ON THE CROSS-BRANCH CORRELATION OF QUANTIZATION ERROR IN CLASSIFIER-FREE GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance runs a network twice per step, with and without the condition, and extrapolates along the difference of the two predictions. Quantized models are judged per branch, yet the guided error equals $E_c + s(E_c - E_u) + s(1+s)V$, where E_c and E_u are the branch errors, s is the guidance scale, and V, the mean squared difference of the two errors, depends on a cross term that no per-branch metric sees. In six inference stacks, 75.8% of the activation quantizers give the branches different scales or random numbers, and on these, 19% to 28% of the values held by both branches receive different codes. We propose PairRound, which lets the branches share what decides their codes: the scale, and the random numbers where rounding is stochastic. Codes then differ only where values differ, and a quadratic in s states when the coarser grid that sharing gives one branch pays off, a condition our measurements meet. At matched bits and seeds, the shared scale lowers the trajectory distance to the BF16 sampler by 60% to 68% on eleven image, video and language backbones, improves FID, VBench or perplexity on every one, and composes with eight published methods and improves each of them.

1 INTRODUCTION

Classifier-free guidance (CFG) steers a diffusion model toward a prompt (Ho & Salimans, 2021) and drives text-to-image, text-to-video and language models (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Team Wan et al., 2025; Sanchez et al., 2024). At each step the network runs twice, with and without the condition, and the sampler combines the conditional and unconditional predictions $\boldsymbol{\epsilon}_c$ and $\boldsymbol{\epsilon}_u$ into $(1+s)\boldsymbol{\epsilon}_c - s\boldsymbol{\epsilon}_u$, with a guidance scale s usually between 4 and 11. Two runs per step double the cost, so deployed models are quantized to eight or four bits (Li et al., 2023a; 2025), and each quantizer is judged branch by branch, by the error of a quantized tensor or of one forward pass. This assumes that two accurate branches give an accurate guided output, but the sampler never uses a branch alone, and a scaled difference of two accurate terms can be inaccurate.

To see how inaccurate, write $\boldsymbol{e}_c$ and $\boldsymbol{e}_u$ for the quantization errors of the two branches and E_c, E_u for their mean squared norms. The guided error is $E_c + s(E_c - E_u) + s(1+s)V$, where V is the mean squared difference of the two errors (Proposition 1). The per-branch errors enter linearly in s, but V is multiplied by $s(1+s)$, which is 48.75 at the common guidance weight $w = 1 + s = 7.5$, and $V = E_c + E_u - 2C$ depends on the cross term $C = \mathbb{E}\langle \boldsymbol{e}_c, \boldsymbol{e}_u \rangle$, which no per-branch metric sees. In Figure 1, four couplings of the rounding of the two branches give per-branch errors within 8% of each other, yet at $s = 6.5$ their trajectory distances to the BF16 sampler differ by up to 58 times, and at $s = 0$ they coincide. Guided quality thus depends on a term that current metrics do not measure.

This term is set at the activation quantizers, where a tensor is rounded onto a grid whose step is its scale σ. The two branch errors agree only if a value held by both branches gets the same code in both, and two properties decide this: the rounding rule, deterministic or stochastic, and the scale, shared by the branches or computed per branch. Among the $1{,}753$ activation quantizers that we surveyed in six inference stacks, 74.7% round deterministically with per-branch scales, as per-token dynamic quantization does (Dettmers et al., 2022; Yao et al., 2022), 1.1% round stochastically with separate random numbers, and only 24.2% are consistent by design. On the inconsistent quantizers, 19% to 28% of the coordinates that carry the same value in both branches receive different codes. If these mismatches cause the loss, their effect must grow with the guidance scale, and it does on each of the backbones on which we vary their guidance scale (Figures 1 and 3, Section 5.4 and Appendix C.1).

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.

