SEEDS TEACH EACH OTHER THROUGH PROMPT TOKENS: REWARD-FREE COMPOSITIONAL TEXT-TO-IMAGE SELF-IMPROVEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

When a text-to-image model samples one compositional prompt with several seeds, different samples get different requirements right: one has the correct count, another the correct spatial relation. Reward and preference fine-tuning score each image as a whole, so they raise the probability of the best complete image and discard the correct parts of the others, and the parts cannot be mixed in pixel space because each seed lays out its objects differently. We observe that joint-attention transformers already provide positions that correspond across seeds: the prompt tokens. After one joint-attention block, the state of each prompt token has absorbed what its own sample rendered for that word, while its position denotes the same word in every sample. VOTEX (Visual-to-Text Exchange) uses this correspondence during training. For G samples of a prompt, the next block lets every image query of a target sample attend to the prompt-token states of all G samples, with a $-\log G$ logit shift that keeps attention unchanged when the samples coincide. The resulting prediction is a dense target that we distill into the ordinary single-sample model, so architecture and inference cost stay the same. We prove that query-wise reading covers K requirements with $O(\log(K/\delta))$ samples under a fixed evidence margin, whereas whole-image selection needs exponentially many, and that the $-\log G$ shift is the unique scale-preserving correction. On SD3.5 Medium, FLUX.1-dev, and Qwen-Image-2512, VOTEX improves six compositional benchmarks by 2.2 to 8.9 points without any reward model, with all 18 run-level 95% intervals above zero. Filling the group with copies or other prompts removes at least 90% of the gain, and masking the sample that rendered a missed requirement correctly lowers accuracy on it four times more than masking an unrelated sample.

1 INTRODUCTION

Text-to-image transformers such as SD3.5, FLUX.1, and Qwen-Image render single objects with high fidelity, yet they still miss requirements when a prompt combines several of them (Esser et al., 2024; Black Forest Labs, 2024; Wu et al., 2025). The GenEval2 prompt "four white bicycles in front of three plastic cows" contains seven atomic requirements on objects, counts, colors, materials, and a spatial relation (Kamath et al., 2025), and a sample often gets one of them wrong. Sampling the prompt with several seeds shows a clear regularity: *the misses differ from seed to seed*. One sample draws three cows beside the bicycles; another gets the relation right and draws two cows. Within a few samples, the correct requirements are usually all present, spread over different images (Section 2, Figure 2).

Current post-training learns from whole images and cannot use this spread. Reward fine-tuning (Black et al., 2024; Fan et al., 2023; Clark et al., 2024), preference optimization (Wallace et al., 2024), and group-relative policy optimization (Liu et al., 2025; Luo et al., 2026; Li et al., 2026a) assign one score or rank to each sample and raise the likelihood of the better ones. Such a signal identifies which image is better and carries no information about which region of a worse image is correct. The correct count of one sample and the correct relation of another reach the model only through a sample that contains both, which for K requirements of probability p appears with probability p^K: the signal is rarest on the prompts that most need it. The correct parts cannot be pasted together either, because

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



Figure 1: Base and VoTEX samples for four compositional prompts, four seeds each. The prompts fix counts and vertical order, bind attributes to objects, place objects, and order colors by depth. VoTEX keeps the variation across seeds and satisfies these requirements in more of the samples.

independent seeds place the objects at different positions, so the same patch index shows unrelated content in two samples.

Joint-attention transformers already contain positions that correspond across seeds, namely the prompt tokens. In an MMDiT block, image and text tokens attend to each other (Esser et al., 2024), so after one block the state of the token "cows" encodes what its own sample rendered for "cows" (Shin et al., 2025; Li et al., 2026b), and this token sits at the same position in every sample. Prompt tokens therefore index a group of samples *by word*, independently of layout: an image region that renders the cows can look up the "cows" token of every sample and use the one that matches it best.

Visual-to-Text Exchange (VoTEX) turns this observation into a training method. For each prompt, the current model samples G trajectories. Of two adjacent joint-attention blocks, the first runs unchanged, so each sample's prompt tokens absorb its own image. In the second, every image query of a target sample attends to its own image tokens and to the prompt tokens of all G samples, with a $-\log G$ shift of the text logits that leaves attention unchanged when the samples coincide. The rest of the network outputs a velocity for the target that has consulted the whole group; this dense prediction becomes the regression target of the ordinary single-sample path at the same noisy state. After training, the group path is removed, and the model samples one image with the original architecture, sampler, and cost. No reward model, preference label, or image decoding enters training.

On SD3.5 Medium, FLUX.1-dev, and Qwen-Image-2512, VoTEX improves all six compositional benchmarks by 2.2 to 8.9 points and exceeds five reward-based fine-tuning methods in every column they report (Table 1). Filling the group with copies of the target or with samples of other prompts removes at least 90% of the gain, and masking the one sample that rendered a missed requirement correctly lowers accuracy on it by 19.9 points, four times the drop from masking an unrelated sample. Our contributions are threefold: **(i)** the measurement that correct requirements are spread across samples, with prompt-token states as their shared, word-indexed coordinates; **(ii)** VoTEX, a normalized query-wise read of the group's prompt tokens distilled into the ordinary model, with coverage, routing, and transfer guarantees; and **(iii)** controls on three backbones that connect the gains to complementary samples.

2 Correct Requirements Are Spread Across Samples

Setting. A pretrained denoiser $f_\theta(x_t, t, c)$ maps a noisy latent x_t, a time t, and a prompt c to a velocity through a stack of joint-attention (MMDiT) blocks (Esser et al., 2024). For one prompt, G seeds give synchronized states $\mathcal{X}_t = \{x_t^g\}_{g=1}^G$. The benchmarks decompose a prompt into atomic

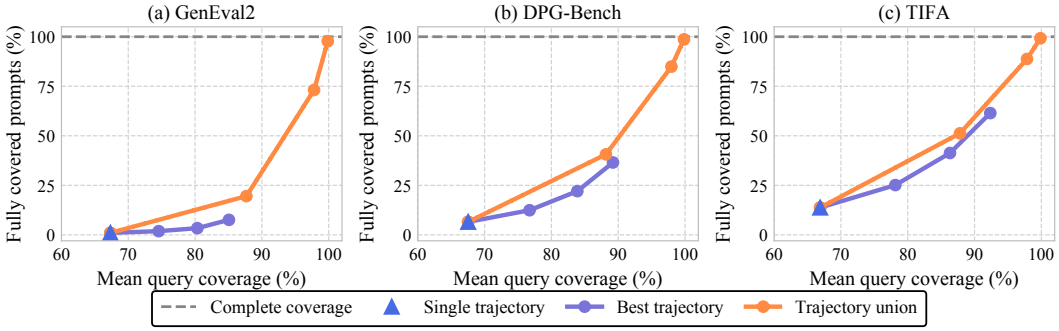


Figure 2: Query coverage versus complete-prompt success for $G = 1, 2, 4, 8$ samples per prompt. Curves connect the best single sample (purple) or the union of the samples (orange); the triangle is one sample. On GenEval2, more samples barely raise the complete success of the best sample, while their union covers almost all prompts at $G = 8$.

requirements $q \in Q_c$ (an object, an attribute, a count, or a relation), each checked by a question (Kamath et al., 2025; Hu et al., 2023; 2024). Let $z_{gq} \in \{0, 1\}$ record whether sample g satisfies atom q. We seek a model that satisfies more atoms at once from one seed, learned without rewards or preference labels.

Complementarity. Two coverages separate a group with one strong image from a group whose correct atoms are split across images. The union coverage $U_c(G) = |Q_c|^{-1} \sum_q \max_g z_{gq}$ counts an atom as covered if any sample satisfies it, and the best-member coverage $B_c(G) = \max_g |Q_c|^{-1} \sum_q z_{gq}$ keeps only the best single image. Their difference $D_{\text{comp}} = U_c - B_c$ is positive only when correct atoms are distributed across samples; finding one better image leaves it unchanged. Before any training, with $G = 4$, D_{comp} ranges from 2.75 to 15.40 points across the nine backbone–benchmark pairs, while the same statistic with samples from other prompts stays between 0.02 and 0.07. On GenEval2, going from three to ten atoms per prompt raises D_{comp} from 1.86–7.16 to 12.47–21.67 points, while the fraction of single samples that satisfy the whole prompt falls from 65.3–89.3% to 1.0–7.0%.

Figure 2 shows what this spread means for learning. From $G = 1$ to 8, the complete success of the best single sample on GenEval2 rises only slightly, whereas the union of the same samples covers nearly all prompts at $G = 8$; DPG-Bench and TIFA, with fewer coupled atoms, show a smaller gap. A method that selects or upweights whole images learns from the purple curve, while the evidence for complete prompts lies on the orange curve. With K independent atoms that each hold with probability p, whole-image selection needs

$$G_{\text{whole}} = \left\lceil \frac{\log \delta}{\log(1 - p^K)} \right\rceil = \Theta\left(p^{-K} \log \frac{1}{\delta} \right) \tag{1}$$

samples to observe one complete success with probability $1 - \delta$, which is exponential in K. Reusing the correct atoms of different samples requires only that each atom appear somewhere in the group, which needs $O(\log(K/\delta))$ samples under a fixed evidence margin (Theorem 1). Because image positions do not correspond across samples, VoteX exchanges this evidence through prompt tokens: token ℓ is the same word in every sample, and after one joint-attention block its state has read that sample's image.

3 Method

VoteX turns the spread of correct atoms into a training target for the ordinary model (Figure 3). Its three steps answer the three needs of the problem: a correspondence across samples, a per-region choice among them, and a single-sample model at deployment.

Notation. In block b, sample g has image-token states $I_b^g \in \mathbb{R}^{N \times d}$ and prompt-token states $E_b^g \in \mathbb{R}^{L \times d}$. Per head of width d_h, image tokens give Q_g^I, K_g^I, V_g^I and prompt tokens give K_g^T, V_g^T. We call (K_j^T, V_j^T) the text bank of sample j, and the samples $j \neq g$ the donors of target g. A frozen copy $\bar{\theta}$ of the current parameters produces the group and the teacher prediction $T_{\bar{\theta}}$; sg stops gradients.

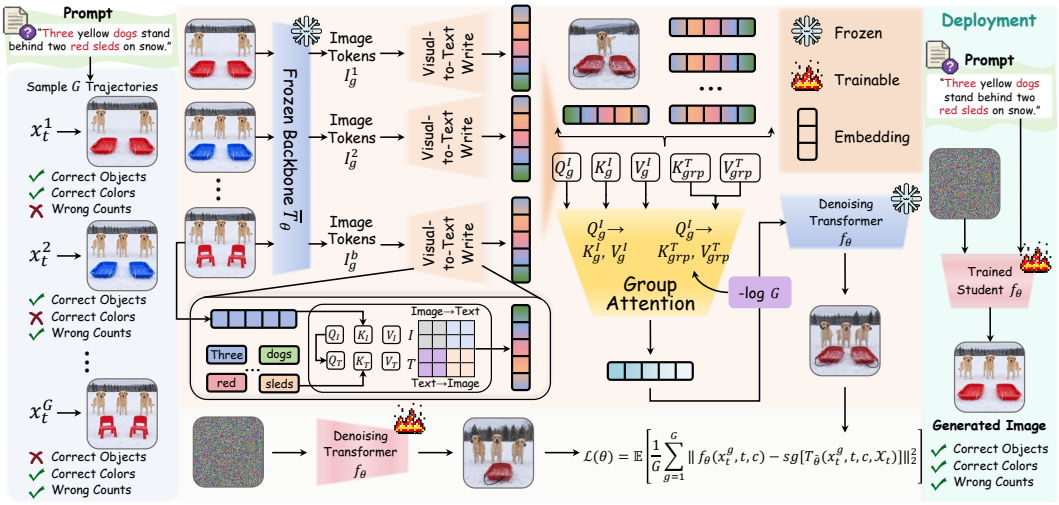


Figure 3: Overview of VOTEX. A frozen snapshot samples G trajectories for one prompt; each gets different requirements right. Block b writes each sample's image into its own prompt-token states. In block $b{+}1$, the image queries Q_g^I of target g attend to its own image keys and values and to the text keys and values $K_{\mathrm{grp}}^T, V_{\mathrm{grp}}^T$ of all samples, with $-\log G$ added to the text logits. The frozen suffix outputs the teacher prediction, which supervises the trainable single-sample model f_θ through Equation (5). Deployment uses f_θ alone.

