}$$

$\boldsymbol{\Delta}$ accumulates information across layers up to l, including subject position. Direct transfer can therefore carry source pose and layout along with identity. We transfer this gap through the three interfaces of Sec. 3.2, using complementary mechanisms to reduce positional carryover. **SRT** (*State Read Transfer*) expresses the source–target gap at the *attention input* rather than transferring it directly to the stream: it changes what the layer reads and lets the layer decide what is written. **R^{2}SA** (*RoPE-Rewound Subject Attention*) makes no explicit correction to the residual stream. It adds source keys and values to the target's softmax and rewinds the positional rotary band only in the source keys to reduce source-location bias. **PWT** (*Pre-Render Write Transfer*) mixes source and target layer increments instead of full states:

$$\mathbf{h}_s^{\mathrm{tgt},l} \;\leftarrow\; \mathbf{h}_s^{\mathrm{tgt},l-1} + (1-a)\,\mathrm{d}\mathbf{z}_s^{\mathrm{tgt},l} + a\,\mathrm{d}\mathbf{z}_s^{\mathrm{src},l}, \qquad \mathrm{d}\mathbf{z}_s^l = \mathbf{h}_s^l - \mathbf{h}_s^{l-1}, \tag{3}$$

which is equivalent to adding $a\,(\boldsymbol{\Delta}_s^l - \boldsymbol{\Delta}_s^{l-1})$: the accumulated state before layer l is not transferred directly; only the current-layer increment difference is added. Across contiguous layers, the gap differences telescope. Our probes show that increments are less tied to position than full states, supporting their use for identity transfer (Sec. 4).

Shared subject sites. The first image in each group comes from an unmodified pass that only *records* activations (Fig. 1, top). At each (s, l), entity saliency $\omega_{s,l}$ is computed per token by averaging the block's text cross-attention over heads and the subject noun phrase. We retain the *site set*

$$\mathcal{I}_{s,l} \;=\; \mathrm{top}_k\left(\omega_{s,l}\right), \qquad k = \min(k_{\max},\, n_s). \tag{4}$$

At these sites, we store the attention input for SRT and the layer write for PWT. The R^{2}SA bank uses the same sites (Sec. 3.5), giving all three modules shared subject coordinates. A scalar gate $g_s \in [0, 1]$, a thresholded normalised cross-correlation between the target's accumulated code map at scale s and the source's final one (App. A), scales SRT and PWT: it stays closed while the target's composition has not yet settled and opens once the two maps agree, so injection never precedes the layout. R^{2}SA is ungated.

3.4 SRT: State Read Transfer

For $l \in \mathcal{L}_{\mathrm{SRT}}$, SRT moves the post-AdaLN attention input of Eq. (1) toward its source counterpart at the selected sites:

$$\mathbf{x}_s^{\mathrm{tgt},l}[\mathcal{I}_{s,l}] \;\leftarrow\; \mathbf{x}_s^{\mathrm{tgt},l}[\mathcal{I}_{s,l}] \;+\; a_s^{\mathrm{S}}\, g_s \underbrace{\left(\mathbf{x}_s^{\mathrm{src},l}[\mathcal{I}_{s,l}] - \mathbf{x}_s^{\mathrm{tgt},l}[\mathcal{I}_{s,l}]\right)}_{\textit{read gap}}, \tag{5}$$

where a_s^{S} decays linearly across the scale schedule. Since the skip connection in Eq. (1) still adds $\gamma_1 \odot \mathrm{SA}(\mathbf{x}_s^l)$ to the unmodified $\mathbf{h}_s^{l-1}$, the read gap reaches the residual stream through the layer's projections, softmax, and attention output. This keeps SRT active at layout-committing scales without directly modifying the residual state; the modified read also enters the value projection and the layer's cache. The small layer set $\mathcal{L}_{\mathrm{SRT}}$ is split evenly between the *pre-render layer band*, at layers that also receive PWT, and the subsequent *render stage*. Earlier reads can affect layout, while later reads have less influence on identity. Raising a_s^{S} with this split improves identity at a lower diversity cost than using either stage alone (Sec. 4).

3.5 R^{2}SA: RoPE-Rewound Subject Attention

At every layer of an active scale $s \in \mathcal{A}$, we prefix a small bank of source keys and values to the target's. A *single* softmax processes the concatenation without additional bank-specific weights, temperatures, or biases:

$$\mathbf{K}' = \left[\tilde{\mathbf{K}}^{\mathrm{src},l}[\mathcal{B}_s];\ \mathbf{K}^{\mathrm{tgt},l}_{\le s}\right], \quad \mathbf{V}' = \left[\mathbf{V}^{\mathrm{src},l}[\mathcal{B}_s];\ \mathbf{V}^{\mathrm{tgt},l}_{\le s}\right], \quad \mathrm{SA} = \mathrm{softmax}\Big(\mathbf{q}\,\mathbf{K}'^{\top}/\sqrt{d_h}\Big)\,\mathbf{V}', \tag{6}$$

for each query row $\mathbf{q}$ of a head with head width d_h, all keys and queries carrying their RoPE phase; $\mathbf{K}^{\mathrm{tgt},l}_{\le s}$ is the target cache over scales $0, \ldots, s$ and the tilde denotes the rewound keys of Eq. (7). The *bank* $\mathcal{B}_s$, $|\mathcal{B}_s| = B$, summarises source scales $0, \ldots, s$ using positions sampled at uniform stride on each grid. A fraction ρ of each scale j's slice is replaced by *entity-voted* positions: those appearing most often across $\{\mathcal{I}_{j,l}\}_l$ of Eq. (4). All layers share the bank index set but use keys and values from their own caches (ρ in Sec. 4).

Rewinding the source-key phase. Bank keys are retrieved *after* RoPE and retain the rotary phase $\varphi_p(\mathrm{pos}_k)$ of the source position in pair p. The logit term $\cos\big(\angle q_p - \angle k_p + \varphi_p(\mathrm{pos}_q) - \varphi_p(\mathrm{pos}_k)\big)$ depends on the query–key grid *offset* and can favour source tokens at matching grid locations, encouraging layout copying. We remove the source-position phase from bank keys within a band $\mathcal{P}$ by applying the inverse rotation:

$$\left[\tilde{k}_{p,0},\ \tilde{k}_{p,1}\right] = \left[\cos\varphi_p\, k_{p,0} + \sin\varphi_p\, k_{p,1},\ \ -\sin\varphi_p\, k_{p,0} + \cos\varphi_p\, k_{p,1}\right], \qquad p \in \mathcal{P}, \tag{7}$$

leaving all other pairs and the values unchanged. The inverse uses the model's rotary cache and preserves key norms. $\mathcal{P}$ is the fast-rotating half of the pairs on both axes, as ablations (Tab. 2, App. B) suggest that positional effects concentrate in this band while useful identity information is not confined to this band, motivating selective rewinding rather than filtering identity by band. *Limitation.* The query-side phase remains intact. Although shared by all bank keys for a given query, it allows the bank's overall affinity to vary with query position.

Scale-dependent bank influence. The bank contains at most B source positions from scales $0, \ldots, s$, while the target's key count N_s grows across scales. Assuming comparable attention scores for source and target keys, the bank's attention share can be approximated from their counts:

$$\mathrm{share}(s) \ \approx\ \frac{\min(B, N_s)}{\min(B, N_s) + N_s}. \tag{8}$$

At coarse scales, where $N_s \le B$, source and target contribute equal numbers of keys, giving an estimated bank share of one half. Once $N_s > B$, this estimate decreases as $B/(B + N_s)$, from 36% at scale 3 to 0.3% at the final scale for $B = 32$, so R^{2}SA uses no explicit gate (Sec. 3.7).

3.6 PWT: Pre-Render Write Transfer

PWT modifies the block output. Let $\mathrm{d}\mathbf{z}^l_s = \mathbf{h}^l_s - \mathbf{h}^{l-1}_s$ be the layer increment, and $\mathcal{L}_{\mathrm{PWT}}$ and $\mathcal{W}$ the layers and scales that receive the write. For $l \in \mathcal{L}_{\mathrm{PWT}}$, $s \in \mathcal{W}$, and $i \in \mathcal{I}_{s,l}$,

$$\mathbf{h}^{\mathrm{tgt},l}_s[i] \ \leftarrow\ \mathbf{h}^{\mathrm{tgt},l}_s[i] \ +\ a^{\mathrm{W}}_s\, g_s\, w_l\, \tau_s(i)\ \underbrace{\big(\mathrm{d}\mathbf{z}^{\mathrm{src},l}_s[i] - \mathrm{d}\mathbf{z}^{\mathrm{tgt},l}_s[i]\big)}_{\textit{write gap}}, \tag{9}$$

where a^{W}_s decays like a^{S}_s, and $w_l = w_{\mathrm{tot}}/|\mathcal{L}_{\mathrm{PWT}}|$ assigns uniform layer weights under a fixed budget w_{tot}, separating layer selection from transfer strength. The spatial gate $\tau_s(i) \in [0, 1]$ comes from the target's own cross-attention to its subject tokens. In addition to the cancellation in Eq. (3), $\mathrm{d}\mathbf{z}^{\mathrm{tgt},l}_s$ is recomputed from the current block input and output, making the correction relative to the target's current behaviour. At every written scale except the last, $n_s \le k_{\max}$, so $\mathcal{I}_{s,l}$ contains all tokens and τ_s determines localisation. We place writes at the front of the empirically identified pre-render layer band. When decoded and compared with the final image codes using *render NCC*, the normalised cross-correlation of Sec. 4, layer writes are anti-correlated early, weakly positive in a middle band, and strongly positive late. The pre-render band has the lowest measured positional load while its contributions have not yet been rendered. Within this band, we use the front half, which has the largest measured downstream effect per unit of injected norm (Sec. 4).