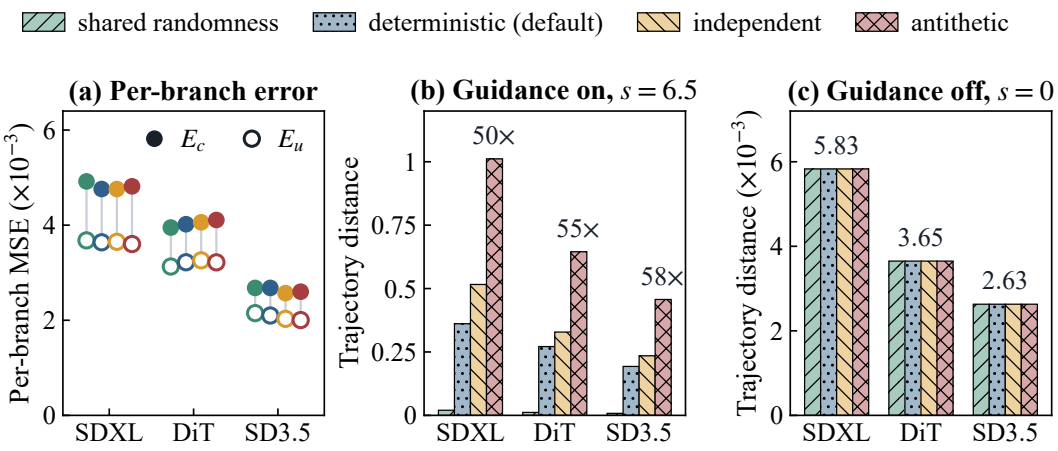


Figure 1: The guided output separates couplings that per-branch errors cannot tell apart. Four couplings of the rounding of the two branches on SDXL, DiT and SD3.5 at W4A8: stochastic rounding with shared, independent or antithetic random numbers, and deterministic rounding with per-branch scales, the deployed default. (a) Per-branch errors E_c (filled) and E_u (open). (b) Trajectory distance to the BF16 sampler at $s = 6.5$; labels give the antithetic over the shared distance. (c) The same at $s = 0$, where only one branch runs; labels give the common distance.

It follows that the branches should share whatever decides their codes. Where rounding is stochastic, PairRound gives both branches the same random numbers, which lifts the cross term to its Fréchet–Hoeffding upper bound (Fréchet, 1951; Hoeffding, 1940). Where scales are per branch, PairRound gives both branches the larger of the two scales, so that equal values land on one code, at the price of a coarser grid for one branch. Under either rule, the values x_c and x_u held at a coordinate, when at most one step apart, receive different codes with probability $p = |x_c - x_u|/\sigma$, and the difference error $\sigma^2 p(1-p)$ vanishes where they agree (Proposition 3). A quadratic in s states when the coarser grid is worth its price (Proposition 4), and the larger scale is the finest shared grid that clips neither branch (Corollary 1), so the rule has no free parameter and needs neither calibration nor retraining.

On eleven backbones, seven image, one video and three language models, at matched bits and seeds, the shared scale of PairRound lowers the trajectory distance to the BF16 sampler by 60% to 68% and improves FID, VBench or perplexity, and the preference or win rate, on every one (Table 1 and Appendix C.2). It removes over 60% of the distance left by each of eight published methods, which target the per-branch errors while PairRound targets V (Table 2); its measured coefficients meet the condition of Proposition 4 (Table 4); and it adds 12 μs per step and branch. We contribute the decomposition, the survey, the two rules with their analyses, and ablations of each design decision.

2 Related Work

Classifier-free guidance. CFG (Ho & Salimans, 2021) replaced the classifier of Dhariwal & Nichol (2021) by a second run of the same network. Later work sets when and how strongly it acts, with weight schedules (Wang et al., 2024), a limited interval (Kynkäänniemi et al., 2024), a weaker second model (Karras et al., 2024) or a manifold constraint (Chung et al., 2025), and language models use the same rule (Sanchez et al., 2024). All keep the guided prediction a scaled difference of two runs, the structure our analysis needs, whereas guidance distillation folds both runs into one (Meng et al., 2023). We keep the guidance rule unchanged and study how both of its runs are quantized together.

Quantization of diffusion models. Post-training methods calibrate per time step (Li et al., 2023a; Shang et al., 2023), correct the noise of a branch (He et al., 2023), preserve temporal features (Huang et al., 2024a), mix precisions (Zhao et al., 2024), handle the outliers of diffusion transformers (Chen et al., 2025; Wu et al., 2024), cover video (Zhao et al., 2025) or absorb outliers into a low-rank branch (Li et al., 2025); others fine-tune under the quantizer (He et al., 2024; Sui et al., 2024). Per-token dynamic scales come from language-model quantization (Dettmers et al., 2022; Yao et al., 2022; Xiao et al., 2023). These methods target E_c and E_u within each branch, whereas PairRound acts on the pair of branches, so we compare it with these methods and also apply it underneath each of them.