3.1 STEP 1: PROMPT TOKENS RECORD WHAT EACH SAMPLE RENDERED

The text encoder output E_0^g depends only on the prompt and is identical for all seeds. In every block up to and including b, which runs unchanged, the text queries attend to the image tokens of their own sample, giving $E_{b+1}^g = \mathrm{TextBlock}_\theta^b(E_b^g, I_b^g, t)$. The states of token ℓ therefore differ across g only through what each sample's image contributed to that word, while position ℓ still denotes the same word in every sample. The native attention of the backbone has thus written unaligned image content into aligned slots, with no new parameters. Donor image tokens lack this alignment and cost N^2 instead of NL attention; reading them recovers less than half of the gain (Table 3).

3.2 STEP 2: EACH IMAGE QUERY READS THE PROMPT TOKENS OF THE WHOLE GROUP

At block $b{+}1$, the text queries and the image bank stay local to sample g; only the text keys and values that an image query sees are replaced by those of all G samples. For image query q_n of target g, with image logits $\alpha_{nm} = \langle q_n, k_{gm}^I\rangle/\sqrt{d_h}$ and text logits $\beta_{nj\ell} = \langle q_n, k_{j\ell}^T\rangle/\sqrt{d_h}$ for sample j and token ℓ, the group attention output is

$$o_n^{\mathrm{grp}} = \frac{\sum_{m=1}^{N} e^{\alpha_{nm}} v_{gm}^I + \frac{1}{G}\sum_{j=1}^{G}\sum_{\ell=1}^{L} e^{\beta_{nj\ell}} v_{j\ell}^T}{\sum_{m=1}^{N} e^{\alpha_{nm}} + \frac{1}{G}\sum_{j=1}^{G}\sum_{\ell=1}^{L} e^{\beta_{nj\ell}}}, \tag{2}$$

which is a softmax over the concatenated keys $[K_g^I; K_1^T; \ldots; K_G^T]$ with $-\log G$ added to every text logit. At $G = 1$, Equation (2) is the original attention. When every donor duplicates the target, the G text sums are equal and the $1/G$ factor returns the original output, so the teacher departs from the base model only through content differences between samples. The weights over the GL text keys are recomputed for every query, so for the prompt of Figure 3, "three yellow dogs stand behind two red sleds," the queries that render the sleds and those that render the dogs can draw on different samples.

Why each query chooses its own sample. Whole-image methods weight samples with one vector per prompt, as in a reward-weighted average or a best-of-G choice. Suppose K image queries each need a different sample, so the ideal query-by-sample routing matrix is I_K. Any mixer built from at most M global weight vectors has rank at most M, and the Eckart–Young theorem gives $\inf_{\mathrm{rank}(\Pi)\le M} K^{-1}\|I_K - \Pi\|_F^2 = 1 - M/K$; whole-image weighting has $M = 1$. The per-query softmax in Equation (2) realizes a different row for each query, with error at most $K(K-1)e^{-2\gamma}$ when the matching logits exceed the others by γ (Appendix A). Query-wise reading must also place enough attention on the relevant evidence:

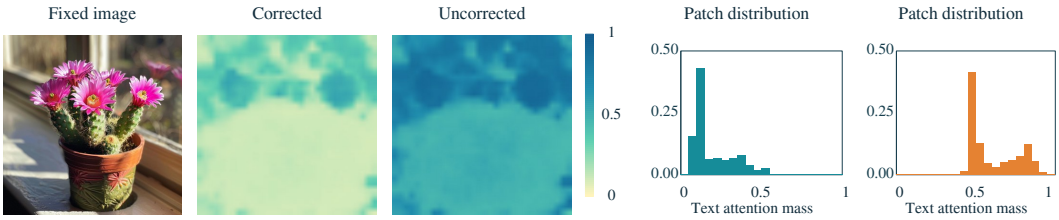


Figure 4: Duplicated text banks at $G = 8$ for a fixed image, query set, and image bank. With the $-\log G$ shift (corrected), each patch keeps the text attention mass of the original model; without it (uncorrected), the bank counted eight times takes most of the attention. Maps share the $[0, 1]$ scale; histograms share bins.

Theorem 1 (Coverage by query-wise reading)**.** *Fix the target input $X = (x_t^g, t, c)$, relabel the banks so that the target's own bank is $j = 1$, and consider an image query that renders requirement k, with text logits $\beta_{kj\ell}$ and relevant prompt positions R_k. For bank j, let $A_{kj} = \sum_{\ell \in R_k} e^{\beta_{kj\ell}}$, $B_{kj} = \sum_{\ell \notin R_k} e^{\beta_{kj\ell}}$, let z_k^I be the image-bank mass, and let $\rho_{k,G} = \sum_j A_{kj}/(Gz_k^I + \sum_j (A_{kj} + B_{kj}))$ be the attention fraction on relevant evidence. For $\tau_k \in (0, 1)$, define $W_{kj} = (1 - \tau_k)A_{kj} - \tau_k B_{kj}$ and $h_{k,G} = (G\tau_k z_k^I - W_{k1})/(G-1)$. Then $\rho_{k,G} \geq \tau_k$ if and only if $(G-1)^{-1} \sum_{j=2}^{G} W_{kj} \geq h_{k,G}$. If the donors are conditionally independent given X, $W_{kj} \in [w_k^-, w_k^+]$, and $\mu_{W,k} = \mathbb{E}[W_{kj} \mid X] > h_{k,G}$, then*

$$\Pr(\rho_{k,G} < \tau_k \mid X) \leq \exp\left[-\frac{2(G-1)(\mu_{W,k} - h_{k,G})^2}{(w_k^+ - w_k^-)^2}\right]. \tag{3}$$

Since $h_{k,G} \to \tau_k z_k^I$, a larger group helps exactly when donors carry more relevant mass than the threshold on average. With a fixed evidence margin, a union bound shows that $G = 1 + O(\log(K/\delta))$ samples cover all K requirements with probability $1 - \delta$ (Appendix A), against the exponential G_{whole} of Equation (1).

Why the text logits are shifted by $\log G$. Adding G text banks multiplies the text mass by about G, so the target image context loses weight as the group grows (Figure 4). For one query, write u_k^I, z_k^I for the numerator and mass of the image bank and r_{kj}, y_{kj} for those of text bank j, and let $o_{k,G} = (u_k^I + c_G \sum_j r_{kj})/(z_k^I + c_G \sum_j y_{kj})$ with a text scale c_G.

Theorem 2 (Normalized group attention)**.** *Requiring the original attention at $G = 1$ and invariance to copying all text banks gives $c_G = 1/G$ uniquely. Given X, fix bank 1 and let banks $2{:}G$ be conditionally independent with means $\mu_{r,k}, \mu_{y,k}$. With $o_{k,\infty} = (u_k^I + \mu_{r,k})/(z_k^I + \mu_{y,k})$, $\xi_{kj} = (r_{kj} - \mu_{r,k}) - o_{k,\infty}(y_{kj} - \mu_{y,k})$, and $\sigma_{\xi,k}^2 = \mathbb{E}[\|\xi_{kj}\|_2^2 \mid X] < \infty$ for $j \geq 2$,*

$$\mathbb{E}\big[\|o_{k,G} - o_{k,\infty}\|_2^2 \mid X\big] \leq \frac{\|\xi_{k1}\|_2^2 + (G-1)\sigma_{\xi,k}^2}{G^2 (z_k^I)^2}. \tag{4}$$

A Λ_X-Lipschitz frozen suffix passes this bound to the teacher output with factor Λ_X^2.

The $-\log G$ shift is thus the only additive correction that keeps the attention scale of the base model, and the teacher converges at rate $O(1/G)$ to a limit set by the prompt and the target state. We compute Equation (2) with two fused attention calls, one over the local image bank and one over the concatenated text banks, merged by their log-sum-exp normalizers (Algorithm 1). The added work per head is $O(G(G-1)NLd_h)$, against $O(G(G-1)N^2 d_h)$ for attention across image tokens, since $L \ll N$.

3.3 Step 3: Distill the Group Prediction into the Single-Sample Model

After block $b{+}1$, each target continues through the frozen suffix and yields $T_{\bar{\theta}}(x_t^g, t, c, \mathcal{X}_t)$, a velocity on the same grid as the student's output. The student evaluates the same state with its ordinary path and minimizes

$$\mathcal{L}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{g=1}^{G}\Big\| f_\theta(x_t^g, t, c) - \mathrm{sg}\big[T_{\bar{\theta}}(x_t^g, t, c, \mathcal{X}_t)\big]\Big\|_2^2\right], \tag{5}$$

with the expectation over prompts, seeds, and time. A dense target corrects each position separately, so a sample that already renders the dogs correctly is corrected only where it renders the sleds. The snapshot and group are refreshed after every update, so the group reflects the current errors of the model.

What the student can learn. The student sees only X, whereas the teacher also depends on the sampled donors D_G. Write the teacher correction as $\Delta_G = T_{\bar{\theta}}(X, D_G) - f_{\bar{\theta}}(X)$ and split it into its predictable part $m_G(X) = \mathbb{E}[\Delta_G \mid X]$ and the donor-specific remainder $\epsilon_G = \Delta_G - m_G(X)$. For every single-input correction $\psi(X)$,

$$\mathbb{E}\|\Delta_G - \psi(X)\|_2^2 = \mathbb{E}\|\epsilon_G\|_2^2 + \mathbb{E}\|m_G(X) - \psi(X)\|_2^2. \tag{6}$$

The first term is the same for every student, which can remove only the second. With parameter Jacobian J_X, the best first-order update removes exactly $\|P_{\mathcal{T}} m_G\|_{L^2}^2$, where $P_{\mathcal{T}}$ projects onto the closed span $\mathcal{T}$ of $X \mapsto J_X \omega$ (Appendix A). The teacher–student gap thus has two sources, donor-specific corrections and adapter capacity. Figure 5 shows the first: on a held-out target, the residual of a student fitted on other targets stays around the boots, where the correction depends on the drawn donors, and does not shrink from step 2 to step 80 of fitting.

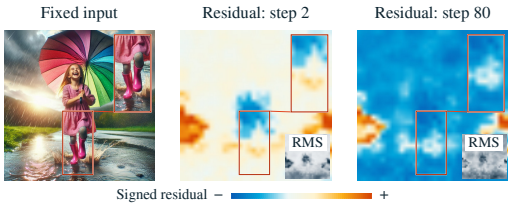


Figure 5: Residual of one shared student on a held-out target after 2 and 80 fitting steps; boxes mark the boots region.

4 Experiments

We test whether the deployed model improves and whether the gain follows the path of Section 3.

4.1 Setup

We fine-tune the public SD3.5 Medium, FLUX.1-dev, and Qwen-Image-2512 checkpoints at 1024^2 with $G = 4$ and rank-64 LoRA for 300 updates of eight prompt groups each, using Pick-a-Pic prompts deduplicated against every test set (Kirstain et al., 2023). Held-out calibration prompts select the write/read blocks 11/12, 9/10, and 29/30. We evaluate on GenEval (Ghosh et al., 2023), GenEval2 (Kamath et al., 2025), T2I-CompBench++ (Huang et al., 2025), DPG-Bench (Hu et al., 2024), GenAI-Bench with VQAScore (Lin et al., 2024), and TIFA (Hu et al., 2023), with official prompts, evaluators, and aggregation (TIFA uses the 4,073 of 4,081 prompts with official question–answer pairs). Three training runs are the repeated units for paired-run t intervals; prompt-clustered sign-flip tests use Holm correction per backbone (Appendix B.3). Flow-GRPO (Liu et al., 2025), MixGRPO (Li et al., 2026a), DGPO (Luo et al., 2026), DiffusionNFT (Zheng et al., 2026), and TreeGRPO (Ding & Ye, 2026) run on the backbones their official code supports and are rescored with the same evaluators.

4.2 Main Results

VoteX improves all 18 headline scores in Table 1, from 2.22 points (Qwen-Image TIFA) to 8.93 points (FLUX GenEval). On GenEval2, the gain is 6.97 points for SD3.5 (run-level 95% interval $[6.24, 7.70]$), 7.10 for FLUX ($[6.49, 7.71]$), and 4.34 for Qwen-Image ($[3.75, 4.93]$); every run improves, and Holm-adjusted p values are at most 0.00025 for scalar outcomes and 0.0004 for T2I categories. Without a reward model, VoteX exceeds the strongest of the five reward-based methods in every column by 0.2 to 2.0 points, with the largest margins on GenEval2 and DPG, the two benchmarks built from dense multi-requirement prompts.

The gain tracks the complementarity of Section 2: on GenEval2, SD3.5 and FLUX have $D_{\text{comp}} = 15.40$ and 14.60 before training and gain about 7 points, while Qwen-Image, whose samples already satisfy more atoms, has $D_{\text{comp}} = 6.69$ and gains 4.34. Gains are larger where the samples of a model leave more correct evidence unused.