Figure 2: **Qualitative comparison of consistent image generation.** SPIN preserves distinctive subject identity across diverse scene prompts while retaining prompt-driven variations in pose, composition, and background. Compared with prior methods, it reduces identity drift without collapsing the generated images into repeated configurations.

3.7 Where each module acts

At layout-committing scales, few tokens shape identity and composition (the *diversity-sensitive scales* in Fig. 1, middle). We test four placement choices based on positional effects (Tabs. 2 and 4). **R^{2}SA skips these scales**: early attention improves identity but sharply reduces layout diversity (Tab. 2b). Eq. (8) suggests comparable source and target influence, so $\mathcal{A}$ starts at the first scale with $n_s > B$ (hence $N_s > B$), giving the bank a minority share that decreases thereafter. **PWT covers the excluded scales**: layer increments under the target's spatial gate correct identity with less direct transfer of accumulated source geometry. **No module acts at** 1×1, where any transfer is global. Both gated modules are inactive because $g_0 = 0$ and τ_s is degenerate; forcing either on yields no identity gain and reduces layout diversity in ablations. **Rewind acts only at the first attend scale**, where the estimated bank share remains large before declining at later scales.

4 Experiments

Setup. We evaluate 191 prompt groups (1,100 images) from ConsiStory+ (Liu et al., 2025) with a fixed seed. We measure segmented-subject identity (CLIP-I (Radford et al., 2021) ↑, DreamSim (Fu et al., 2023)↓), full-image prompt fidelity (CLIP-T (Radford et al., 2021)↑, VQAScore (Lin et al., 2024)↑), and within-group pose diversity following CoDi (Gao et al., 2026) and subject centroid spread (both ↑), referenced to the unmodified backbone (App. C). On Infinity-2B, PWT uses layers 4–9 and scales 1–5; SRT uses layers 4, 6, 9 (pre-render) and 16, 24, 31 (render), with a_s^{S} decaying

Method	TF	Reference	Backbone	CLIP-I↑	DS↓	CLIP-T↑	VQA↑	Pose↑	Centroid↑
SDXL backbone (no module)	✓	–	SDXL	0.8668	0.3120	0.3232	0.8381	0.0967	0.1574
1Prompt1Story	✓	–	SDXL	0.8980	0.2309	0.3041	0.7687	0.0662	0.1118
ConsiStory	✓	–	SDXL	0.8943	0.2272	0.3186	0.8396	0.0785	0.1403
StoryDiffusion	✓	–	SDXL	0.8904	0.2447	0.3141	0.8157	0.0750	0.1312
CoDi	✓	–	SDXL	0.8907	0.2364	0.3151	0.8192	0.0947	0.1396
PhotoMaker	×	Img	SDXL	0.8487	0.3495	0.3138	0.8116	**0.1194**	**0.1733**
IP-Adapter	×	Img	SDXL	0.9218	0.1720	0.3058	0.7750	0.0680	0.1006
FLUX backbone (no module)	✓	–	FLUX.1-dev	0.8757	0.2760	0.3045	0.8077	0.1007	0.1235
Zigzag	✓	–	FLUX.1-dev	0.9180	0.1728	0.3086	0.8404	0.0684	0.0768
CharaConsist	✓	–	FLUX.1-dev	0.8872	0.2510	0.3013	0.8103	0.0892	0.1253
Diptych	✓	Img (Grid)	FLUX.1-dev	0.8787	0.2738	0.2916	0.7405	0.0853	0.0960
OminiControl	×	Img	FLUX.1-schnell	0.9132	0.1845	0.2976	0.7981	0.0862	0.1068
FLUX.1 Kontext	×	Img	FLUX.1 Kontext	**0.9509**	**0.0947**	0.2892	0.7628	0.0446	0.0585
Infinity-2B backbone (no module)	✓	–	Infinity-2B	0.8945	0.2348	0.3037	0.8311	0.0842	0.1052
Infinite Story	✓	–	Infinity-2B	0.9213	0.1592	0.2983	0.8052	0.0614	0.0728
SPIN-2B (Ours)	✓	–	Infinity-2B	0.9142	0.1821	0.3042	0.8254	0.0750	0.0945
Infinity-8B backbone (no module)	✓	–	Infinity-8B	0.8958	0.2299	0.3209	0.8575	0.1008	0.1097
SPIN-8B (Ours)	✓	–	Infinity-8B	0.9142	0.1886	**0.3219**	**0.8607**	0.0913	0.0954

Table 1: **Comparison with consistent image generation methods.** TF: training-free. Reference: an input image (Img) is required at inference. Vanilla backbone rows are shown for reference and excluded from best/second-best ranking. **Bold** / underline: best / second best among consistent generation methods.

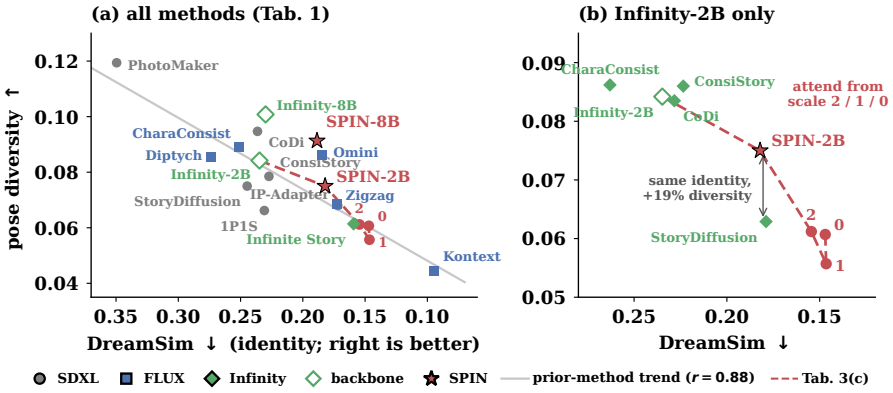

Figure 3: **The identity–diversity trade-off.** Identity (DreamSim, right is better) against pose diversity. *(a)* Consistent-generation methods in Tab. 1, excluding the vanilla backbone references, lie along one descending trend ($r = 0.88$), gaining identity in proportion to the diversity they give up; SPIN sits above it. *(b)* Infinity-2B only: the re-implemented mechanisms of Tab. 3 (App. A) and, dashed, our own method move along the trend when R^{2}SA is activated earlier (Tab. 2b; numbers: first attend scale). At the identity of StoryDiffusion, SPIN keeps 19% more diversity.

from 0.36 to 0.12 and a_s^{W} from 0.30 to 0.10; R^{2}SA starts at scale 3 and rewinds only scale 3, with $B = 32$, $k_{\max} = 64$, and $\rho = 0.375$. See App. A for 8B, other backbones, and the gate.

Comparison with prior methods. The twelve prior methods in Tab. 1 show an identity–diversity trade-off ($r = 0.88$ between DreamSim and pose diversity; Fig. 3a). FLUX.1 Kontext has the highest identity and lowest pose and centroid diversity; IP-Adapter, Infinite Story, and Zigzag follow in identity, while PhotoMaker reverses this pattern. SPIN lies above this empirical trend. SPIN-8B approaches IP-Adapter and Zigzag in identity (0.914 vs. 0.918–0.922 CLIP-I), with 34% more pose diversity and the highest CLIP-T and VQA. On Infinity-2B, Infinite Story gains 0.007 CLIP-I over SPIN-2B but has 18% lower pose and 23% lower centroid diversity, with lower prompt fidelity on both metrics. SPIN-2B gains 0.020 CLIP-I over the backbone while retaining 89% of its pose diversity. Qualitative comparisons appear in Fig. 2 and App. L.