Rounding and coupled randomness. Stochastic rounding is unbiased (Gupta et al., 2015; Connolly et al., 2021; Croci et al., 2022), dither decouples the rounding error from the signal (Roberts, 1962; Gray & Stockham, 1993), and the uniform model of rounding error dates back to Bennett (1948). Common random numbers and antithetic variates are the two extreme couplings of two Monte Carlo estimates (Glasserman & Yao, 1992; Hammersley & Morton, 1956), and the Fréchet–Hoeffding bounds give the full range of dependence at fixed marginals (Fréchet, 1951; Hoeffding, 1940; Nelsen, 2006). To our knowledge, the tools above have not yet been applied to quantized guided sampling.

3 Attributing the Guided Error

At step t the sampler runs the network on the state $\boldsymbol{x}_t$ with the condition c and without it, and combines the two predictions $\boldsymbol{\epsilon}_c, \boldsymbol{\epsilon}_u \in \mathbb{R}^D$ with the guidance weight $w = 1 + s$:

$$\hat{\boldsymbol{\epsilon}}_t = (1+s)\,\boldsymbol{\epsilon}_c(\boldsymbol{x}_t, c) - s\,\boldsymbol{\epsilon}_u(\boldsymbol{x}_t). \tag{1}$$

A quantizer maps an activation x to $\sigma\ \mathrm{rnd}(x/\sigma)$, where the scale σ is the grid step and rnd rounds $y = x/\sigma$ to the nearest integer (deterministic rounding) or rounds it up with probability equal to its fractional part $f = y - \lfloor y \rfloor$, using a uniform random number u (stochastic rounding). The scale is fixed per tensor or channel, or computed per token at run time. A quantized network returns $\boldsymbol{\epsilon}_c + \boldsymbol{e}_c$ and $\boldsymbol{\epsilon}_u + \boldsymbol{e}_u$, so the guided prediction carries the error $\hat{\boldsymbol{e}} = (1+s)\boldsymbol{e}_c - s\boldsymbol{e}_u$. We write $E_c = \mathbb{E}\|\boldsymbol{e}_c\|^2$ and $E_u = \mathbb{E}\|\boldsymbol{e}_u\|^2$ for the per-branch errors, $C = \mathbb{E}\langle \boldsymbol{e}_c, \boldsymbol{e}_u \rangle$ for their cross term, $\boldsymbol{\delta} = \boldsymbol{e}_c - \boldsymbol{e}_u$ for the branch difference and $V = \mathbb{E}\|\boldsymbol{\delta}\|^2$ for its mean square, with expectations over prompts, seeds and rounding randomness. Section 5 measures how far these errors move the whole sampling trajectory.

3.1 Decomposing the Guided Error

Proposition 1 (Decomposition of the guided error)**.** *For any branch errors with finite second moments and any guidance scale $s \geq 0$, the error of the guided prediction satisfies*

$$\boxed{\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 = (1+s)^2 E_c + s^2 E_u - 2s(1+s)\,C = E_c + s\,(E_c - E_u) + s(1+s)\,V.} \tag{2}$$

With E_c and E_u fixed, the guided error can range over a closed interval of length $4s(1+s)\sqrt{E_c E_u}$.

The proof writes $\hat{\boldsymbol{e}} = \boldsymbol{e}_c + s\boldsymbol{\delta}$ and applies the Cauchy–Schwarz inequality to C (Appendix A.1). The first two terms depend on each branch alone and are linear in s, while the last depends on the pair and grows like $s(1+s)$; at the first layer of SDXL under the default, it is 96% of the guided error (Appendix A.6). For an ensemble with convex weights the same identity says that disagreement lowers the error (Krogh & Vedelsby, 1995); the negative weight of guidance flips the sign, so disagreement between the branch errors raises the guided error, with the largest coefficient of the three terms once $s > 0.62$, which includes all guidance scales in common use (Appendix A.5).