The category results in Table 2 show which requirements move. Numeracy rises by 4.7, 5.1, and 3.9 points and spatial relations by 3.3, 3.8, and 2.6 points, while color, shape, and texture change

Table 1: Matched whole-system comparison across six compositional benchmarks. Means are on a 0–100 scale (↑). External methods use the same backbone and evaluation protocol; training objectives differ. A dash denotes a result not reported in this comparison. Bold marks the best value within each backbone; blue identifies our method.

Method	GenEval	GenEval2	T2I-Comp.++	DPG	GenAI	TIFA
	Overall	GM	Complex	Overall	VQAScore	Overall
SD3.5 Medium (Stability AI, 2024)						
Base	64.9	51.3	87.2	84.0	62.6	80.3
MixGRPO (Li et al., 2026a)	–	56.5	89.7	85.8	65.7	82.1
Flow-GRPO (Liu et al., 2025)	–	56.9	90.0	86.1	66.0	82.3
DGPO (Luo et al., 2026)	–	57.5	90.2	86.4	66.3	82.5
DiffusionNFT (Zheng et al., 2026)	–	57.2	90.1	86.2	66.5	82.6
TreeGRPO (Ding & Ye, 2026)	–	53.5	88.4	84.5	67.0	81.5
VoteX	**70.6**	**58.3**	**90.9**	**88.0**	**67.2**	**83.1**
FLUX.1-dev (Black Forest Labs, 2024)						
Base	65.8	53.6	87.3	83.8	64.8	82.1
MixGRPO (Li et al., 2026a)	–	58.2	90.2	85.8	68.8	83.8
Flow-GRPO (Liu et al., 2025)	–	58.7	90.5	86.1	69.1	84.0
VoteX	**74.8**	**60.7**	**90.8**	**87.1**	**69.9**	**84.7**
Qwen-Image-2512 (Qwen Team, 2025)						
Base	85.8	72.8	92.5	89.0	74.2	87.3
Flow-GRPO (Liu et al., 2025)	–	75.4	94.2	90.5	76.9	88.6
VoteX	**90.0**	**77.1**	**94.8**	**91.7**	**77.7**	**89.5**

Table 2: Category-level evaluation of compositional generation. Seven additional T2I-CompBench++ categories and all five DPG categories are shown; the T2I Complex headline remains in Table 1. Means are multiplied by 100 (↑), with the best value bold within each backbone. Non-sp. denotes non-spatial relationships; 2D denotes complex 2D relationships; Global denotes global detail.

Method	T2I-CompBench++							DPG-Bench				
	Color	Shape	Texture	Spatial	Numeracy	Non-sp.	2D	Global	Entity	Attrib.	Relation	Count
SD3.5 Medium (Stability AI, 2024)												
Base	81.3	58.9	73.3	32.0	40.8	61.7	90.9	79.7	87.5	86.1	84.5	82.6
VoteX	**81.6**	**59.6**	**74.2**	**35.3**	**45.5**	**62.9**	**94.0**	**85.3**	**88.4**	**87.4**	**86.6**	**86.2**
FLUX.1-dev (Black Forest Labs, 2024)												
Base	74.1	57.2	69.2	28.6	38.7	61.9	92.1	79.3	87.2	85.8	84.2	82.7
VoteX	**74.5**	**58.0**	**70.3**	**32.4**	**43.8**	**63.2**	**95.5**	**84.1**	**88.0**	**87.0**	**86.2**	**85.9**
Qwen-Image-2512 (Qwen Team, 2025)												
Base	85.2	67.1	77.5	42.1	51.8	70.2	94.8	92.7	91.3	91.6	92.0	94.3
VoteX	**85.5**	**67.7**	**77.7**	**44.7**	**55.7**	**71.2**	**97.0**	**95.2**	**91.8**	**92.3**	**93.1**	**96.1**

by at most 1.1 points. We attribute this ordering to how each requirement is decided: a count or a relation holds for the whole scene (a sample draws three dogs or four), so other samples often supply what one image misses, while a single object's attribute is already correct in most samples. On DPG, Global detail and Count gain most (5.6 and 3.6 points on SD3.5) and Entity least.

4.3 Where the Gain Comes From

Same-prompt samples with different content are required. The matched source controls in Table 4 (Appendix B.4) keep group size, attention work, updates, and optimizer fixed and change only the states in the donor slots. Full exchange reaches 58.4, 60.9, and 77.1 GenEval2 points. Copies of the target reach 51.3, 53.6, and 72.8, and samples of other prompts reach 51.5, 53.8, and 72.8, close to the base scores of 51.3, 53.6, and 72.8. Copies add no content, and other prompts add content under the wrong words; the gain requires states that share the prompt and differ in content.

Each query uses the sample that rendered its requirement. A frozen-teacher intervention on SD3.5 and FLUX (Table 5) tests the direction of the read. For every atom the target misses, an independent selection bank identifies a donor that answers it correctly, and we mask that donor's text bank at every step. Target-atom accuracy falls from 75.84% to 55.92% on SD3.5 and from 76.50% to 56.60% on FLUX. Masking an unrelated donor matched on other atoms lowers it only to 70.93% and 71.72%, and other atoms move by at most 1.55 points. Replacing it with a bank from another prompt lowers accuracy further, to 44.74% and 45.17%, since the target then reads wrong evidence under the



Figure 6: Scaling one donor's contribution (weight 0 to 1) at the prompt positions of one object in a frozen teacher. *Top:* bottom or top book; only that book takes the donor's color. *Bottom:* floral or chevron mug; only that mug takes the donor's pattern. All weights share the target state and noise.

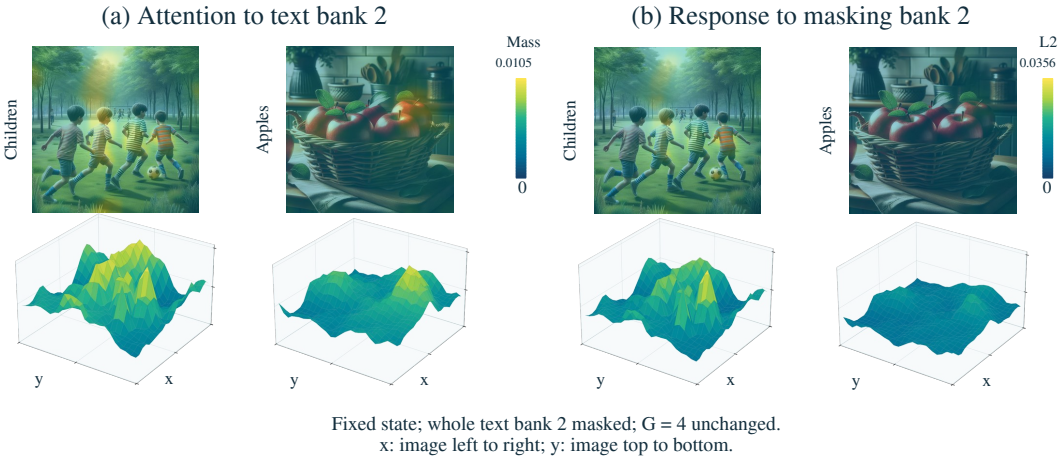


Figure 7: Masking the text bank of donor 2 ($G = 4$, fixed states). **(a)** Attention mass of each image patch on that bank before masking. **(b)** L2 change of the attention output after masking. Surfaces plot the maps above them.

right words. Figure 6 shows this selectivity in images: scaling a donor at the positions of one book or mug changes only that object.

Figure 7 explains why masking acts locally. The patches that attended to donor 2 are the patches whose output changes when it is removed, concentrated on the children and the apples. Removing a donor changes only the regions that read it, the per-query routing that whole-image weights cannot express.

4.4 Design Analysis

Table 3 varies one component at a time on SD3.5 under fixed training and evaluation (quantile analyses in Appendix C.2). Removing the group read gives 51.6, close to the base score of 51.3, so the read is the source of the gain. Blocking the image-to-text attention of the write block (no visual write) gives 53.7: the prompt tokens read at block $b{+}1$ must first absorb the current image of their sample. Reading donor image keys and values gives 54.3 at N/L times the attention cost, because patches of different layouts do not correspond. Removing the $-\log G$ shift costs 2.9 points, as predicted by Theorem 2 and Figure 4. A final-token target reaches 56.1, since it discards the per-position corrections.

Table 3: SD3.5 Medium design studies. Means $\times 100$ (↑); GE2: GenEval2 GM, T2I: Complex, DPG: overall. Bold marks the best in each section.

Condition	GE2	T2I	DPG
$G = 4$: component ablations			
Full VoteX	**58.3**	**90.9**	**88.0**
No visual write	53.7	88.5	85.4
No group read	51.6	87.4	84.1
Image key/value read	54.3	88.8	85.6
No $-\log G$ correction	55.3	88.9	85.9
Final-token loss	56.1	89.7	86.1
Fixed updates: group size			
$G = 1$	51.4	87.2	84.0
$G = 2$	56.0	89.7	86.7
$G = 4$	**58.3**	**90.9**	**88.0**
$G = 8$	57.9	90.7	87.7

At $G = 1$, Equation (2) reduces to the original model, the target equals the student's own prediction, and the score stays at 51.4. It rises to 56.0 at $G = 2$ and 58.3 at $G = 4$ as more atoms become available (Theorem 1) and settles at 57.9 at $G = 8$. With a fixed number of trajectories instead of fixed updates, $G = 2, 4, 8$ give 55.40, 56.77, and 54.31: larger groups leave fewer independent prompts per update and add quadratic group work.

4.5 Transfer to the Single-Sample Model and Deployment

The student recovers most of the teacher gain (Table 6). On GenEval2, base, teacher, and student score 51.34, 59.28, and 58.31 for SD3.5, 53.64, 61.73, and 60.74 for FLUX, and 72.75, 77.63, and 77.09 for Qwen-Image; across GenEval2, DPG, and GenAI, students retain 87.1% to 97.1% of the teacher gain. The remaining gap corresponds to the two terms of Equation (6), donor-specific corrections and adapter capacity. In a blind study with 600 prompts, three judgments per pair, and 54 raters, a rater-random-intercept model prefers VoteX for full-prompt fidelity with probabilities 0.652 $[0.631, 0.673]$, 0.656 $[0.639, 0.672]$, and 0.633 $[0.611, 0.655]$, and for fewer defects with 0.589 to 0.608 (Krippendorff's α 0.282–0.368 and 0.048–0.059). HPS v2 (Wu et al., 2023), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023) stay at or above base, and within-prompt LPIPS (Zhang et al., 2018) falls by 0.014–0.015, a small diversity reduction. The exported checkpoint holds only student parameters; over 1,000 alternating batch-one A100 runs, p90 latency changes by -0.05%, 0.06%, and 0.09%, and peak memory by less than 1 MB.

5 Related Work

Improving compositional generation after pretraining. Text-to-image post-training has moved from terminal rewards to relative and reward-free signals. DDPO and DPOK optimize a reward through the denoising policy (Black et al., 2024; Fan et al., 2023), DRaFT differentiates the reward through sampling (Clark et al., 2024), and Diffusion-DPO learns from pairwise preferences (Wallace et al., 2024). Group-relative methods such as Flow-GRPO, DGPO, MixGRPO, TreeGRPO, and DiffusionNFT compare on-policy samples of one prompt and turn their scores into advantages (Liu et al., 2025; Luo et al., 2026; Li et al., 2026a; Ding & Ye, 2026; Zheng et al., 2026; Chen et al., 2025). CoMat and SILMM replace the reward model with image-to-text matching or VQA (Jiang et al., 2024; Qu et al., 2025), and distillation, consistency training, and self-correction learn from model trajectories or clean targets (Agarwal et al., 2024; Fang et al., 2026; Jiang et al., 2026; Qin et al., 2026; Feng et al., 2024; Lu & Song, 2025). These methods optimize rewards, targets, or trajectories without matching corresponding prompt tokens across seeds to transfer one sample's correct requirement to another. VoteX builds the supervision inside the network, per image query, from the other samples of the same prompt.

Computation across samples and cross-modal attention. Group Diffusion shares attention across images during joint denoising (Mo et al., 2026), Scaling Group Inference recombines test-time candidates (Parmar et al., 2025), and Diffusion Mental Averages aligns trajectories before averaging them (Thawatdamrongkit et al., 2026); they keep the group at inference and combine image-level or aligned features. Within one sample, MMDiT introduced bidirectional text–image attention (Esser et al., 2024), analyses trace its cross-modal paths and the object information held by prompt positions (Shin et al., 2025; Li et al., 2026b; Helbling et al., 2025), and TACA and REPA modify modal attention and visual features (Lv et al., 2025; Yu et al., 2025) of latent diffusion transformers (Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2023). Cross-sample computation needs a correspondence between samples that differ in layout; VoteX takes it from the prompt positions that these analyses show to hold object information, and uses the group only in training.