Configuration	Subject identity		Prompt fidelity		Diversity	
	CLIP-I↑	DreamSim↓	CLIP-T↑	VQA↑	Pose↑	Centroid↑
Infinity-2B backbone (no module)	0.8945	0.2348	0.3037	0.8311	0.0842	0.1052
Ours: SRT (3 pre-render + 3 render) + R^{2}SA (from scale 3) + PWT (scales 1–5)	0.9142	0.1821	0.3042	0.8254	0.0750	0.0945
(a) Contribution of each module						
PWT only	0.9023	0.2113	0.3044	0.8239	0.0824	**0.0944**
PWT + R^{2}SA	0.9070	0.1979	0.3039	0.8295	0.0791	0.0942
(b) R^{2}SA: first attend scale $\min \mathcal{A}$ (ours: 3, the first scale with $n_s > B$)						
from scale 0	**0.9242**	0.1470	0.2999	0.8044	0.0607	0.0644
from scale 1	**0.9252**	0.1465	0.3000	0.8090	0.0557	0.0642
from scale 2	**0.9240**	0.1545	0.3001	0.8104	0.0612	0.0698

Table 2: **Ablation on Infinity-2B (main blocks).** Each row changes one switch relative to Ours; (a) adds the modules one at a time and (b) varies the first attend scale of Sec. 3.7. The remaining switches (PWT scales, rewind band and scale, SRT allocation) are in Tab. 4 (App. B). **Bold** / underline: significantly higher / lower than Ours under a paired Wilcoxon test over the 191 groups ($p < 0.05$). The test acts on per-group differences, so a marker can point against the displayed means; it indicates the direction of a change, not the best value in a column.

Same-backbone comparison. Fig. 3(b) and Tab. 3 (App. A) compare four training-free mechanisms re-implemented on Infinity-2B. Three provide no identity gain over the backbone. StoryDiffusion matches SPIN's identity (0.9154 vs. 0.9142 CLIP-I), but its pose and centroid diversity are 0.063 and 0.084, versus 0.075 and 0.095 for SPIN, with lower prompt fidelity on both metrics.

Ablation. Tab. 2 tests modules and placement; blocks (c)–(e) appear in Tab. 4 (App. B). *(a)* Modules cumulatively improve identity from the Infinity-2B backbone (DreamSim $0.235 \rightarrow 0.211 \rightarrow 0.198 \rightarrow 0.182$) while gradually reducing diversity. *(b)* Before scale 3, the bank forms at least half the key set (Eq. 8); attention reaches 0.925 CLIP-I, above other training-free methods (Tab. 1), but lowers prompt fidelity and diversity (up to 26% in pose and 32% in centroid). This trade-off (dashed in Fig. 3b, labelled by first attend scale) reflects identity gains through layout copying; the deployed setting avoids its steep diversity cost. Without rewinding, bank attention roughly follows Eq. (8) (0.31, 0.18, 0.10 at scales 3–5 versus 0.36, 0.21, 0.11 predicted; App. G). *(c)* PWT gains little beyond scale 5, so $\mathcal{W}$ ends there to limit cached source writes. Forcing it on at 1×1 gives little identity gain but lowers both diversity metrics (Sec. 3.7). *(d)* High-frequency rewinding helps most; rewinding the low half or scale 4 instead of scale 3 resembles no rewind, while using all pairs adds little. *(e)* At fixed read strength, six pre-render reads match Ours in DreamSim (0.182) with 7% less pose diversity; six render-stage reads retain diversity but worsen DreamSim (0.192). The even split retains similar diversity and the highest centroid spread; increasing its strength restores identity at a 2% pose-diversity cost, with 8% more pose diversity than all-pre-render reads at similar identity. Removing SRT at scales 1–2 lowers identity without diversity gains; forcing it on at 1×1 mirrors (c).

Module complementarity. Under teacher forcing with fixed codes, we record each module's individual residual increment at every PWT layer (App. D). A group-held-out linear classifier predicts the module from one token's increment direction with balanced accuracy 0.88 (8B) and 0.89 (2B; chance 0.33) at the first PWT layer and above 0.73 throughout. R^{2}SA has low cosine similarity with PWT (≤ 0.19) and SRT (≤ 0.30). PWT and SRT overlap more, consistent with their three shared read/write layers (Sec. 3.4), but peak cosines (0.44 on 2B, 0.38 on 8B) still imply angles above $60°$, consistent with complementary gains in Tab. 2(a).

Position dependence and PWT layer choice. Fig. 4 (left) tests the claim in Sec. 3.3 on source passes. Pre-render writes have positional share η^2 of 0.057 (8B) and 0.346 (2B), below the corresponding residual states (0.123, 0.426)—indicating lower positional dependence in increments than in states—and writes before and after the band (0.203/0.130, 0.528/0.529; $p < 0.001$). Render NCC remains low (0.10 and 0.03, versus 0.24 and 0.14 after the band), supporting transfer before the writes become image-like (App. F). Writes injected at the front PWT layers (right) are amplified $2.0\times$ (8B) and $2.1\times$ (2B) at the output, versus $1.2\times$ at the back ($p < 0.001$). Earlier layers amplify more ($3.4\times$, $3.7\times$) but carry more position.

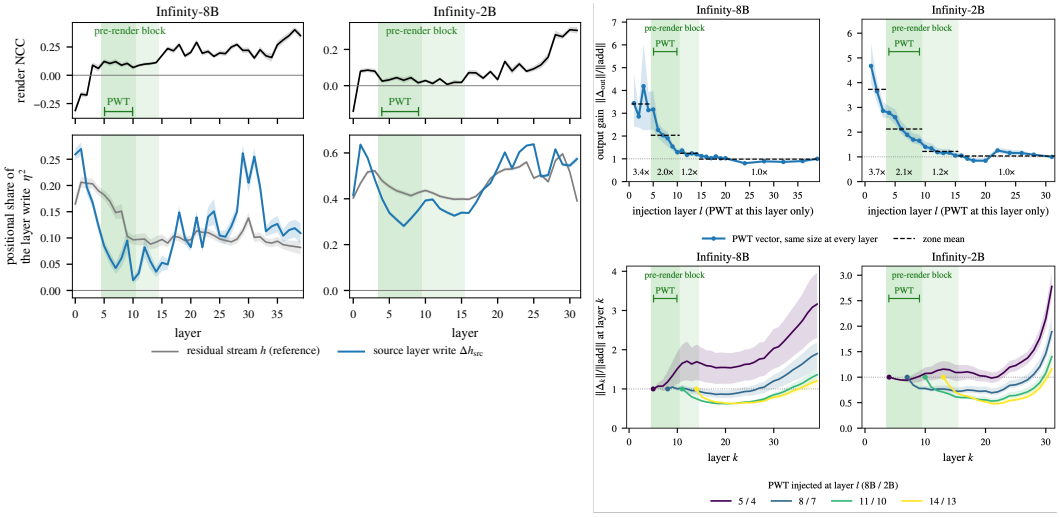

Figure 4: **Why PWT writes in the pre-render layer band, and at its front.** *Left*, source pass: render NCC of each layer's write $\mathrm{d}\mathbf{z}^{\mathrm{src},l}$ decoded through the model head (top), and the share of its variance explained by grid position (η^2, held-out) against the residual stream $\mathbf{h}$ (bottom). *Right*: a write of fixed relative size injected at a single layer l; output gain versus l (top; dashed: zone means) and its norm across depth for four l (bottom). Green: pre-render layer band; dark green: PWT layers. Bands: 95% group-bootstrap intervals. Details in App. E.

Effect of the rewind band. Fig. 9 (App. G) measures the effect of Eq. (7). At scale 3, rewinding the high-frequency half reduces same-cell attention from $12.1\times$ to $3.1\times$ the mean, removes a same-position logit bonus of 2.14, and lowers bank attention from 0.31 to 0.22. The remaining broad preference is consistent with coarse position in low-frequency pairs. The low half leaves the peak at 12.0; all pairs yield little further reduction (3.0), consistent with Tab. 4(d) and the band choice in Sec. 3.5. Preference grows across scales (12.1, 18.8, 36.5 at scales 3–5), but bank share declines (Eq. 8), motivating rewinding only the first attend scale.

Transfer across backbones. Tab. 6 (App. H) applies the same recipe to Switti, HART, STAR, and both Infinity sizes, adjusting only the resolution-anchored first attend scale. Identity improves on all five (DreamSim -12% to -22%; Infinity-2B with SPIN exceeds plain 8B in identity). CLIP-T changes by at most 0.002 on four backbones (STAR: -0.016) and VQA by at most 0.01; pose and centroid diversity retain 76–91% and 75–99% of their backbone values.

User study. Tab. 7 (App. I) compares preferences against Infinite Story, CoDi, CharaConsist, and FLUX.1 Kontext. FLUX.1 Kontext ranks second in identity (21.5%) but last in pose diversity (5.4%); CoDi receives relatively more pose than identity votes. SPIN leads in both, with the largest margins in pose diversity (46.2%) and prompt alignment (51.5%).