Proposition 1 implies that per-branch metrics cannot rank quantizers for guided sampling, as Figure 1 shows for four couplings. The same holds layer by layer on SDXL, where independent rounding gives 12 to 26 times the guided error of shared rounding at nearly equal per-branch errors (Figure 2a).

3.2 Sources of Inconsistent Rounding

A value held by both branches gets the same code only if both round it on the same grid with the same outcome. Deterministic rounding with a shared scale is consistent by design. Per-token dynamic quantization (Dettmers et al., 2022; Yao et al., 2022) computes the scale of each token from its own values, and the activations of the two branches differ, so a value they share lands on two grids; stochastic rounding with separate random numbers can round it in two ways even on one grid. We classified all 1,753 activation quantizers on the path of one guided step in six stacks: diffusers (von Platen et al., 2022), TensorRT (NVIDIA, 2024), vLLM on its guidance path (Kwon et al., 2023), a ComfyUI quantization plugin (comfyanonymous, 2023), the official SANA kernel (Xie et al., 2025), and one vendor stack. Deterministic rounding with per-branch scales holds 69% to 80% of the quantizers of every stack and 74.7% overall, stochastic rounding 1.1%, and consistent quantizers 24.2% (Figure 2c and Appendix B.1). On the inconsistent quantizers, 19% to 28% of the coordinates that carry the same value in both branches receive different codes, against none on the consistent ones.

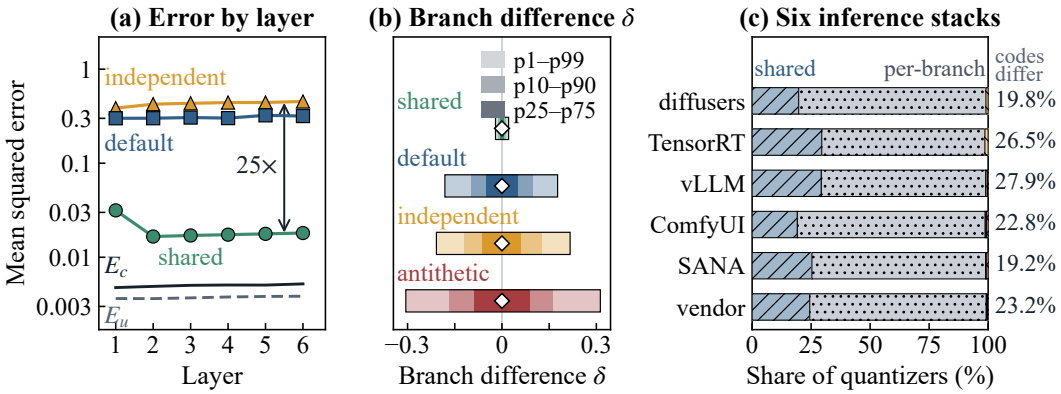


Figure 2: Where the guided error comes from. (a, b) SDXL at W4A8 and $s = 6.5$. (a) Per-branch errors and the guided error $\mathbb{E}\|\hat{e}\|^2$ of three couplings at six layers; the arrow gives the ratio of independent to shared randomness between layers 5 and 6. (b) Branch difference δ over coordinates: bands span the 1st to 99th, 10th to 90th and 25th to 75th percentiles, and diamonds mark the median. (c) Activation quantizers of six stacks with a shared scale (blue) or per-branch scales (gray); stochastic types are thin slivers at the right. Numbers give the share of same-value coordinates that receive different codes on the inconsistent quantizers.

The resulting branch difference is centered at zero but wide: under the default, its central 98% is 8.6 times as wide as under shared randomness (Figure 2b). In the deployed models the cross term is thus set mostly by per-branch scales, and a repair needs to share the scale as well as the random numbers.