6 Conclusion

Samples of one compositional prompt get different requirements right. VoteX exchanges these partial successes through the prompt tokens of a joint-attention transformer, the positions that correspond across samples, and distills the group prediction into the ordinary model, which generates from one seed at unchanged cost. The gains track the complementarity of the base samples, shrink by at least 90% without new same-prompt content, and depend on the specific sample that rendered each missed requirement correctly.

AI USE STATEMENT

Generative AI tools assisted with related-work retrieval, research ideation, drafting sections of the paper, and polishing prose and LaTeX. They suggested candidate papers and analyses. The authors are responsible for checking cited sources, the method, reported results, and all scientific claims.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Black Forest Labs. FLUX.1-dev model card. Hugging Face model card, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Accessed 2026-09-18.

Renjie Chen, Wenfeng Lin, Yichen Zhang, Jiangchuan Wei, Boyuan Liu, Chao Feng, Jiao Ran, and Mingyu Guo. Towards self-improvement of diffusion models via group preference optimization. *arXiv preprint arXiv:2505.11070*, 2025. URL https://arxiv.org/abs/2505.11070.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17400.

Zheng Ding and Weirui Ye. TreeGRPO: Tree-advantage GRPO for online RL post-training of diffusion models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.08153.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024. URL https://arxiv.org/abs/2403.03206.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 79858–79885, 2023. URL https://arxiv.org/abs/2305.16381.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, and Feng Zhao. Flow-OPD: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*, 2026. URL https://arxiv.org/abs/2605.08063.

Weilun Feng, Chuanguang Yang, Zhulin An, Libo Huang, Boyu Diao, Fei Wang, and Yongjun Xu. Relational diffusion distillation for efficient image generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 205–213, 2024. URL https://arxiv.org/abs/2410.07679.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52132–52152, 2023. URL https://arxiv.org/abs/2310.11513.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22946–22963, 2025. URL https://arxiv.org/abs/2502.04320.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. URL https://arxiv.org/abs/2403.05135.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20349–20360, 2023. URL https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025. doi: 10.1109/TPAMI.2025.3531907. URL https://doi.org/10.1109/TPAMI.2025.3531907.

Dengyang Jiang, Xin Jin, Dongyang Liu, Zanyi Wang, Mingzhe Zheng, Ruoyi Du, Xiangpeng Yang, Qilong Wu, Zhen Li, Peng Gao, Harry Yang, and Steven Hoi. D-OPSD: On-policy self-distillation for continuously tuning step-distilled diffusion models. *arXiv preprint arXiv:2605.05204*, 2026. URL https://arxiv.org/abs/2605.05204.

Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. CoMat: Aligning text-to-image diffusion model with image-to-text concept matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 76177–76209, 2024. URL https://arxiv.org/abs/2404.03653.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023. URL https://arxiv.org/abs/2305.01569.

Junzhe Li, Yutao Cui, Tao Huang, Chuxuan Zeng, Weijie Kong, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, and Liefeng Bo. MixGRPO: Unlocking flow-based GRPO efficiency with mixed ODE–SDE. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2507.21802.

Maohua Li, Qirui Li, Yanke Zhou, Yiduo Li, Zhaosheng Chi, Chao Xu, Cuifeng Shen, Yixuan Xu, Hanlin Tang, Kan Liu, Tao Lan, Lin Qu, and Shao-Qun Zhang. Text template tokens are implicit semantic registers in diffusion transformers. *arXiv preprint arXiv:2607.19139*, 2026b. URL https://arxiv.org/abs/2607.19139.

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. URL https://arxiv.org/abs/2404.01291.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.05470.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11081.

Yihong Luo, Tianyang Hu, and Jing Tang. Reinforcing diffusion models by direct group preference optimization. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.08425.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5934–5943, 2025. URL https://arxiv.org/abs/2506.07986.

Sicheng Mo, Thao Nguyen, Richard Zhang, Nicholas Kolkin, Siddharth Srinivasan Iyer, Eli Shechtman, Krishna Kumar Singh, Yong Jae Lee, Bolei Zhou, and Yuheng Li. Group diffusion: Enhancing image generation by unlocking cross-sample collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35702–35712, 2026. URL https://arxiv.org/abs/2512.10954.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. *arXiv preprint arXiv:2508.15773*, 2025. URL https://arxiv.org/abs/2508.15773.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://arxiv.org/abs/2212.09748.

You Qin, Linqing Wang, Hao Fei, Roger Zimmermann, Liefeng Bo, Qinglin Lu, and Chunyu Wang. SOAR: Self-correction for optimal alignment and refinement in diffusion models. *arXiv preprint arXiv:2604.12617*, 2026. URL https://arxiv.org/abs/2604.12617.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. SILMM: Self-improving large multimodal models for compositional text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18497–18508, 2025. URL https://arxiv.org/abs/2412.05818.

Qwen Team. Qwen-Image-2512 model card. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen-Image-2512. Accessed 2026-09-18.

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*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19492–19502, 2025. URL https://arxiv.org/abs/2508.07519.

Stability AI. Stable Diffusion 3.5 Medium model card. Hugging Face model card, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium. Accessed 2026-09-18.

Phonphrm Thawatdamrongkit, Sukit Seripanitkarn, and Supasorn Suwajanakorn. Diffusion mental averages. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35713–35725, 2026. URL https://arxiv.org/abs/2603.29239.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024. URL https://arxiv.org/abs/2311.12908.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15903–15935, 2023. URL https://arxiv.org/abs/2304.05977.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06940.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.16117.

A PROOFS

This appendix proves Theorem 1, the routing statement of Section 3.2, Theorem 2, and the transfer decomposition of Section 3.3. The notation follows Section 3. We fix a prompt, a time, the target input $X = (x_t^g, t, c)$, and one image query of the target, which renders requirement k. Text banks are indexed by $j \in \{1, \dots, G\}$ with $j = 1$ the target's own bank and $j \geq 2$ the donors. Probabilities and moments are taken over the donor seeds conditionally on X, so bank 1 is fixed. All denominators are positive because they are sums of exponentiated logits.

A.1 COVERAGE AND ROUTING

Exact coverage condition. Write $\tau = \tau_k$. Multiplying $\rho_{k,G} = \sum_j A_{kj} / (G z_k^I + \sum_j (A_{kj} + B_{kj}))$ by its denominator gives

$$\rho_{k,G} \geq \tau \iff \sum_{j=1}^{G} A_{kj} \geq \tau \Big[G z_k^I + \sum_{j=1}^{G} (A_{kj} + B_{kj}) \Big] \iff W_{k1} + \sum_{j=2}^{G} W_{kj} \geq G \tau z_k^I. \tag{7}$$

Dividing by $G - 1$ yields the equivalence in Theorem 1 with $h_{k,G} = (G\tau z_k^I - W_{k1})/(G-1)$.

Concentration. Given X, the variables $W_{k2}, \dots, W_{kG}$ are independent, have common mean $\mu_{W,k}$, and lie in $[w_k^-, w_k^+]$ with width $\mathcal{R}_k = w_k^+ - w_k^- > 0$ (for $\mathcal{R}_k = 0$ the donor average is deterministic). If $\mu_{W,k} > h_{k,G}$, Hoeffding's inequality gives

$$\begin{aligned} \Pr(\rho_{k,G} < \tau_k \mid X) &= \Pr\Big(\frac{1}{G-1} \sum_{j=2}^{G} W_{kj} - \mu_{W,k} < -(\mu_{W,k} - h_{k,G}) \,\Big|\, X \Big) \\ &\leq \exp\Big[- \frac{2(G-1)(\mu_{W,k} - h_{k,G})^2}{\mathcal{R}_k^2} \Big], \end{aligned} \tag{8}$$

which is Equation (3). If $\mu_{W,k} < h_{k,G}$, the upper-tail form bounds $\Pr(\rho_{k,G} \geq \tau_k \mid X)$ by the same expression with $h_{k,G} - \mu_{W,k}$ in place of $\mu_{W,k} - h_{k,G}$. Since $h_{k,G} = \frac{G}{G-1}\tau_k z_k^I - \frac{W_{k1}}{G-1} \to \tau_k z_k^I$, the conditional strong law makes coverage converge to probability one when $\mu_{W,k} > \tau_k z_k^I$ and to zero when $\mu_{W,k} < \tau_k z_k^I$.

Binary evidence model and union bound. Let $e_{kj} \sim \text{Bernoulli}(p_k)$ for $j \geq 2$ indicate that donor j carries the evidence for requirement k, with $A_{kj} = a_{0,k} + (a_{1,k} - a_{0,k})e_{kj}$, $a_{1,k} > a_{0,k} \geq 0$, and $B_{kj} = b_k$. The count $n_{k,G} = \sum_{j \geq 2} e_{kj}$ is $\text{Binomial}(G-1, p_k)$, and Equation (7) becomes $n_{k,G}/(G-1) \geq f^*_{k,G}$ with

$$f^*_{k,G} = \frac{G\tau_k z^I_k - W_{k1} + (G-1)\big[\tau_k b_k - (1-\tau_k)a_{0,k}\big]}{(G-1)(1-\tau_k)(a_{1,k} - a_{0,k})}. \tag{9}$$

For $0 < f^*_{k,G} < p_k < 1$, the Chernoff bound gives $\Pr(\rho_{k,G} < \tau_k \mid X) \leq \exp[-(G-1)D_{\text{KL}}(f^*_{k,G} \| p_k)]$; $f^*_{k,G} \leq 0$ makes coverage certain and $f^*_{k,G} > 1$ makes it impossible in this model. Assume every requirement with $f^*_{k,G} > 0$ satisfies $0 < f^*_{k,G} < p_k < 1$ and $D_{\text{KL}}(f^*_{k,G} \| p_k) \geq d_0 > 0$. A union bound over $k \in [K]$, which needs no independence between requirements within a sample, gives

$$\Pr\left(\rho_{k,G} \geq \tau_k \text{ for all } k \mid X\right) \geq 1 - Ke^{-(G-1)d_0}, \tag{10}$$

so $G \geq 1 + d_0^{-1}\log(K/\delta)$ suffices for failure probability δ. The logarithmic rate rests on the margin d_0 of the evidence, which the binary model makes explicit.

For whole-image selection, let q_K be the probability that one sample satisfies all K requirements. Independent samples show at least one complete success with probability $1 - (1-q_K)^G$. With independent requirements of probability p, $q_K = p^K$, and solving $(1-p^K)^G \leq \delta$ gives Equation (1); since $-\log(1-p^K) \sim p^K$ as $p^K \to 0$, $G_{\text{whole}} = \Theta(p^{-K}\log(1/\delta))$.

Low-rank obstruction. Let $0 \leq M \leq K$. All K singular values of I_K equal one, so the Eckart–Young theorem gives

$$\inf_{\text{rank}(\Pi) \leq M} \|I_K - \Pi\|_F^2 = \sum_{i=M+1}^{K} 1 = K - M, \tag{11}$$

and division by K gives the bound of Section 3.2. The infimum ranges over all real matrices and therefore also bounds nonnegative row-normalized mixers. One whole-image weight vector repeated at every query yields identical rows and rank one; M global weight vectors span rank at most M.

Query-wise softmax. Aggregate each bank by log-sum-exp and suppose query k has matching score s_k while every other bank scores at most $s_k - \gamma$. The row-wise softmax matrix Π_γ satisfies $(\Pi_\gamma)_{kk} \geq 1/(1 + (K-1)e^{-\gamma})$ and $(\Pi_\gamma)_{kj} \leq e^{-\gamma}$ for $j \neq k$. Each row therefore has squared error at most $[(K-1)^2 + (K-1)]e^{-2\gamma}$, and

$$K^{-1}\|I_K - \Pi_\gamma\|_F^2 \leq K(K-1)e^{-2\gamma}. \tag{12}$$

Equation (2) computes a separate softmax row for every image query, so it realizes such a matrix; Theorem 1 controls how much of the query's mass reaches the relevant text positions.