5 Conclusion

Stronger source information transfer often couples identity with layout, yielding a common identity–diversity trend across prior methods (Fig. 3). SPIN improves this trade-off in scale-wise VAR by transferring the source–target state gap through three block interfaces with layout-aware placement. On Infinity-2B, it lowers DreamSim from 0.235 to 0.182 while retaining 89% of the backbone's pose diversity; the same recipe improves identity across all five tested next-scale backbones (Tab. 6). Human raters also prefer SPIN for identity, prompt alignment, and pose diversity (Tab. 7). Three measurements guide the design: increments carry less position than states (Fig. 4), source keys' same-position preference concentrates in the high-frequency rotary band (Fig. 9), and the three modules induce largely distinct residual directions, supporting their complementary gains (Tab. 2(a); App. D). Limitations, including inference cost, are discussed in App. K.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to assist with language editing, translation, literature search and summarization, reference formatting, and software implementation and debugging. We also used generative AI tools to provide feedback on the presentation and organization of experiments. We did not use generative AI tools to generate experimental results, to develop the conceptual framework of the proposed method, or to determine the scientific claims or conclusions of this work. Generating synthetic data, cleaning or reformatting datasets, and qualitative or thematic data analysis are not applicable to this work. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was reviewed and revised by the authors. All methodological choices, experimental settings, reported results, and scientific conclusions were determined and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies consistent visual autoregressive generation using publicly available models and benchmark datasets. Our experiments do not involve personally identifiable information or newly collected sensitive data. The user study involves voluntary human evaluation of generated images; participants were informed of the study procedure, and no personally identifiable information was collected. We do not foresee significant societal risks beyond those generally associated with image generation models, and we adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All results use the public ConsiStory+ benchmark (191 prompt groups) with a fixed seed and public backbones (Infinity-2B/8B, Switti, HART, STAR). The three modules are specified by Eqs. (5), (6), (7) and (9); their placement by Sec. 3.7; every layer set, scale set, strength schedule and the bank construction by Sec. 4 (Setup) and App. A; and the metrics, segmentation and diversity definitions by App. C. The protocols of the three analyses are given in App. D, E and G, the baseline re-implementation in App. A, the backbone transfer rule in App. H, and the user-study protocol and interface in App. I. Source code and the evaluation scripts will be released.

REFERENCES

Black Forest Labs. FLUX, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19822–19835, 2021.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Zhanxin Gao, Beier Zhu, Liang Yao, Jian Yang, and Ying Tai. CoDi: Subject-consistent and pose-diverse text-to-image generation. In *International Conference on Learning Representations*, 2026.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15890–15902, 2023.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Mingxiao Li, Mang Ning, and Marie-Francine Moens. Consistent story generation: Unlocking the potential of zigzag sampling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 122630–122656, 2025.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8640–8650, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384, 2024.

Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *International Conference on Learning Representations*, 2025.

Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Biye Li, Huaian Chen, and Yi Jin. STAR: Scale-wise text-conditioned autoregressive image generation, 2024.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16784–16804, 2022.

Jihun Park, Kyoungmin Lee, Jongmin Gim, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, and Sunghoon Im. Infinite-story: A training-free consistent text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8278–8286, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8821–8831, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7986–7996, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950, 2025.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. HART: Efficient visual generation with hybrid autoregressive transformer. In *International Conference on Learning Representations*, 2025.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics*, 43(4):52:1–52:18, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865, 2024.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16058–16067, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110315–110340, 2024.

A IMPLEMENTATION DETAILS

Backbones and schedules. Infinity-2B and Infinity-8B use 13 scales with grids $1, 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, 48, 64$; Switti, HART and STAR use 14 scales with grids $1, 2, 3, 4, 5, 7, 9, 12, 16, 21, 27, 36, 48, 64$ (all $h_s = w_s$, 0-indexed). All images are generated with a fixed seed. The source pass is the group's first image and is identical to an unmodified backbone pass; the recorded quantities are the attention inputs and layer writes at the site set $\mathcal{I}_{s,l}$ and the per-layer key/value cache at the bank positions.

Module placement. PWT writes at layers 4–9 on Infinity-2B and 5–10 on Infinity-8B (the front half of the pre-render layer band identified in App. E) on scales $\mathcal{W} = \{1, \dots, 5\}$, with a uniform per-layer weight $w_l = w_{\text{tot}}/|\mathcal{L}_{\text{PWT}}|$ under $w_{\text{tot}} = 6$ on 2B ($w_l = 1.0$) and 9 on 8B ($w_l = 1.5$), and a_s^{W} decaying linearly from 0.30 at scale 1 to 0.10 at the last scale; the effective coefficient of the write gap in Eq. (9) is $a_s^{\text{W}}\, g_s\, w_l$. SRT reads at six self-attention layers on scales ≥ 1, split evenly between the pre-render layer band and the render stage: $\mathcal{L}_{\text{SRT}} = \{4, 6, 9, 16, 24, 31\}$ on Infinity-2B (pre-render layer band 4–9, render stage 16–31) and $\{5, 8, 10, 15, 27, 39\}$ on Infinity-8B (pre-render layer band 5–10, render stage 15–39), with a_s^{S} decaying linearly from 0.36 at scale 1 to 0.12 at the last scale. R^{2}SA attends from the first scale whose token count exceeds the bank, $n_s > B$, which is scale 3 (6×6) on Infinity and scale 5 (7×7) on the other backbones; the rewind of Eq. (7) is applied at that scale only, over the high-frequency half of the rotary pairs on both axes (pairs 0–15 of the y axis and 32–47 of the x axis under $\theta_j = 10{,}000^{-j/32}$). The bank holds at most $B = 32$ tokens drawn at uniform stride over scales $0, \dots, s$ (all of them when $N_s \leq B$); within each scale's slice a fraction $\rho = 0.375$ is replaced by the positions that receive the most votes across $\{\mathcal{I}_{j,l}\}_l$. The site set uses $k_{\max} = 64$, so that $\mathcal{I}_{s,l}$ is the full grid on scales 1–4 ($n_s \leq 64$) and the 64 most salient tokens from scale 5 ($n_5 = 144$).

Gate. The gate compares the target's accumulated code map at scale s with the source's final one, both continuous maps in the VAE latent space, by a mixture of a global and a local normalised cross-correlation, $\text{sim}_s = 0.35\,\text{NCC}_{\text{global}} + 0.65\,\text{NCC}_{\text{local}}$, where the local term averages the correlation over 8×8 patches at stride 4. It is then thresholded and clamped, $g_s = \text{clamp}((\text{sim}_s - \tau_s)/0.20,\ m_s,\ 1)$ with $\tau_s = 0.08 - 0.06\, s/S$ and $m_s = 0.03$ for $s/S \geq 0.35$ and 0 otherwise, so the threshold is highest at the coarse scales. g_s increases with the similarity: it is 0 at scale 0, 0.91 at scale 1 and 1 from scale 2 on, i.e., injection is withheld until the target's composition has settled rather than reduced for targets that already resemble the source.

Re-implementation of baselines on Infinity-2B (Tab. 3). StoryDiffusion, ConsiStory, CoDi and CharaConsist are ported to Infinity-2B by attaching their attention-level mechanism to the same self-attention layers used by R^{2}SA, keeping each method's published sampling ratio and layer range. Methods that require training, a reference image or a diffusion-specific schedule are compared only in Tab. 1 on their native backbone.

	Subject identity		Prompt fidelity		Diversity	
Method	CLIP-I ↑	DreamSim ↓	CLIP-T ↑	VQA ↑	Pose ↑	Centroid ↑
Infinity-2B (backbone)	0.8945	0.2348	0.3037	0.8311	0.0842	0.1052
StoryDiffusion	**0.9154**	**0.1789**	0.3010	0.8191	0.0629	0.0841
ConsiStory	0.8927	0.2235	0.3032	**0.8308**	0.0860	0.0904
CoDi	0.8930	0.2283	0.3032	0.8305	0.0835	0.1009
CharaConsist	0.8828	0.2630	0.3028	0.8261	**0.0862**	**0.1165**
SPIN (Ours)	0.9142	0.1821	**0.3042**	0.8254	0.0750	0.0945

Table 3: **Same-backbone comparison.** Prior training-free mechanisms re-implemented on Infinity-2B with their published settings (App. A). **Bold**: best among consistency methods.