4 PairRound: Sharing the Rounding of the Two Branches

PairRound keeps the rounding rule of every quantizer and changes only what the two branches share. For a pair of corresponding tensors $\boldsymbol{X}_c$ and $\boldsymbol{X}_u$ with per-branch scales σ_c and σ_u, both branches are rounded with one scale and one stream of random numbers:

$$\hat{\boldsymbol{X}}_b = \bar{\sigma}\ \mathrm{rnd}\big(\boldsymbol{X}_b/\bar{\sigma};\, u\big), \quad b \in \{c, u\}, \qquad \bar{\sigma} = \max(\sigma_c, \sigma_u), \tag{3}$$

where the random number u of stochastic rounding is drawn once for both branches and is unused under deterministic rounding. The rule needs no calibration and no training. Tensors are paired only when they correspond position by position, as latent activations and decoded tokens do; quantizers whose tensors do not, such as those holding the two prompts, keep their own scales. When a step runs one branch, at $w = 1$ or in a guidance-distilled model, nothing changes. The maximum can be taken per coordinate, or pooled per channel or per tensor with less synchronization and a smaller gain. On SDXL the exchange of scales takes $11.7\,\mu$s per step and branch, about 0.1% of the 11.4 ms kernel time, and the throughput of each granularity is within 1.5% of the default (Algorithm 1, Appendix F).

4.1 Sharing the Randomness

At a stochastic quantizer the error of a value with fractional part f is $e = \sigma(\mathbf{1}[u < f] - f)$, which has mean zero (Croci et al., 2022). The first rule of PairRound gives both branches the same u by keying the random stream by quantizer and step instead of by quantizer, branch and step; on SDXL this costs $7.0\,\mu$s per step and branch, while each branch keeps its error distribution (Proposition 2).

Proposition 2 (Coupling of stochastic rounding)**.** *Let the two branches round values with fractional parts f_c and f_u stochastically on grids of scales σ_c and σ_u, with uniform random numbers u_c and u_u. For every joint distribution of (u_c, u_u), each branch error keeps its own distribution, and*

$$\mathbb{E}[e_c e_u] = \sigma_c \sigma_u \big(\Pr[u_c < f_c,\, u_u < f_u] - f_c f_u\big). \tag{4}$$

This cross term is largest for shared random numbers ($u_u = u_c$), zero for independent ones, and smallest for antithetic ones ($u_u = 1 - u_c$). On one grid, shared ones give equal values equal errors.

The joint probability in Eq. (4) obeys the Fréchet–Hoeffding bounds, which shared and antithetic random numbers attain (Appendix A.2). Since the marginals do not move, Proposition 1 turns the largest cross term into the smallest V: shared randomness gives the lowest guided error among all couplings of the random numbers with these marginals, and it does so at every guidance scale $s > 0$.

4.2 Sharing the Scale

At a deterministic quantizer with per-branch scales, the second rule gives both branches $\bar{\sigma} = \max(\sigma_c, \sigma_u)$ for each pair of corresponding scales. To see why sharing lowers the difference error, we analyze one coordinate under the uniform model of rounding error, in which the fractional part of x/σ is uniform over the values of a quantizer and independent of the difference between the branches (Bennett, 1948; Widrow & Kollár, 2008). Let $\eta = x_c - x_u$ be the difference of the values the two branches carry at a coordinate, and let $\delta = e_c - e_u$ be the difference of their rounding errors there.

Proposition 3 (Code mismatch and the difference error)**.** *(a) If both branches use one scale σ and round either to the nearest code under the uniform model or stochastically with shared random numbers, then for $|\eta| \le \sigma$ their codes differ with probability $p = |\eta|/\sigma$, and*

$$\mathbb{E}[\delta^2] = \sigma^2 p\,(1-p), \tag{5}$$

which is zero when $x_c = x_u$. (b) If $x_c = x_u = x$ but $\sigma_c = (1+\gamma)\sigma_u$ with $\gamma > 0$, nearest rounding under the uniform model gives different codes with probability $p = |x|\gamma/((1+\gamma)\sigma_u)$ when $p < 1$, and $\mathbb{E}[\delta^2] = \sigma_u^2 p(1-p)$ up to a remainder $O(\sigma_u^2\gamma(p+\gamma))$ that is small when $\gamma \ll \min(p, 1-p)$.

Both parts express the difference error through the probability that the two codes differ (Appendix A.3). Rounding with a shared random number is flooring after a common random shift, so for stochastic rounding part (a) holds for every pair of values and needs no uniform model. Per-branch grids make p positive where the branches agree, and p grows with the code magnitude $|x|/\sigma_u$, which reaches 127 at eight bits: a 1% scale gap already gives $p \approx 0.4$ at $|x|/\sigma_u = 40$. A shared grid, with a shared random shift where rounding is stochastic, makes p depend on the value difference alone, and this is what the two rules of PairRound provide. The price is a coarser grid for one branch. On a coordinate that both branches share, widening the grid of the unconditional branch by the factor $1+\gamma$ adds about $\gamma\sigma_u^2/6$ to its error, while sharing removes $\sigma_u^2 p(1-p)$ from the difference error. The next result states when this pays off for a whole layer, in terms of three coefficients that Section 5.4 measures.