A.2 Normalization and teacher stability

Uniqueness of $c_G = 1/G$. For one query, let $S_r = \sum_{j=1}^G r_{kj}$ and $S_y = \sum_{j=1}^G y_{kj}$. Copying all banks κ times changes the sums to $\kappa S_r, \kappa S_y$, and invariance requires

$$\frac{u^I_k + c_G S_r}{z^I_k + c_G S_y} = \frac{u^I_k + \kappa c_{\kappa G} S_r}{z^I_k + \kappa c_{\kappa G} S_y} \iff (c_G - \kappa c_{\kappa G})\,(z^I_k S_r - u^I_k S_y) = 0. \tag{13}$$

Banks with $z^I_k S_r \neq u^I_k S_y$ exist, so invariance for all banks forces $c_G = \kappa c_{\kappa G}$. With $G = 1$ and $c_1 = 1$ (the original attention), $c_n = 1/n$ for every positive integer n; conversely, $c_G = 1/G$ satisfies $\kappa c_{\kappa G} = c_G$. The additive logit form is $\log c_G = -\log G$.

Rate of the group attention output. With $c_G = 1/G$, let $\bar{r}_{k,G} = G^{-1}\sum_j r_{kj}$ and $\bar{y}_{k,G} = G^{-1}\sum_j y_{kj}$. Subtracting $o_{k,\infty}$ gives the identity

$$o_{k,G} - o_{k,\infty} = \frac{(\bar{r}_{k,G} - \mu_{r,k}) - o_{k,\infty}(\bar{y}_{k,G} - \mu_{y,k})}{z^I_k + \bar{y}_{k,G}} = \frac{G^{-1}\sum_{j=1}^G \xi_{kj}}{z^I_k + \bar{y}_{k,G}}. \tag{14}$$

Given X, ξ_{k1} is fixed and $\xi_{k2}, \dots, \xi_{kG}$ are independent with mean zero, so the cross terms vanish and $\mathbb{E}[\|\sum_j \xi_{kj}\|_2^2 \mid X] = \|\xi_{k1}\|_2^2 + (G-1)\sigma_{\xi,k}^2$. Replacing the denominator by the smaller z_k^I proves Equation (4). The rate is tight: with $y_{kj} \equiv \mu_{y,k}$, $r_{k1} = \mu_{r,k}$, and $\mathbb{E}[\|r_{kj} - \mu_{r,k}\|_2^2 \mid X] > 0$ for $j \geq 2$, the denominator is constant and the error equals $\frac{G-1}{G^2}\,\mathbb{E}[\|r_{kj} - \mu_{r,k}\|_2^2 \mid X]/(z_k^I + \mu_{y,k})^2 = \Theta(G^{-1})$.

Propagation through the frozen suffix. Write $T_G = T_{\bar\theta}(X, D_G)$ for the teacher with G banks. Stack the group attention outputs $o_{k,G}$ of all queries into O_G and their limits into O_∞. The other suffix inputs are fixed by X and absorbed into the suffix map F_X, so $T_G = F_X(O_G)$, and we set $T_\infty = F_X(O_\infty)$. Let

$$\mathcal{B}_G(X) = \sum_k \frac{\|\xi_{k1}\|_2^2 + (G-1)\sigma_{\xi,k}^2}{G^2 (z_k^I)^2}. \tag{15}$$

If F_X is Λ_X-Lipschitz on a set containing O_∞ and every realization of O_G, then $\mathbb{E}[\|T_G - T_\infty\|_2^2 \mid X] \leq \Lambda_X^2\,\mathbb{E}[\|O_G - O_\infty\|_F^2 \mid X] \leq \Lambda_X^2 \mathcal{B}_G(X)$. The squared Frobenius norm is the sum over queries, so no independence across queries is used. A global $O(1/G)$ rate follows under an integrable envelope such as $\mathbb{E}[\Lambda_X^2 \sum_k \sigma_{\xi,k}^2/(z_k^I)^2] < \infty$ together with the analogous target-bank term.

A.3 What a single-sample student can learn

Let $\Delta_G = T_G - f_{\bar\theta}(X)$ and $\Delta_\infty = T_\infty - f_{\bar\theta}(X)$, both square integrable; Δ_∞ is a function of X. Recall $m_G(X) = \mathbb{E}[\Delta_G \mid X]$ and $\epsilon_G = \Delta_G - m_G(X)$.

Orthogonality. For every square-integrable single-input correction $\psi(X)$, $\mathbb{E}\langle \epsilon_G, m_G - \psi\rangle = \mathbb{E}\langle \mathbb{E}[\epsilon_G \mid X], m_G - \psi\rangle = 0$, and expanding the square proves Equation (6). Conditional expectation is the L^2 projection onto functions of X; taking $h = \Delta_\infty$ and applying conditional Jensen gives

$$\mathbb{E}\|\epsilon_G\|_2^2 \leq \mathbb{E}\|\Delta_G - \Delta_\infty\|_2^2, \qquad \|m_G - \Delta_\infty\|_{L^2}^2 \leq \mathbb{E}\|\Delta_G - \Delta_\infty\|_2^2 \leq \mathbb{E}[\Lambda_X^2 \mathcal{B}_G(X)], \tag{16}$$

so both the donor-specific term and the distance of m_G to its limit vanish as $G \to \infty$.

Projection onto the parameter tangent space. In the Hilbert space of square-integrable dense outputs over X, let J_X be the Jacobian of $f_{\bar\theta}(X)$ with respect to the P trained parameters (the LoRA parameters here), and let $\omega \in \mathbb{R}^P$ be a parameter step, assume $\mathbb{E}\|J_X\|_F^2 < \infty$, and let $\mathcal{T} = \overline{\{X \mapsto J_X\omega : \omega \in \mathbb{R}^P\}}$ with projection $P_\mathcal{T}$. Since $m_G - P_\mathcal{T} m_G \perp \mathcal{T}$ and $P_\mathcal{T} m_G - J_X\omega \in \mathcal{T}$,

$$\mathbb{E}\|\Delta_G - J_X\omega\|_2^2 = \mathbb{E}\|\epsilon_G\|_2^2 + \|m_G - P_\mathcal{T} m_G\|_{L^2}^2 + \|P_\mathcal{T} m_G - J_X\omega\|_{L^2}^2. \tag{17}$$

The unchanged student has error $\mathbb{E}\|\epsilon_G\|_2^2 + \|m_G\|_{L^2}^2$, so the largest first-order reduction is exactly $\|P_\mathcal{T} m_G\|_{L^2}^2$, attained up to null directions because $\mathcal{T}$ is finite-dimensional. For the nonlinear student, the Taylor remainder $R_\omega(X) = f_{\bar\theta+\omega}(X) - f_{\bar\theta}(X) - J_X\omega$ satisfies

$$\Big|\, \|\Delta_G - [f_{\bar\theta+\omega} - f_{\bar\theta}]\|_{L^2} - \|\Delta_G - J_X\omega\|_{L^2} \Big| \leq \|R_\omega\|_{L^2}, \tag{18}$$

so optimization error enters through the last term of Equation (17) and departure from linearity through R_ω.

Scope of each statement. The coverage equivalence, the rank identity, and the uniqueness of $1/G$ are deterministic. Hoeffding's bound uses bounded donor mass; the $O(1/G)$ rate uses finite second moments and conditional independence of donors; Lipschitz continuity carries that rate through the frozen suffix. The sets R_k serve the analysis and the measurements, and the method itself reads all prompt positions. The transfer statements concern the dense regression error and its local parameter geometry, and Section 4 measures their consequence for final images.

B Implementation and Reproducibility

This appendix gives the procedures behind Section 4. Information passes between samples only through prompt-token keys and values at the read block, and every reported image is generated by the single-sample student.

Algorithm 1 VOTEX training and deployment

Require: pretrained denoiser f_{θ_0} with joint-attention blocks $1{:}D$; write/read pair $(b, b{+}1)$; prompt distribution $\mathcal{D}$; group size G; prompt groups per update B; updates S; time grid $\{t_i\}_{i=1}^{n}$; learning rate η

Initialize: $\theta \leftarrow \theta_0$, with the LoRA update initialized to zero so that $f_\theta = f_{\theta_0}$
for $s = 1, \dots, S$ **do**
  $\bar{\theta} \leftarrow \mathrm{sg}(\theta)$ ▷ update-local frozen snapshot
  Sample prompts $c_1, \dots, c_B \sim \mathcal{D}$ and one index $i \sim \mathcal{U}\{1, \dots, n\}$; set $t \leftarrow t_i$
  for $r = 1, \dots, B$ **do** ▷ one group per prompt
    Sample $\varepsilon^g \sim \mathcal{N}(0, \mathbf{I})$ and roll out $f_{\bar{\theta}}(\cdot, \cdot, c_r)$ on $\{t_i\}$ from ε^g, $g = 1, \dots, G$; keep $\mathcal{X}_t = \{x_t^g\}_{g=1}^{G}$
    $(I_b^g, E_b^g) \leftarrow \mathrm{Blocks}_{\bar{\theta}}^{1:b-1}(x_t^g, t, c_r)$ for all g ▷ native prefix
    $(I_{b+1}^g, E_{b+1}^g) \leftarrow \mathrm{Block}_{\bar{\theta}}^{b}(I_b^g, E_b^g, t)$ for all g ▷ write: text reads its own image
    $(K_j^T, V_j^T) \leftarrow \mathrm{Proj}_{\bar{\theta}}^{b+1}(E_{b+1}^j)$ for all j; $K_{\mathrm{grp}}^T \leftarrow [K_1^T; \dots; K_G^T]$, $V_{\mathrm{grp}}^T \leftarrow [V_1^T; \dots; V_G^T]$
    for $g = 1, \dots, G$ **do** ▷ read: every image query sees all text banks
      $(Q_g^I, K_g^I, V_g^I) \leftarrow \mathrm{Proj}_{\bar{\theta}}^{b+1}(I_{b+1}^g)$
      $(o_I, \lambda_I) \leftarrow \mathrm{Attn}(Q_g^I, K_g^I, V_g^I)$, $(o_T, \lambda_T) \leftarrow \mathrm{Attn}(Q_g^I, K_{\mathrm{grp}}^T, V_{\mathrm{grp}}^T)$
      $\lambda_T \leftarrow \lambda_T - \log G$; $\lambda^\star \leftarrow \max(\lambda_I, \lambda_T)$; $o_g^{\mathrm{grp}} \leftarrow \dfrac{e^{\lambda_I - \lambda^\star} o_I + e^{\lambda_T - \lambda^\star} o_T}{e^{\lambda_I - \lambda^\star} + e^{\lambda_T - \lambda^\star}}$ ▷ Equation (19)
      Text queries of g attend to $[K_g^I; K_g^T]$; finish block $b{+}1$ with image-attention output o_g^{grp}
      $y^g \leftarrow \mathrm{sg}\left[T_{\bar{\theta}}(x_t^g, t, c_r, \mathcal{X}_t)\right] = \mathrm{sg}\left[\mathrm{Blocks}_{\bar{\theta}}^{b+2:D}(\cdot)\right]$ ▷ frozen suffix, dense target
    $\ell_r(\theta) \leftarrow \frac{1}{G} \sum_{g=1}^{G} \left\| f_\theta(x_t^g, t, c_r) - y^g \right\|_2^2$ ▷ ordinary single-sample path
  $\mathcal{L}(\theta) \leftarrow \frac{1}{B} \sum_{r=1}^{B} \ell_r(\theta)$; $\mathbf{g} \leftarrow \nabla_\theta \mathcal{L}$; $\mathbf{g} \leftarrow \mathbf{g}\, \min\{1, 1/\|\mathbf{g}\|_2\}$
  $\theta \leftarrow \mathrm{AdamW}(\theta, \mathbf{g}; \eta, \beta_1{=}0.9, \beta_2{=}0.95, \lambda{=}0.01)$; discard $\bar{\theta}$ and $\{\mathcal{X}_t\}$
Deploy: sample $x_{t_1} \sim \mathcal{N}(0, \mathbf{I})$ and integrate $f_\theta(x_{t_i}, t_i, c)$ with the original sampler ▷ single sample
return θ

B.1 BACKBONES, GROUP ATTENTION, AND TRAINING

All headline and matched-control experiments generate 1024×1024 images with $G = 4$. The group-size study has two tracks: a fixed-update track with 300 updates of eight prompt groups for $G \in \{1, 2, 4, 8\}$, and a fixed-trajectory track with 9,600 trajectories per run, which gives 600, 300, and 150 updates for $G = 2, 4, 8$. With zero-based block indices of the public implementations, the write/read pairs are blocks 11/12 for SD3.5 Medium (Stability AI, 2024), 9/10 for FLUX.1-dev (Black Forest Labs, 2024), and 29/30 for Qwen-Image-2512 (Qwen Team, 2025). The linked model cards identify the respective checkpoints. Each backbone uses its own tokenizer, text encoder, VAE, scheduler, and weights, with automatic prompt rewriting disabled. Sampling uses 28 steps with guidance 4.5 (SD3.5), 28 steps with guidance 3.5 (FLUX), and 20 steps with guidance 1.0 (Qwen-Image).