B REMAINING ABLATIONS

Tab. 4 continues Tab. 2 with the three switches not shown in the main text: the scales that receive the PWT write (c), the rewind band and the scale at which it is applied (d), and the allocation of the

Configuration	Subject identity		Prompt fidelity		Diversity	
	CLIP-I↑	DreamSim↓	CLIP-T↑	VQA↑	Pose↑	Centroid↑
Infinity-2B backbone (no module)	0.8945	0.2348	0.3037	0.8311	0.0842	0.1052
Ours: SRT (3 pre-render + 3 render) + R^2SA (from scale 3) + PWT (scales 1–5)	0.9142	0.1821	0.3042	0.8254	0.0750	0.0945
(c) PWT: written scales $\mathcal{W}$ (ours: 1–5)						
scales 1–3	0.9130	**0.1825**	0.3038	0.8241	**0.0737**	0.0901
scales 1–4	0.9122	**0.1829**	0.3036	0.8259	**0.0750**	0.0915
scales 1–6	0.9129	0.1801	0.3035	0.8243	0.0726	0.0924
scales 1–12	0.9140	0.1800	0.3033	0.8253	0.0738	0.0920
scales 0–5 (forced on at 1×1)	0.9140	0.1801	**0.3037**	0.8274	0.0723	0.0855
(d) R^2SA: rewind band $\mathcal{P}$ and scale (ours: high-frequency half, at scale 3 only)						
high half, at scales 3 and 4	0.9102	**0.1843**	0.3036	0.8233	0.0744	0.0929
high half, at scale 4 only	0.9135	0.1746	0.3034	0.8238	0.0682	0.0857
low half instead of high half	0.9132	0.1772	0.3030	0.8246	0.0676	0.0852
all pairs	0.9105	0.1852	0.3034	0.8250	0.0730	0.0938
no rewind	0.9132	0.1757	0.3033	0.8243	0.0687	0.0882
(e) SRT: allocation of the six read layers between the pre-render layer band (4–9) and the render stage (16–31), and read strength (ours: $3+3$ at a^{S}: $0.36\to0.12$; the three allocation rows use $0.30\to0.10$)						
pre-render 6 + render 0 (4–9)	0.9129	0.1818	0.3035	0.8238	0.0695	0.0912
pre-render 3 + render 3 (4, 6, 9, 16, 24, 31)	0.9089	0.1914	0.3039	0.8261	0.0766	0.0937
pre-render 0 + render 6 (16–21)	0.9099	0.1919	0.3036	0.8269	0.0773	0.0922
scales ≥ 0 (forced on at 1×1)	0.9127	0.1797	0.3038	0.8224	0.0698	0.0862
scales ≥ 3	0.9094	0.1898	0.3042	0.8254	0.0740	0.0920

Table 4: **Ablation on Infinity-2B (remaining blocks).** Continuation of Tab. 2; (c) and (d) test the placement rules of Sec. 3.7 and (e) the read allocation of Sec. 3.4. Markers as in Tab. 2.

six SRT reads between the pre-render layer band and the render stage (e). Each row changes one switch relative to Ours and is discussed in Sec. 4.

C Benchmark and metrics

ConsiStory+ contains 191 prompt groups (1,100 images) that share a subject description and vary the action, background and viewpoint. Subject identity is computed on the segmented subject and averaged over all image pairs of a group, the source included: CLIP-I is the cosine similarity of CLIP image embeddings and DreamSim is the learned perceptual distance, both on the masked crop. Prompt fidelity is computed on the full image: CLIP-T against the target prompt and VQAScore with the target prompt as the question. Diversity is likewise averaged over all image pairs of a group, the source included: pose diversity is the mean pairwise Procrustes distance between subject keypoint configurations, following CoDi (Gao et al., 2026), and centroid diversity is the mean pairwise Euclidean distance between subject-mask centroids, in units of the image side. Both are reported as raw values, with the unmodified backbone providing the ceiling; the ratios in the main text are taken with respect to that ceiling.

D Do the three modules inject the same direction?

Protocol. All passes replay the same sampled codes under teacher forcing at scale 4, so that differences between passes reflect the modules only. For module $m \in \{\mathrm{SRT}, \mathrm{R^2SA}, \mathrm{PWT}\}$ we define the *isolated effect* at layer l as $\mathbf{u}^l_m = \mathbf{h}^l_{\{m\}} - \mathbf{h}^l_{\varnothing}$, the residual state with only m active minus the state with no module, and its *increment* $\mathrm{d}\mathbf{u}^l_m = \mathbf{u}^l_m - \mathbf{u}^{l-1}_m$, what layer l contributes to that effect. At the first PWT layer, $\mathrm{d}\mathbf{u}^l_{\mathrm{PWT}}$ is exactly the injected write gap of Eq. (9). A linear classifier is trained to predict m from the unit direction of a single token's $\mathrm{d}\mathbf{u}^l_m$ (or $\mathbf{u}^l_m$), with prompt groups held out, and we report balanced accuracy; the cosine panel averages the cosine between the two modules' increments over tokens.

Results (Fig. 5). At the first PWT layer the increment classifier reaches 0.88 on Infinity-8B and 0.89 on Infinity-2B (chance 0.33); it peaks at the second PWT layer (0.91, 0.90) and settles at 0.73–0.76 (8B) and 0.75–0.79 (2B) over the back half of the block, where the classifier on the accumulated isolated effects stays slightly higher (0.81–0.83 and 0.77–0.82). R^2SA is the most distinct module throughout: its cosine with PWT never exceeds 0.19 (8B) or 0.17 (2B), and with SRT it rises only from 0.15 to 0.29 (8B) and from 0.18 to 0.30 (2B) across the block. The PWT–SRT cosine is the

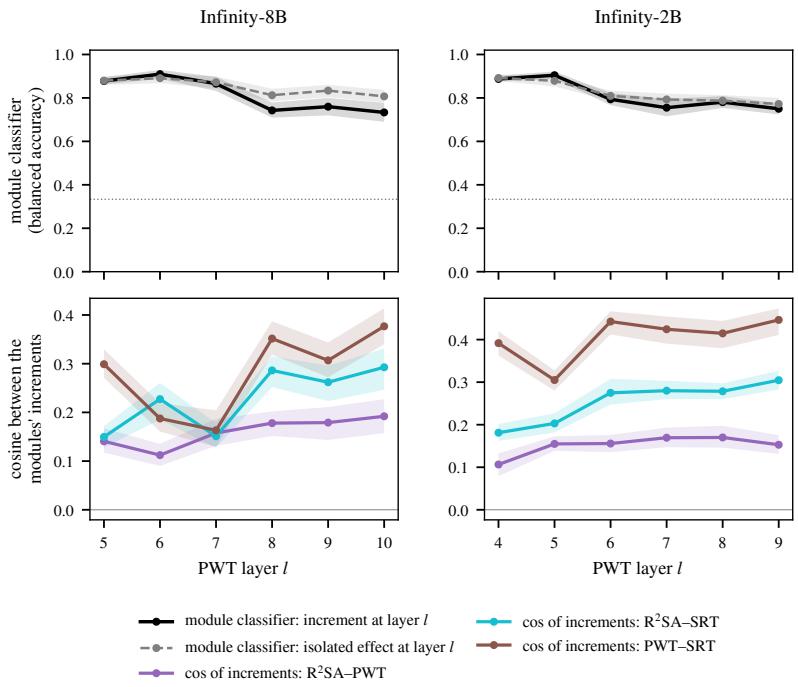

Figure 5: **What each module adds at the PWT layers.** Top: balanced accuracy of a group-held-out linear classifier that predicts the module from the direction of a token's increment $\mathrm{d}\mathbf{u}_m^l$ (solid) or isolated effect $\mathbf{u}_m^l$ (dashed); dotted: chance. Bottom: cosine between the modules' increments; the PWT–SRT curve peaks at the PWT layers that also carry an SRT read (4, 6, 9 on 2B). Bands: 95% group-bootstrap intervals.

largest of the three because three SRT reads are placed at PWT layers (App. A): on 2B it is 0.39 at layer 4, dips to 0.31 at layer 5, where SRT does not read, and returns to 0.42–0.44 from layer 6, the next SRT layer, onward; on 8B it is 0.30 at layer 5, falls to 0.16–0.18 at layers 6–7 and rises to 0.31–0.38 at layers 8–10, again the layers that carry an SRT read (5, 8, 10; App. A). Even at the peak the two increments are more than 60° apart, and the residual confusion of the classifier is between PWT and SRT, the two modules that transfer a source–target gap, rather than with R^{2}SA. Compared with placing the reads away from the write layers, co-locating them raises the PWT–SRT overlap but is what the allocation in Tab. 4(e) selects. These measurements characterise the representation under fixed codes, not the final image; the behavioural contribution of each module is the ablation in Tab. 2(a).

E Locating the pre-render layer band

This appendix gives the full measurements behind Fig. 4. All left-panel quantities are computed on the source pass alone over scales 1–4; the right-panel quantities use scales 1–5 and 12 prompt groups.