Proposition 4 (When sharing the scale pays)**.** *Let $a_c = E_c' - E_c$, $a_u = E_u' - E_u$ and $b = C' - C$ be the changes that sharing the scale causes, and $\Delta V = a_c + a_u - 2b$ the change of V. The guided error changes by*

$$\Delta(s) = a_c + 2s\,(a_c - b) + s^2\,\Delta V. \tag{6}$$

If $\Delta V \le 0$ and $b > a_c$, then $\Delta(s)$ decreases in s and is negative for all $s > s_0$, where $s_0 = 0$ if $a_c \le 0$ and $s_0 \le a_c/(2(b-a_c))$ otherwise; if $\Delta V > 0$, it is positive for all large s. The share of the guided error removed tends to $1 - V'/V$ as $s \to \infty$, where V' is the difference error after sharing.

Corollary 1 (The larger scale is the finest safe grid)**.** *With absolute-maximum scaling, a shared scale clips neither branch if and only if it is at least $\max(\sigma_c, \sigma_u)$. Under the uniform model, the per-branch errors and the difference error $\sigma|\eta| - \eta^2$ of every coordinate with $|\eta| \le \sigma$ increase with σ, so among shared scales that clip neither branch, $\bar{\sigma} = \max(\sigma_c, \sigma_u)$ is the optimal choice for these coordinates.*

Proposition 4 applies Proposition 1 before and after sharing, and Corollary 1 follows from part (a) of Proposition 3 (Appendix A.4). In our measurements, sharing mainly coarsens the unconditional branch, so a_c is small next to b and the condition $\Delta V \le 0$ is close to $a_u \le 2b$: the cross term must gain at least half of what that branch loses. A narrower shared range would lower σ further at the cost of clipping the largest coordinates, the trade that Appendix D.2 explores on a scalar stress test.

5 Experiments

Backbones. Seven image models, SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024), DiT on class-conditional ImageNet (Peebles & Xie, 2023), SD3.5 (Esser et al., 2024), PixArt-Σ (Chen et al., 2024), SANA (Xie et al., 2025) and Hunyuan-DiT (Li et al., 2024); one video model, Wan2.x (Team Wan et al., 2025); and three language models with two-branch guidance (Sanchez et al., 2024), Qwen3-8B (Yang et al., 2025), Llama-3.3-70B (Grattafiori et al., 2024) and Qwen2.5-72B (Yang et al., 2024). The guidance-distilled FLUX.1-dev (Black Forest Labs, 2024) is a single-branch control.
Quantization. Weights are quantized once, identically for both branches. The reference recipe is W4A8 with per-token dynamic activation scales; we also run W8A8, FP8 (E4M3), per-channel W4 and SmoothQuant+W8A8 (Xiao et al., 2023). Q-Diffusion, PTQD, TFMQ-DM, MixDQ, Q-DiT, PTQ4DiT and ViDiT-Q (Li et al., 2023a; He et al., 2023; Huang et al., 2024a; Zhao et al., 2024; Chen

Table 1: Main comparison at matched bits and $w = 7.5$, with 28 sampling steps for the image and video models; quantized rows use W4A8 except SVDQuant (W4A4). Per-backbone columns give the trajectory distance to the BF16 sampler; Agg. is a normalized score that combines it over the backbones of a row; FID and Pref. are means over the image models of a row, with Pref. the preference probability against the default (0.5 is parity); p99 is the 99th percentile of $|\delta|$ over the backbones of a row. Gray rows are references: the BF16 sampler, three couplings that make every quantizer stochastic, and an oracle with one grid for both branches. Among the others, bold marks the best and underline the second best, PairRound and its shared-randomness variant are shaded, and a dash marks a backbone that a method does not cover.