The write block is unmodified: its text queries attend to the image and text states of their own sample. In the read block, only the text bank seen by an image query changes: it contains the G text banks of the group, with $-\log G$ added to every text logit, while the image bank stays local. Text queries, residual paths, feed-forward layers, output projections, and all later blocks keep the backbone computation. The two blocks therefore contain one native image-to-text write followed by one cross-sample text-to-image read.

The logical score tensor of shape $[B, G, H, N, GL]$ is never materialized. For each target, one fused attention call over the local image bank returns the normalized output o_I and its log-normalizer λ_I, and a second call over the concatenated text banks returns o_T and λ_T. With $\lambda_T' = \lambda_T - \log G$ and $\lambda^\star = \max(\lambda_I, \lambda_T')$,

$$o^{\mathrm{grp}} = \frac{e^{\lambda_I - \lambda^\star} o_I + e^{\lambda_T' - \lambda^\star} o_T}{e^{\lambda_I - \lambda^\star} + e^{\lambda_T' - \lambda^\star}}, \tag{19}$$

which equals Equation (2), including its $G = 1$ and duplicate-bank identities. The processor reuses the backbone's query/key normalization, positional transformation, masks, adaptive normalization, and output projection. When a group spans devices, only text keys and values are gathered. For SD3.5 ($L = 333$, $d = 1536$, BF16), the four text banks occupy 3.90 MiB and their keys and values 7.80 MiB.

The default runs use rank-64 LoRA with scaling 64 and AdamW for 300 updates, learning rate 10^{-4}, $\beta = (0.9, 0.95)$, weight decay 0.01, and gradient clipping at norm 1.0. Each update contains eight prompt groups, so a run uses $300 \times 8 \times 4 = 9{,}600$ on-policy trajectories; the training seeds are 17, 29, and 43. Matched controls share the prompt order, target trajectories, sampled states, trainable parameters, precision, gradient accumulation, and optimizer, and each run manifest records these settings, the condition-dropout masks, and the trainable-parameter list. Algorithm 1 gives the complete procedure; sampling and the teacher forward run outside the backward graph.

B.2 Data Separation

Training prompts come from the text field of the Pick-a-Pic v2 training split (Kirstain et al., 2023). Text is normalized with Unicode NFKC and whitespace collapsing, and a training prompt is removed if it matches a normalized test prompt exactly or has token 5-gram Jaccard similarity of at least 0.90 with one. Official validation prompts undergo the same normalization and deduplication. Five percent of the remaining pool, selected by $\mathrm{SHA256}(\text{prompt_id}) \bmod 20 = 0$, forms the calibration set, which is excluded from gradient updates and alone selects group size, block pair, LoRA rank, learning rate, and sampling batch. Checkpoints are selected by validation dense MSE at the cadence recorded in each run manifest, with the earlier checkpoint winning an exact tie.

B.3 Evaluation Units, Estimands, and Inference

The prompt is the independent evaluation unit, and images, questions, categories, and diffusion states of one prompt form one cluster. Official aggregation is applied within a prompt before any comparison across prompts, as illustrated in Figure 8: a prompt is split into observable atoms, each output is checked atom by atom, and the counts enter the benchmark aggregate. Base and VoteX comparisons are paired by prompt and, where the interface permits, by initial noise and generation seed. Scores are computed from unrounded values and rounded only for display.



Figure 8: Atom-level check of one GenAI-Bench prompt on two public outputs. Each output is scored per requirement, and the matched atoms (3/3 and 1/3) enter the aggregate.

Test sets enter only the final evaluation, which uses every official prompt with the official annotations, evaluator, and aggregation. Of the 4,081 TIFA prompts, the 4,073 with official question–answer pairs form the TIFA score. Generation failures, safety blocks, missing images, and evaluator failures receive the benchmark's minimum outcome. External systems use their public checkpoint or official training configuration on natively supported backbones and are rescored with the same locked evaluator. Every observation keeps identifiers for the code and aggregation revisions, container and library versions, backbone and evaluator revisions, prompt, seeds, sampled noise, checkpoint, image index, evaluator query, raw score, hardware, GPU time, and failure status; dataset files and evaluator weights are recorded by SHA-256. Each aggregate can therefore be traced to images, queries, and unrounded raw values.

The complementarity statistic of Section 2 is defined per prompt. With Q_p the official atom set of prompt p and $z_{pgq} \in \{0, 1\}$ the correctness of sample g on atom q,

$$D_{\mathrm{comp},p}(G) = \frac{1}{|Q_p|} \sum_{q \in Q_p} \max_{1 \le g \le G} z_{pgq} - \max_{1 \le g \le G} \frac{1}{|Q_p|} \sum_{q \in Q_p} z_{pgq}, \tag{20}$$

reported as $100 D_{\mathrm{comp}}$. Subtracting the best member makes the statistic positive only when correct atoms are distributed across samples. The cross-prompt reference keeps the target's atom set and

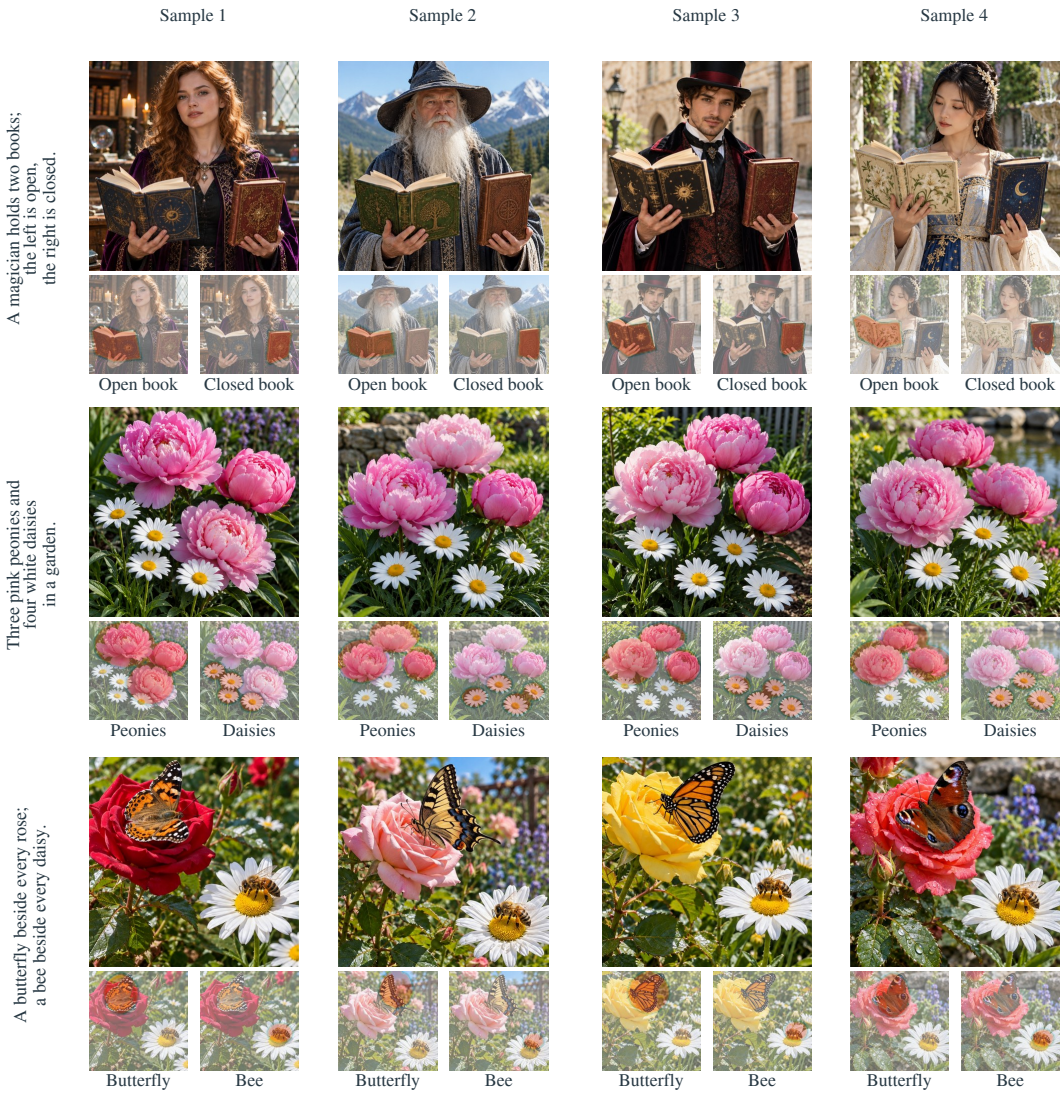


Figure 9: Four samples of the same prompt for three prompts, with crops of the regions that decide each requirement (open and closed book; peonies and daisies; butterfly and bee). The samples differ in which requirements they render correctly, the situation that D_{comp} in Equation (20) measures.

replaces the donors by prompts matched on atom count and token length. For GenEval2 at $G = 4$, the same-prompt values are 15.40, 14.60, and 6.69 points for SD3.5 Medium, FLUX.1-dev, and Qwen-Image-2512, and the cross-prompt values are 0.04, 0.07, and 0.03.

Let d_{rp} be the paired difference for run $r \in \{1, 2, 3\}$ and prompt p after within-prompt aggregation, and $d_r = P_r^{-1} \sum_p d_{rp}$. Main-text intervals treat the three runs as repeated units, $\bar{d} \pm t_{0.975,2}\, s_d/\sqrt{3}$ with $\bar{d} = \frac{1}{3}\sum_r d_r$. A 20,000-draw nested bootstrap resamples runs and, within each run, complete prompt clusters; its 2.5 and 97.5 percentiles give the nested interval. Prompt-conditional significance uses 20,000 sign-flip draws with one Rademacher sign $s_{rp}^{(b)}$ per prompt cluster:

$$T^{(b)} = \frac{1}{3}\sum_{r=1}^{3} \frac{1}{P_r} \sum_{p=1}^{P_r} s_{rp}^{(b)} d_{rp}, \qquad p = \frac{1 + \sum_{b=1}^{20000} \mathbf{1}\{|T^{(b)}| \geq |T_{\text{obs}}|\}}{20001}. \tag{21}$$

Holm correction is applied per backbone to the five scalar outcomes, and separately to the eight T2I-CompBench++ categories, since that benchmark defines no overall score: for sorted raw values $p_{(1)} \leq \cdots \leq p_{(m)}$, $p_{(k)}^{\text{H}} = \max_{j \leq k} \min\{1, (m - j + 1)p_{(j)}\}$.

Figure 9 shows the per-sample annotation that underlies D_{comp}. For each sample, the regions that decide a requirement are cropped and checked separately; the union over samples and the best single sample are then computed from these checks.

B.4 Source Controls and Text-State Interventions

Table 4: GenEval2 source controls. Mean GM $\times 100$ (↑); full model names are given in Table 1.

Source states	SD3.5	FLUX	Qwen
Full exchange	**58.4**	**60.9**	**77.1**
Duplicated target	51.3	53.6	72.8
Cross-prompt donors	51.5	53.8	72.8

The source controls keep $G = 4$, the target trajectory, tensor shapes, attention work, optimizer updates, and evaluation fixed. *Full exchange* uses four independent samples of the same prompt. *Duplicated target* copies the target state into all four slots before the write block, so every donor text bank duplicates the target's. *Cross-prompt donors* fill the three donor slots with states of other prompts matched on token length and object count while keeping the target prompt and trajectory, which preserves real activations and group computation and breaks the shared meaning of the prompt positions. The controls match training prompts, target seeds, diffusion states, total trajectories, updates, trainable parameters, checkpoint selection, and p90 A100 time; cheaper paths execute shape-matched attention whose outputs are discarded before the loss.

Table 5: Frozen-teacher interventions on GenEval2. Mean target-atom accuracy (%, ↑).

Intervention	SD3.5	FLUX
Full read	**75.84**	**76.50**
Helpful donor masked	55.92	56.60
Unrelated donor masked	70.93	71.72
Cross-prompt replacement	44.74	45.17

Interventions act on frozen group teachers, so they change inference through one edge and leave the parameters fixed. An independent selection bank assigns donor roles with the official atomic questions: for target atom q, a helpful donor answers q correctly where the target does not, and a deterministic hash of the prompt ID breaks ties. The selected donor index is transferred to an intervention bank with disjoint generation seeds, and whether the donor again answers q correctly is logged as role repetition, with all attempts kept in the analysis. At every denoising step, the intervention changes only donor text keys and values at the read block. *Helpful masking* removes the selected bank; *unrelated masking* removes an equally sized bank from a donor that is incorrect on q and matched on its other atoms; *cross-prompt replacement* substitutes a length-matched bank from another prompt. The primary outcome is the official correctness of q, and other-atom correctness and the complete-prompt score measure changes elsewhere.