Render NCC. Each layer's write $\mathrm{d}\mathbf{z}^{\mathrm{src},l}$ is decoded independently through the model head and the VAE and compared with the final source image by normalised cross-correlation (the global term of the gate g_s in App. A, applied between a decoded write and the final image). The writes preceding the block are not image-like (-0.10 on 8B, 0.03 on 2B), the block stays low (0.10, 0.03), and the NCC rises after it (0.24, 0.14); render NCC therefore fixes the upper boundary of the block only.

Positional share. η^2 is the held-out fraction of variance in $\mathrm{d}\mathbf{z}^{\mathrm{src},l}$ explained by grid position. It is 0.203 before, 0.057 inside and 0.130 after the block on 8B, and 0.528, 0.346 and 0.529 on 2B (block vs. either side, $p < 0.001$, group bootstrap). Position is not absent inside the block: it remains



Figure 6: **The write gap decoded at each scale (prompt group 21).** Top: the source and the target generated by the full model. Below: the full-token write gap $\mathrm{d}\mathbf{z}^{\mathrm{src},l} - \mathrm{d}\mathbf{z}^{\mathrm{tgt},l}$ at a PWT layer ($l = 9$) and a late layer ($l = 30$), decoded through the model head and the VAE at scales 1–5 (2×2 to 12×12). Cell values: render NCC of that layer's source write at the scale; row labels: the mean over the five scales.

linearly decodable from the residual stream (held-out R^2 0.90 and 0.96). The point is relative: inside the block the write carries a smaller positional share than the stream (0.057 vs. 0.123; 0.346 vs. 0.426), whereas before and after the block it carries a comparable or larger share. This is the empirical content of the claim in Sec. 3.3 that an increment is less position-bound than a state, and it identifies the block as a whole but not its front, where η^2 is similar to the back.

Lever strength. To choose the front, the write gap is injected at a single layer l with the same relative norm at every l, the codes are kept fixed, and the change Δ_k is tracked at every downstream layer k.[1] The output gain $\|\Delta_{\mathrm{out}}\|/\|\mathrm{add}\|$ is $2.03\times$ (8B) and $2.13\times$ (2B) for the PWT layers, $1.24\times$ and $1.22\times$ for the back of the block, and $\approx 1.0\times$ after it; every transition between adjacent zones is significant ($p < 0.001$). The depth profiles show where the difference arises: a write injected at the first PWT layer is preserved or amplified throughout the block, whereas later injections first shrink to 0.5–$0.9\times$ in the middle layers before the final layers amplify all injections by a similar factor. Layers before the block amplify more ($3.4\times$, $3.7\times$) but carry the most position, so they are excluded by the positional criterion. This analysis measures how strongly the network propagates a write, not the direction of the change or its effect on the image.

F What PWT transfers, decoded

Fig. 6 visualises the transferred quantity by decoding the write gap directly; this is a visualisation only and not part of the method. The maps show the full-token difference, whereas the deployed injection is scaled by $a_s^{\mathrm{W}} g_s w_l$, gated by τ_s and, from scale 5, restricted to $\mathcal{I}_{s,l}$. At the PWT layer no recognisable structure appears at any scale (render NCC 0.01 on average, -0.02 to 0.05 per scale). At the late layer the source subject is already visible, including silhouette, dress and hair, at NCC 0.32, rising from 0.11 at 2×2 to 0.34–0.39 from 4×4 onward. A late write therefore already contains the rendered source, whereas a pre-render write does not, even though the latter is the write that shifts identity when transferred (Tab. 2).

[1] Replaying the same codes twice changes the state by at most 2.5×10^{-4} of its norm, far below the injected sizes.

Scale	Rewind	same-position preference	bank mass	centroid distance (cells)	centroid corr.	logit bonus at 0
3	no rewind	12.094 [11.700, 12.493]	0.307 [0.301, 0.312]	3.240 [3.208, 3.269]	0.183 [0.174, 0.193]	0.000 [0.000, 0.000]
3	deployed rewind (high half)	3.073 [2.826, 3.305]	0.223 [0.220, 0.227]	4.252 [4.225, 4.277]	-0.052 [-0.058, -0.047]	-2.143 [-2.167, -2.119]
3	low half	12.029 [11.639, 12.422]	0.305 [0.299, 0.310]	3.257 [3.226, 3.285]	0.178 [0.169, 0.187]	-0.011 [-0.012, -0.011]
3	all pairs	3.038 [2.794, 3.269]	0.223 [0.219, 0.226]	4.269 [4.244, 4.294]	-0.056 [-0.062, -0.051]	-2.154 [-2.179, -2.131]
4	no rewind	18.818 [18.125, 19.514]	0.184 [0.181, 0.187]	7.310 [7.243, 7.371]	0.057 [0.049, 0.063]	0.000 [0.000, 0.000]
4	deployed rewind (high half)	6.051 [5.678, 6.418]	0.145 [0.142, 0.147]	8.460 [8.400, 8.519]	-0.049 [-0.056, -0.044]	-1.833 [-1.864, -1.802]
4	low half	18.705 [18.019, 19.400]	0.183 [0.180, 0.186]	7.333 [7.267, 7.393]	0.054 [0.047, 0.061]	-0.012 [-0.013, -0.012]
4	all pairs	6.034 [5.668, 6.396]	0.144 [0.142, 0.146]	8.487 [8.429, 8.544]	-0.051 [-0.057, -0.045]	-1.844 [-1.877, -1.813]
5	no rewind	36.467 [34.798, 38.029]	0.101 [0.099, 0.102]	12.684 [12.556, 12.798]	-0.025 [-0.039, -0.009]	0.000 [0.000, 0.000]
5	deployed rewind (high half)	8.242 [7.581, 8.909]	0.081 [0.080, 0.082]	14.332 [14.272, 14.396]	-0.034 [-0.037, -0.031]	-4.593 [-4.745, -4.435]
5	low half	36.488 [34.810, 38.057]	0.100 [0.098, 0.101]	12.698 [12.571, 12.812]	-0.025 [-0.039, -0.009]	-0.009 [-0.010, -0.008]
5	all pairs	8.249 [7.588, 8.912]	0.081 [0.080, 0.081]	14.345 [14.287, 14.409]	-0.033 [-0.037, -0.030]	-4.602 [-4.755, -4.443]

Table 5: **The rewind, by scale and band.** Same-position preference, total attention on the bank, the distance between a target token's own position and the attention-weighted mean bank position, the correlation between the two, and the same-position logit bonus relative to the no-rewind setting. Point estimates with 95% group-bootstrap intervals over 17 prompt groups.

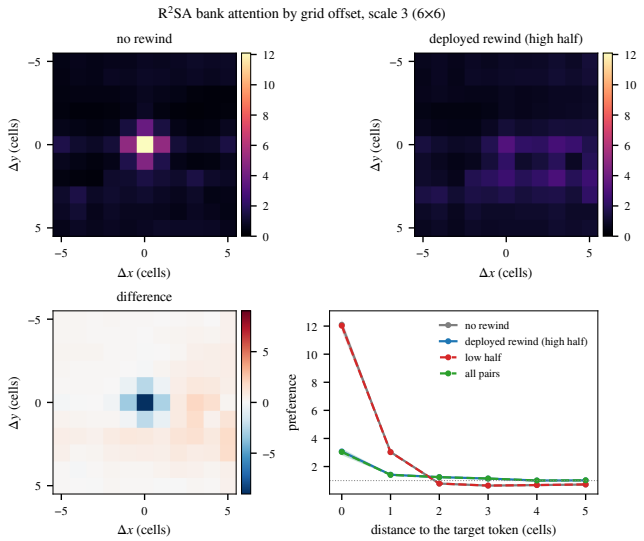

Figure 7: **Offset profile of the bank attention at scale 3.** Mean attention on a bank token at each grid offset, divided by the mean over all offsets: without the rewind (top left), with the deployed rewind (top right), and their difference (bottom left). Bottom right: the same preference against the distance between the two tokens, for the four rewind settings; the curves collapse into two pairs, no rewind with the low half and the deployed rewind with all pairs. Bands: 95% group-bootstrap intervals.

G THE ROPE REWIND

How the rewind is measured. At every attend scale and layer we take the recorded queries and the full key set of Eq. (6) – the bank keys first, then the token's own keys and the cache – and recompute the attention with the model's own softmax scale, which is exact at these grid sizes. The bank keys are then rotated or de-rotated offline to obtain the four settings, so they differ only in the phase of those keys and nothing else in the pass changes. The same-position preference is the attention assigned to the bank token at the query's own grid cell divided by the mean attention over all offsets, averaged over tokens, layers and 17 prompt groups.