B.5 Group Teacher and Single-Sample Student

Table 6: GenEval2 group-to-single transfer. Mean GM $\times 100$ (↑).

Prediction path	SD3.5	FLUX	Qwen
Base	51.34	53.64	72.75
Group teacher	**59.28**	**61.73**	**77.63**
Single student	58.31	60.74	77.09

Table 6 reports final-image GenEval2 scores of the base model, the frozen group teacher, and the exported single-sample student, with the checkpoints of the main comparison. The teacher generates with the group read at every step, and the student generates from one ordinary trajectory. The student improves over the base model on all three backbones and stays within 1.0 point of its teacher. Figure 10 shows ten teacher and ten student samples of one prompt. The student samples match the teacher samples in scene layout and in how they place the children, hats, and balloons, and they vary in identity, position, and background to a similar extent, so the student reproduces the group behavior from one trajectory.

B.6 Human Evaluation and Quality Checks

The blind study uses 600 prompts, 100 per benchmark, selected before generation by SHA-256 order of prompt IDs. Base and VoteX images are paired by backbone, prompt, and initial noise, and raters see the prompt and two randomized sides without method, backbone, or seed information. Three backbones and three runs give 5,400 pairs and 16,200 judgments from 54 raters. Raters answer separate questions on complete-prompt fidelity (objects, attributes, relations, counts) and visible defects (malformed structure, duplicated parts, text artifacts, texture, composition), each with a tie option; practice examples give no feedback on study images. The preference model treats the prompt as the sampling unit, includes a rater random intercept, and separates training runs; the audit



Figure 10: Ten samples of the group teacher (training) and of the single-sample student (deployment) for one prompt.

record keeps per-rater decisions, win and tie rates, prompt-cluster intervals, Krippendorff's α, and attention-check exclusions fixed before analysis. HPS v2.1 (Wu et al., 2023), PickScore (Kirstain et al., 2023), and ImageReward (Xu et al., 2023) are computed on the official-evaluation images, and LPIPS is aggregated within prompts and read together with fidelity, defects, and atom-answer diversity.

B.7 Deployment Measurement

Export removes the snapshot, group sampler, text buffers, and group-read path. A fresh process loads only the student and the base single-prompt stack with matched tokenizer, encoder, VAE, scheduler, guidance, precision, input length, and 1024×1024 output; matched prompt and noise inputs give identical latents and images in the export and in the group-disabled training code. After 100 warm-up batches on A100 GPUs, base and VoteX alternate for at least 1,000 synchronized batch-one images per backbone, and we report the relative p90 change $100[Q_{0.90}(T_{\mathrm{VoTEx}}) - Q_{0.90}(T_{\mathrm{Base}})]/Q_{0.90}(T_{\mathrm{Base}})$ with paired peak memory. The training ledger separates rollouts, teacher forward, student forward and backward, text-state communication, and peak memory.

C Additional Analyses

C.1 How the Group Read Uses Donor Evidence

Figure 11 follows the read of Section 3.2 back to donor pixels. A query on the red pair of either target reaches the donor mainly through the prompt tokens that the donor's red pair wrote into, and a query on the green pair reaches the donor's green pair, although the pairs sit at different image positions in the three samples. A query on the bench spreads over the donor's bench. The two-step path multiplies the read weights of block $b{+}1$ with the write weights of block b, so it shows which donor content the prompt tokens carry to each target region; the residual, feed-forward, and projection layers act on top of it.

Figure 12 shows the image-token projections that the read block uses. In all three projections, the boat, the dolphins, and the birds take distinct colors, so image queries at different objects point in different directions and select different prompt tokens in Equation (2). The fields change across depth, which is why the write/read pair is selected on calibration prompts per backbone (Appendix B.1); Figure 17 compares depths directly.

Figure 13 separates the two terms of Equation (6) on real teacher corrections. The conditional mean, which is the part m_G that a single-sample student can learn, follows the objects of the target: the individual pastries in the first prompt and the sails in the second. The two leading modes of donor variation are nearly uniform over the image, and moving along them (points A–D) shifts the tone of the correction while the object layout stays fixed. The donor-specific part ϵ_G is therefore dominated by global, low-structure variation, and the object-specific content of the correction lies in its predictable part, which the student absorbs (Table 6).

Figure 14 shows the read when the group lacks the needed evidence. The group read keeps each target close to its local-read output: the yellow pots stay filled and the number of red cranes stays unchanged, with the rest of each image preserved. VoteX reuses what some sample of the group

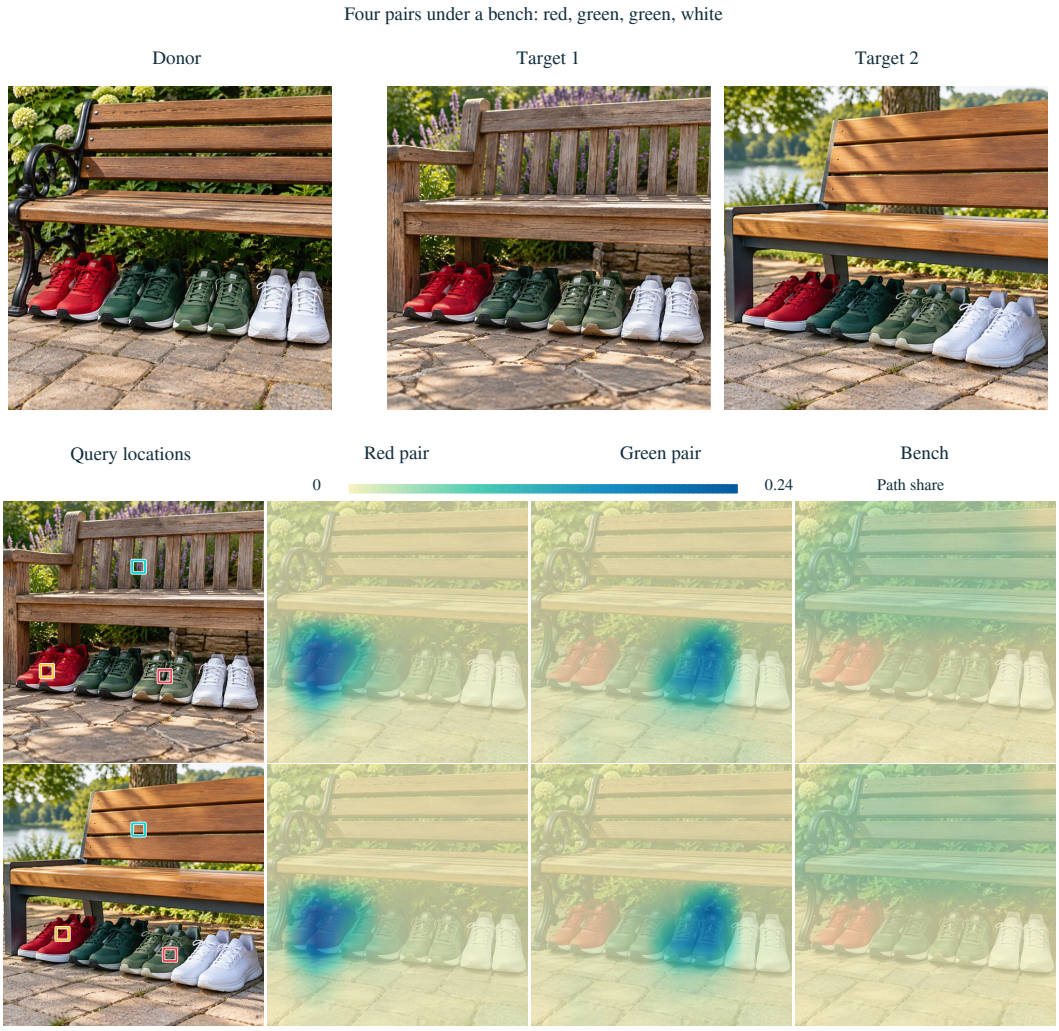


Figure 11: Correspondence through prompt tokens for "four pairs under a bench: red, green, green, white." *Top:* a donor and two targets with different layouts. *Bottom:* for three query locations in each target (boxes), the share of the two-step path target query → prompt token → donor image patch that ends at each donor patch, on a common scale.

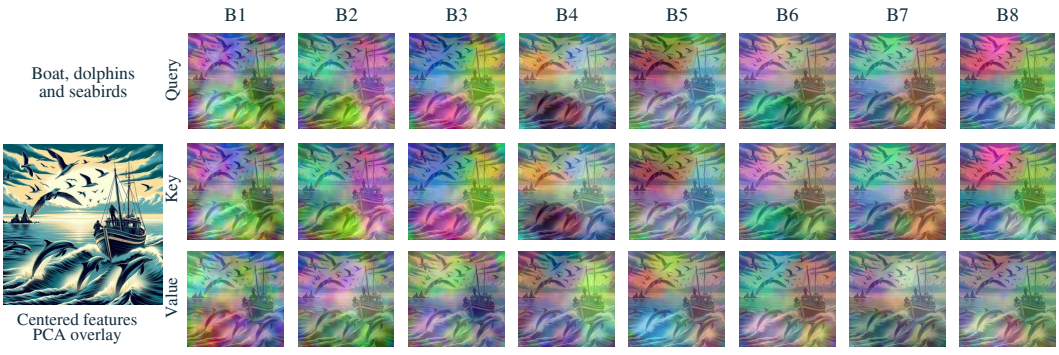


Figure 12: Centered query, key, and value fields of the image tokens over eight consecutive blocks B1–B8 for one image, projected on a shared three-component PCA basis and overlaid on the image.

already renders, which is the quantity that D_{comp} measures, and larger groups or refreshed on-policy samples raise the chance that each requirement appears in the group (Theorem 1).

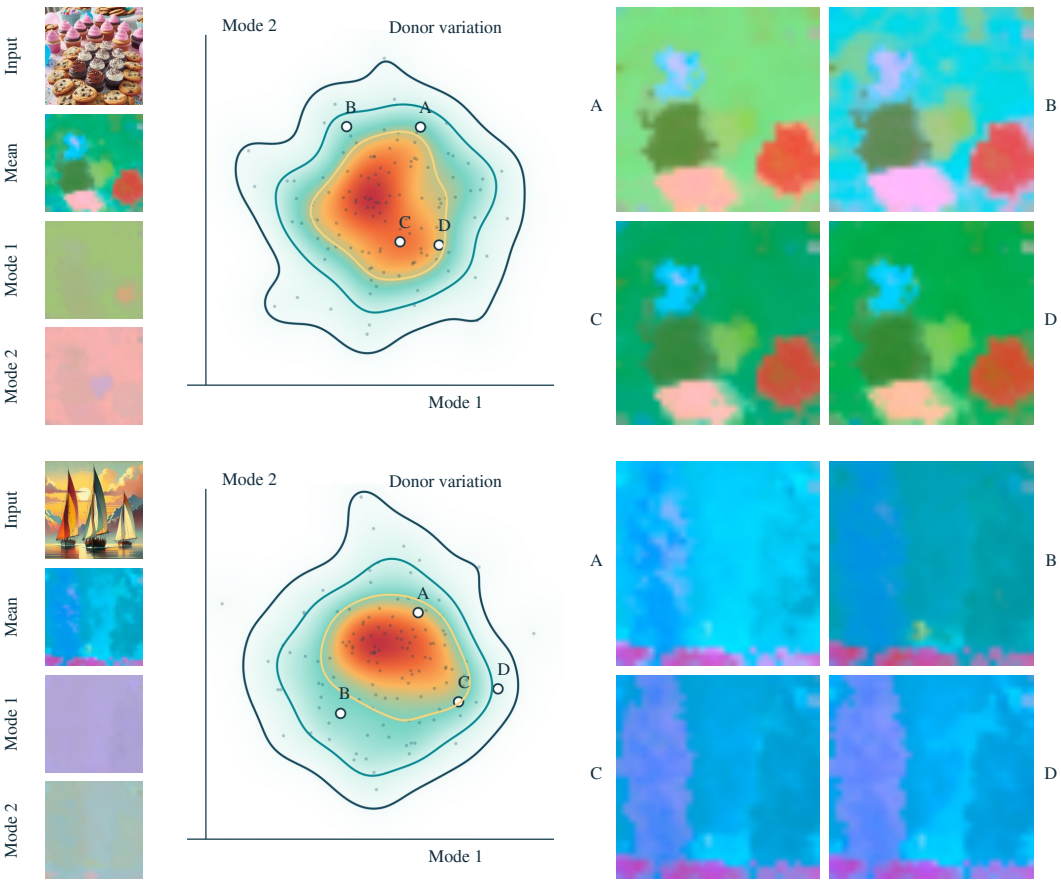


Figure 13: Variation of the teacher correction over 128 donor draws for a fixed target (two prompts). *Left:* input, conditional mean, and the two leading modes of variation. *Center:* density of the draws in the plane of the two modes. *Right:* rank-two reconstructions at the marked points A–D.



Figure 14: Local and group read on the same targets when no donor carries the missing evidence. *Left:* all green pots hold tulips and all yellow pots are empty; no donor renders an empty yellow pot. *Right:* five cranes with exactly one red; no donor renders exactly one red crane.

C.2 Design Studies

Figure 15 shows where each component acts in the score distribution. VoteX shifts every quantile upward, by about 10 points in the upper part of the distribution, while removing the $-\log G$ shift, reading donor image keys and values, or using a final-token target yields shifts of a few points that fall to zero over wide ranges. The right panel compares the geometric mean GM, which drops sharply when any atom is missed, with the arithmetic mean AM. The VoteX gain in GM exceeds its gain in AM over the upper three quarters of the distribution, by more than for the final-token target: the

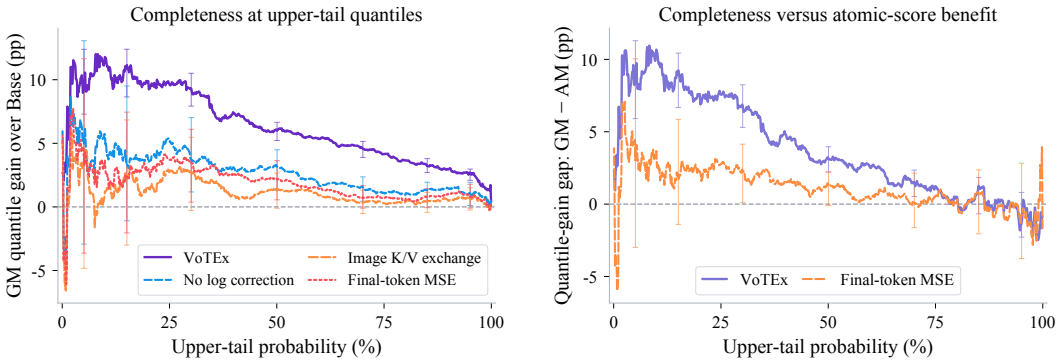
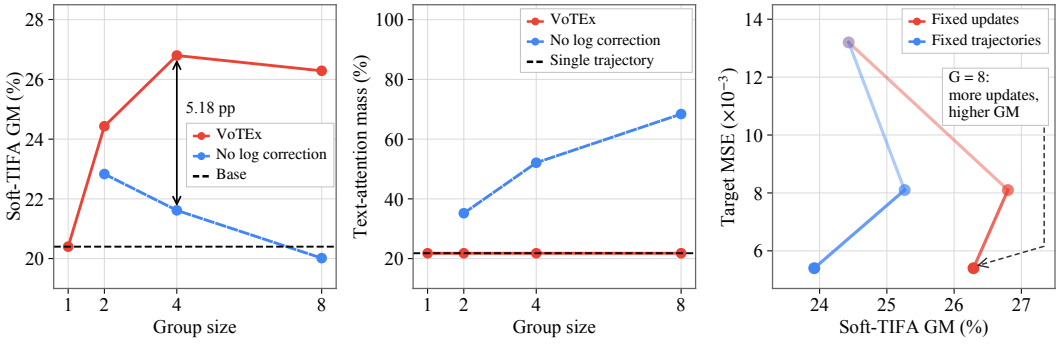


Figure 15: Component and objective contrasts on SD3.5 Medium, GenEval2. *Left:* GM quantile shifts relative to Base for VoTEX and three alternatives; at upper-tail probability u, the evaluated quantile is $1 - u$. *Right:* GM quantile shift minus AM quantile shift for VoTEX and a final-token MSE target. Bars are pointwise 95% nested seed-and-prompt bootstrap intervals (20,000 resamples) at seven positions.

dense group target improves the joint satisfaction of all atoms of a prompt more than the average atom score, the effect predicted by per-query exchange of complementary atoms.



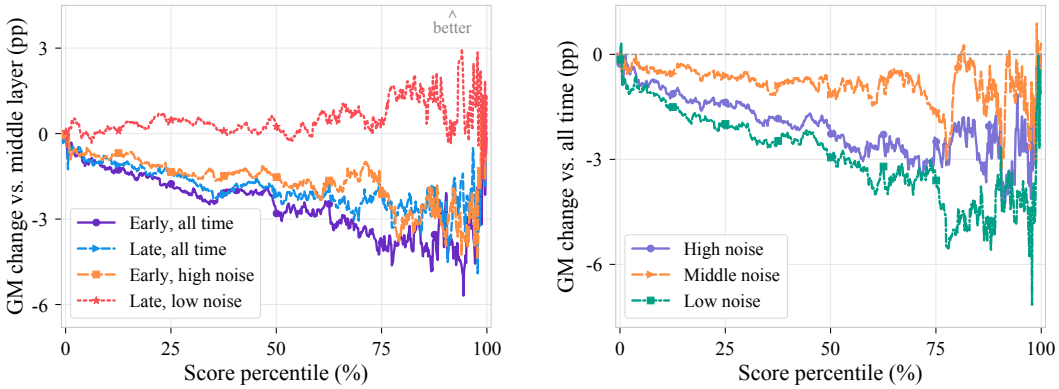
(a) SD3.5: group-size response (b) Multiplicity normalization (c) Budget-dependent trajectories

Figure 16: Group size on SD3.5 Medium. **(a)** Completeness with and without the $-\log G$ shift. **(b)** Text-attention mass of the image queries. **(c)** Dense-target MSE versus completeness under fixed-update and fixed-trajectory training, connected in the order $G = 2, 4, 8$, with opacity increasing with G.

Figure 16 connects the group-size rows of Table 3 to Theorem 2. With the shift, the text-attention mass stays at its single-sample value for every G (b), and completeness peaks at $G = 4$ (a). Without it, the text mass grows with G, the target image context loses weight, and completeness falls as the group grows. Panel (c) places each configuration by its dense-target MSE and score; the two budgets trace different paths as G grows, so the regression error alone does not rank group sizes, and we select G by the final score on calibration prompts.

Figure 17 supports the default of a middle-depth pair applied at all sampled times. Early pairs are worse than the middle pair in every window, since their image tokens have not yet formed object-specific queries (Figure 12). A late pair is worse over all times and better only within the low-noise window, where the layout is fixed and the read refines details. Restricting a middle pair to one noise window lowers completeness in every case, most for the low-noise window, because counts and relations are decided early in sampling and need the exchange at high noise.

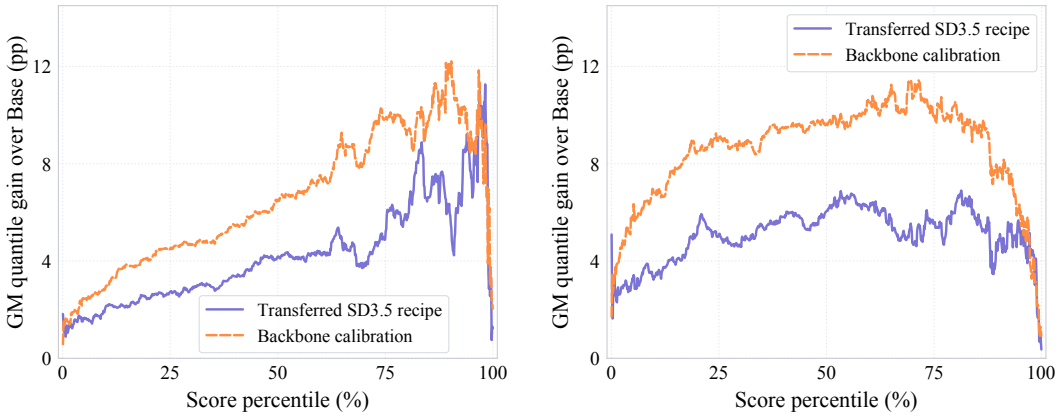
Figure 18 tests how much of the method depends on per-backbone tuning. The SD3.5 recipe, used unchanged on FLUX and Qwen-Image, improves every percentile of both backbones. Calibrating the block pair and hyperparameters on the target backbone adds a further gain across most percentiles. The mechanism thus carries across joint-attention architectures, and calibration on held-out prompts, which uses no test data, sets its size.



(a) Depth effects within each noise window. The late-layer advantage depends on the window and on the score percentile.

(b) Window effects at middle depth. The low-noise-only window incurs the largest average completeness penalty.

Figure 17: Depth and noise window of the write/read pair on SD3.5 Medium. **(a)** GM quantile contrasts of early and late pairs against the middle pair within the same time window. **(b)** At middle depth, restricted noise windows against the all-time default. Each contrast retains all 800 order statistics after seed averaging.



(a) FLUX.1-dev

(b) Qwen-Image-2512

Figure 18: Transfer of the SD3.5 Medium recipe to FLUX.1-dev and Qwen-Image-2512. Curves show GM quantile gains over each backbone's Base for the transferred recipe and for calibration on the target backbone; both panels share the vertical scale.

C.3 Additional Samples

Figure 19 shows five samples per prompt and model, so that compliance and diversity can be read together. Both rows of a block share the seeds, so the difference within a column reflects the model change. The VoteX samples vary in subject, pose, framing, and background to a similar extent as the base samples, consistent with the small within-prompt LPIPS change in Section 4.5.

Figure 20 follows three requirements over training. The requirements change at different checkpoints: the yellow pillows of the first two checkpoints give way to blue ones, while the mug count and the chimney requirement change over later checkpoints. The scene of each prompt stays similar across checkpoints, so the updates act on the specific requirements and leave the overall composition in place.



Figure 19: Five independent samples of Base and VoTEX for six prompts with clothing, color, count, and relation requirements. Each block uses the same five seeds for both models.

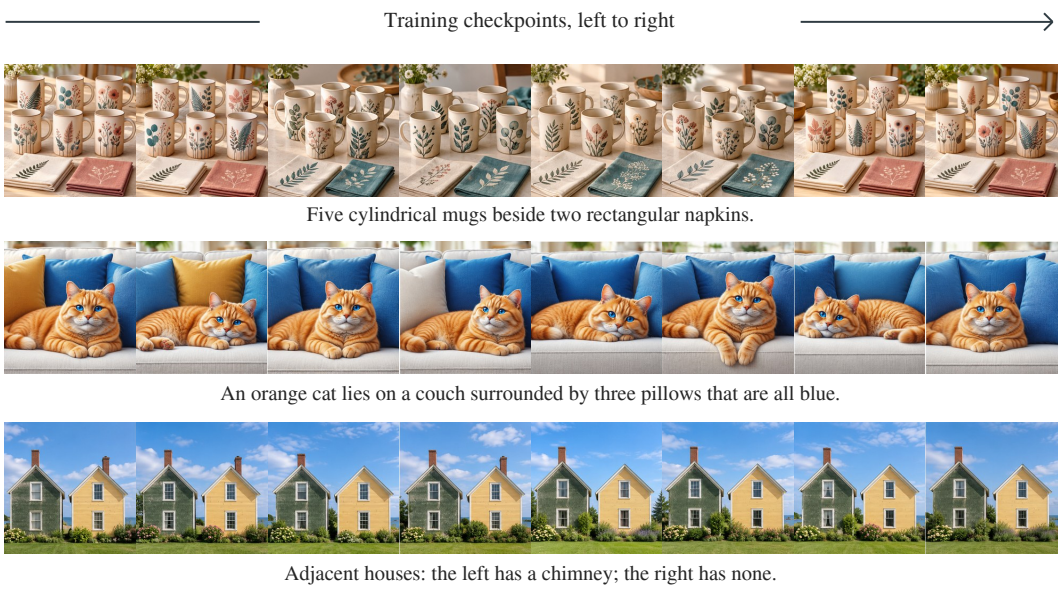


Figure 20: Samples of three prompts at eight training checkpoints, left to right, with count (five mugs, two napkins), universal attribute (three pillows, all blue), and negation (only the left house has a chimney) requirements.