Results (Tab. 5, Figs. 9, 7 and 8). The effect of the deployed rewind is of the same order at every attend scale ($12.1 \to 3.1$ at scale 3, $18.8 \to 6.1$ at scale 4, $36.5 \to 8.2$ at scale 5), and it also reduces the total attention on the bank ($0.31 \to 0.22$, $0.18 \to 0.15$, $0.10 \to 0.08$). At scale 3 the attention-weighted bank position moves from 3.2 to 4.3 cells away from the target token's own position, and its correlation with that position falls from 0.18 to -0.05. The remaining preference is broad (Fig. 7, top right and bottom right): the low-frequency pairs encode coarse position, so

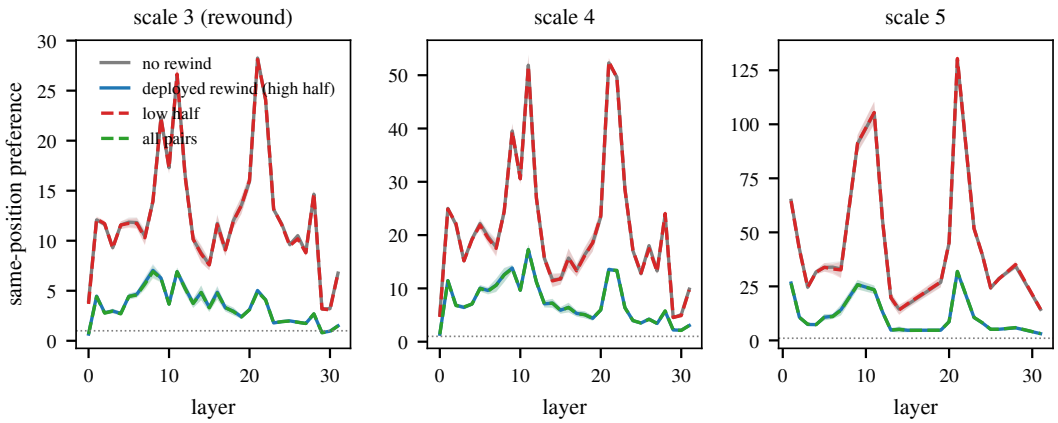

Figure 8: **Same-position preference layer by layer.** The value at offset $(0, 0)$ for every layer, at the rewound scale (left) and at the two later scales, where the deployed model does not rewind; note the different vertical scales. The four settings form the same two pairs at every layer, and the preference is concentrated in a few layers rather than spread evenly. Bands: 95% group-bootstrap intervals.

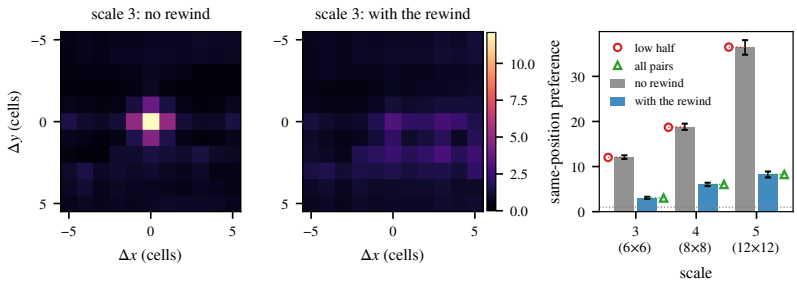

Figure 9: **The rewind removes the same-position preference.** Attention from a target token to a bank token at scale 3 as a function of their grid offset, normalised by the mean over offsets (1: no preference), without (left) and with (middle) the deployed rewind. Right: the preference at offset $(0, 0)$ per scale, with two control bands (low half: red; all pairs: green). Only scale 3 is rewound in the deployed model. 95% group-bootstrap intervals over 17 groups.

the rewind removes fine alignment rather than position altogether, which is the limitation stated in Sec. 3.5. The two control bands stay within 0.11 of the settings they match at every scale, against a rewind effect of 9–28, so the fine positional alignment is carried by the high-frequency pairs alone; this is the attention-level counterpart of the band ablation in Tab. 4(d), where the low half behaves like no rewind and all pairs like the deployed model. Fig. 8 shows that the preference is not uniform over depth: it peaks in the middle layers and falls to about 1 in the last ones, while the pairing of the bands holds at every layer. Because the preference grows with scale while the bank's share of attention shrinks by Eq. (8), the rewind is applied only at the first attend scale; moving it to scale 4 raises identity and lowers diversity in Tab. 4(d), i.e., it behaves like no rewind. Two limits are worth stating: these are attention-level counterfactuals on the deployed pass, not images generated under another band, and the deployed recipe rewinds the first attend step only, so the preference at scales 4 and 5 is left in place.

H TRANSFER ACROSS BACKBONES

The only quantity that cannot be copied verbatim between backbones is the first attend scale, because the scale schedules differ (App. A). We therefore anchor it to resolution: R^2SA is enabled from the first scale with at least 6×6 tokens, which is index 3 (6×6) on Infinity and index 5 (7×7) on Switti, HART and STAR. This leaves the coarsest scales, at which the layout is committed, unaffected on every backbone. The relative layer rule (front half

	Subject identity		Prompt fidelity		Diversity	
Backbone	CLIP-I↑	DreamSim↓	CLIP-T↑	VQA↑	Pose↑	Centroid↑
Switti	0.8732	0.2839	0.3194	0.8197	0.1118	0.1362
+ ours	0.8895	0.2402	0.3189	0.8172	0.0852	0.1140
HART	0.8817	0.2742	0.3134	0.8277	0.0930	0.1341
+ ours	0.8916	0.2415	0.3116	0.8215	0.0808	0.1000
STAR	0.8793	0.2807	0.3185	0.8270	0.1142	0.1009
+ ours	0.9034	0.2229	0.3028	0.8176	0.1001	0.0995
Infinity-2B	0.8945	0.2348	0.3037	0.8311	0.0842	0.1052
+ ours	0.9142	0.1821	0.3042	0.8254	0.0750	0.0945
Infinity-8B	0.8958	0.2299	0.3209	0.8575	0.1008	0.1097
+ ours	0.9142	0.1886	0.3219	0.8607	0.0913	0.0954

Table 6: **Transfer to other next-scale backbones.** Same layer rule and gate on every backbone; only the first attend scale is set by resolution ($\geq 6\times6$) because the schedules differ.

of the pre-render layer band) and the gate are unchanged; no other quantity is searched per model. Per backbone, the gains in Tab. 6 are $+0.016/+0.010/+0.024/+0.020/+0.018$ CLIP-I and $-15\%/-12\%/-21\%/-22\%/-18\%$ DreamSim for Switti/HART/STAR/Infinity-2B/Infinity-8B; pose diversity retains $76\%/87\%/88\%/89\%/91\%$ and centroid diversity $84\%/75\%/99\%/90\%/87\%$ of the respective backbone.

I User study details

I.1 Study setup

We conducted a user study to evaluate **identity consistency**, **prompt alignment**, **pose diversity**, and **image quality** in multi-image generation. We compare five methods: Infinite Story, CoDi, CharaConsist, FLUX.1 Kontext, and our proposed method, SPIN.

We randomly selected 10 subjects from the benchmark dataset, covering diverse categories including humans, animals, and objects. For each subject, three prompts were randomly selected, and each method generated a sequence of three images corresponding to these prompts. Participants were shown the five generated sequences and asked to select the best-performing sequence according to each of the four evaluation criteria described below.

To avoid positional and method-order bias, the five methods were anonymized as A–E and their row order was randomized independently for each question. Therefore, the method corresponding to a given label (e.g., A) could differ across questions. Participants were not informed of the method identities.

A total of 52 participants were recruited for the study. The study was conducted online through Google Forms. All participants provided informed consent before participation. Participation was voluntary, and all responses were collected anonymously.

I.2 Evaluation criteria

Participants evaluated each three-image sequence according to the following four criteria.

1. Identity consistency: Participants were asked to determine whether the subject continued to appear as the same subject across all three images, even as the scene changed. For human or character subjects, they were instructed to consider facial structure, hairstyle, color, and other distinctive characteristics. For animals and objects, they considered overall shape, patterns, relative proportions, and other identifying features.

Participants were instructed not to regard sequences in which all images appeared nearly identical as a desirable result. They were asked to consider whether each image naturally depicted a different situation or scene while preserving enough of the subject's distinctive characteristics for it to remain recognizable as the same subject across the sequence.

2. Prompt alignment: Participants were asked to carefully inspect all three images and select the method whose images best reflected the content and intent of the corresponding text prompts. They were instructed to consider whether prompt-specified elements such as objects, backgrounds, actions, colors, mood, and style were properly represented.

Participants were further instructed not to judge alignment solely by the presence of individual keywords, but also by whether the intended scene or situation was expressed naturally and coherently in the generated image.

3. Pose diversity: Participants were asked to select the method whose three images exhibited the greatest diversity in the subject's pose, action, and position within the image. They considered variations in facing direction, body posture, limb movement, gaze, action, and spatial placement, such as whether the subject appeared on the left, center, or right of the image, or at different depths within the scene.

Participants were instructed not to treat changes caused only by camera zoom or background variation as meaningful pose diversity. Instead, they were asked to consider whether the subject's pose, action, and spatial arrangement changed substantially across the three images while remaining natural in each scene.

4. Image quality: Participants were asked to select the method that produced the images with the highest overall visual quality and completeness. They considered whether the images appeared clear and natural, whether colors and composition were harmonious, and whether the subjects and backgrounds were rendered without noticeable artifacts, distortions, or malformed regions.

Participants were also asked to consider the overall level of visual polish and aesthetic appeal across the three generated images.

For each criterion, participants selected exactly one of the five methods. The study therefore followed a forced-choice protocol, with one preferred method selected for each subject and evaluation criterion.

I.3 User study interface

Figure 10 shows an example of the Google Forms interface used in the study. For each question, the five methods were displayed as rows labeled A–E, with each row containing a sequence of three generated images. The assignment of methods to rows was randomized across questions, so the method corresponding to a particular row label was not fixed throughout the study.

Participants compared all five rows before selecting the best-performing method for identity consistency, prompt alignment, pose diversity, or image quality.

I.4 Evaluation protocol

For each of the 10 subjects and each of the four evaluation criteria, we recorded the method selected by each participant as the best-performing method. The preference score for each method was computed as the percentage of votes assigned to that method over all participant–subject evaluations for the corresponding criterion; Tab. 7 reports the results.

Method	Identity	Text	Pose	Quality
Infinite Story	15.38	16.92	7.69	16.15
CoDi	9.23	4.62	18.46	9.23
CharaConsist	12.31	<u>18.46</u>	<u>22.31</u>	<u>22.31</u>
FLUX.1 Kontext	<u>21.54</u>	8.46	5.38	13.08
SPIN (Ours)	**41.54**	**51.54**	**46.15**	**39.23**

Table 7: **Human preference (%).** 52 participants, 10 subjects, three prompts each; **bold** / <u>underline</u>: best / second best.

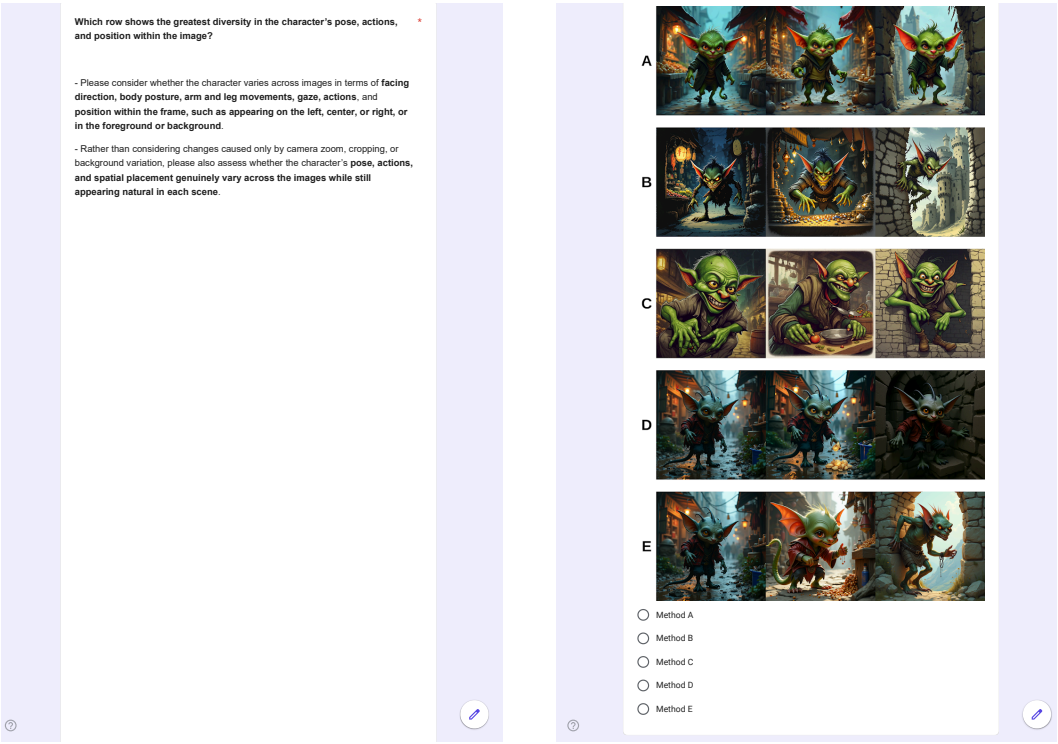

Figure 10: **User study interface.** Example of the Google Forms interface. Each row contains a three-image sequence from one method. Methods are anonymized as A–E, and their row order is randomized across questions.

Table 8: **Inference cost.** Wall-clock time per generated image and peak GPU memory. SPIN introduces moderate overhead from the source pass and activation capture, while requiring only a small increase in peak memory.

Model	Method	Time / image (s)	Peak memory (GiB)
Infinity-2B	Vanilla	1.32	15.75
	SPIN	2.66 (2.02×)	16.55 (+0.80)
Infinity-8B	Vanilla	2.54	37.08
	SPIN	3.84 (1.51×)	38.41 (+1.33)

J Inference cost

We additionally report the inference overhead of SPIN in terms of wall-clock time and peak GPU memory. All measurements are performed on the same prompt groups using a single NVIDIA Blackwell GPU. Timing includes only the generation call, excluding model loading, and the first generated image is discarded as warm-up. Since SPIN first generates the source image and records the required activations before generating each target image, the reported per-image time includes the source pass and the corresponding state/write capture overhead.

K Limitations

Inference cost. SPIN adds a source pass and activation capture, increasing runtime per image to 1.5–2.0× and peak memory by 3–5% (Tab. 8, App. J). The overhead is bounded rather than sequence-proportional: the bank holds a fixed 32 key/value tokens per layer per attend scale, gathered in place from the backbone's own cache, so it adds a constant term to the key length of Eq. (6);

the SRT and PWT records are captured once per group during the source pass, read by every target of that group, and released when the group ends.

Residual positional preference. The rewind of Eq. (7) removes the source-side rotary phase from the bank keys but leaves the query-side phase intact (Sec. 3.5); shared by all bank keys of a query, it does not favour any one source position, but it lets the bank's overall affinity vary with the query's position. The low-frequency pairs also retain coarse position, so the rewind removes fine alignment rather than position altogether (App. G).

L Additional Qualitative Results

We provide additional qualitative comparisons across diverse subjects and prompts in Figs. 11–14. Each figure compares SPIN with existing consistent image generation methods under the same prompt group, further illustrating the identity–diversity behavior observed in Fig. 3. Across human, animal, and robotic subjects, SPIN largely preserves the prompt-driven pose and composition variation of vanilla Infinity while maintaining a consistent subject identity. In contrast, existing methods often exhibit either restricted pose variation or visible identity drift, as detailed in the figure captions.



Figure 11: **Additional qualitative comparison (1/4).** SPIN preserves the wizard's characteristic appearance, including his hat, beard, and clothing, while largely retaining the pose and scene variation of vanilla Infinity across the target prompts. In contrast, Infinite Story shows more limited pose and scene variation and fails to clearly place the subject on the mountain peak specified by the second prompt. CharaConsist and OmniControl introduce visible changes in identity-defining details, such as the wizard's hat, across the generated images.



Figure 12: **Additional qualitative comparison (2/4).** SPIN preserves the dog's characteristic appearance while largely retaining the pose variation of vanilla Infinity across different actions. In contrast, Infinite Story, FLUX.1 Kontext, and Zigzag tend to produce similar upright poses and compositions across the target prompts, resulting in limited pose variation. StoryDiffusion, CoDi, and PhotoMaker exhibit greater changes across the prompts but show noticeable identity drift, including variations in facial appearance, ear shape, and coat patterns.



Figure 13: **Additional qualitative comparison (3/4).** SPIN largely preserves the diverse poses and scene compositions of vanilla Infinity while maintaining a more consistent identity of the robotic dog. In contrast, Infinite Story, OmniControl, IP-Adapter, and Zigzag tend to retain similar poses and body configurations across different prompts, resulting in limited pose variation. ConsiStory, CoDi, and CharaConsist allow larger changes in pose and composition but exhibit noticeable identity drift, including changes in the robot's head, body structure, and overall design.



Figure 14: **Additional qualitative comparison (4/4).** SPIN largely preserves the pose variation of vanilla Infinity, including changes in the subject's viewing direction, head orientation, and body pose across different scenes. In contrast, Infinite Story, CharaConsist, OmniControl, FLUX.1 Kontext, and Zigzag tend to maintain similar frontal poses and viewing directions across the target prompts, resulting in more limited pose variation.