		Image							Video	Language			Summary			
	Coupling or method	SD1.5	SDXL	DiT	SD3.5	PixArt	SANA	Hunyuan	Wan2.x	Qwen3	Llama	Qwen2.5	Agg.↓	FID↓	Pref.↑	p99↓
	BF16 (reference)	0	0	0	0	0	0	0	0	0	0	0	0	12.60	0.728	0
Couplings	Det., shared per-tensor	0.090	0.187	0.113	0.080	0.113	0.085	0.164	0.096	0.123	0.122	0.117	0.483	17.02	0.627	0.123
Couplings	Default (per-branch)	0.202	0.361	0.271	0.193	0.229	0.196	0.313	0.214	0.246	0.258	0.267	0.973	19.16	0.500	0.180
Couplings	Stoch., independent	0.258	0.516	0.328	0.235	0.338	0.239	0.460	0.278	0.348	0.337	0.332	1.330	20.39	0.410	0.212
Couplings	Stoch., shared rand.	0.008	0.020	0.012	0.008	0.013	0.008	0.019	0.009	0.012	0.014	0.013	0.046	13.92	0.721	0.018
Couplings	Stoch., antithetic	0.483	1.011	0.645	0.457	0.657	0.464	0.904	0.566	0.681	0.691	0.684	2.545	24.23	0.171	0.304
Couplings	Oracle, one grid	0.008	0.019	0.013	0.009	0.011	0.006	0.019	0.008	0.012	0.014	0.013	0.044	14.01	0.722	0.020
Ours	Shared scale, per coord.	0.067	0.140	0.089	0.064	**0.087**	**0.063**	0.126	0.075	0.094	0.091	0.088	0.359	16.46	0.653	0.109
Ours	Shared scale, per channel	0.082	0.159	0.100	0.072	0.099	0.072	0.141	0.085	0.111	0.106	0.101	0.410	16.79	0.640	0.114
Ours	Shared scale, per tensor	0.092	0.180	0.116	0.081	0.114	0.083	0.164	0.096	0.117	0.118	0.113	0.459	17.00	0.629	0.129
Ours	Shared scale, output layer	0.178	0.310	0.233	0.168	0.187	0.161	0.271	0.191	0.207	0.223	0.236	0.851	18.72	0.535	0.165
Ours	Shared scale + rand.	**0.064**	**0.137**	**0.086**	**0.061**	0.089	**0.063**	**0.122**	**0.071**	**0.091**	**0.088**	**0.087**	**0.343**	**16.43**	**0.659**	**0.105**
Published	Q-Diffusion	0.180	0.315	–	–	–	–	–	–	–	–	–	0.881	18.22	0.540	0.177
Published	PTQD	0.168	0.304	–	–	–	–	–	–	–	–	–	0.824	18.21	0.551	0.169
Published	TFMQ-DM	0.158	0.278	–	–	–	–	–	–	–	–	–	0.784	18.02	0.561	0.162
Published	MixDQ	0.155	0.273	–	–	–	–	–	–	–	–	–	0.739	17.75	0.571	0.162
Published	Q-DiT	–	–	0.197	0.146	0.167	0.139	0.238	–	–	–	–	0.746	18.46	0.559	0.151
Published	PTQ4DiT	–	–	0.199	0.137	0.161	0.135	0.226	–	–	–	–	0.699	18.33	0.564	0.151
Published	ViDiT-Q	–	–	–	–	–	–	–	0.175	–	–	–	0.770	–	–	0.173
Published	SVDQuant (W4A4)	0.352	0.634	0.467	0.328	0.394	0.315	0.548	–	–	–	–	1.680	21.64	0.331	0.225

et al., 2025; Wu et al., 2024; Zhao et al., 2025) run at W4A8, and SVDQuant (Li et al., 2025) at W4A4. In the results, PairRound denotes its shared scale per coordinate, and a separate row adds shared randomness at the few quantizers that are stochastic. **Guidance.** We use $w \in \{1, 2, 3, 5, 7.5, 12\}$, that is $s \in \{0, 1, 2, 4, 6.5, 11\}$, with 4 to 100 steps; the reference setting is $w = 7.5$ with 28 steps, on SDXL at W4A8 when one backbone is studied. **Metrics.** The trajectory distance is the ℓ_2 distance between the states of the quantized and the BF16 sampler from the same noise, averaged over steps and samples; Agg. is a normalized score that combines it over the backbones of a row. We report FID (Heusel et al., 2017) against COCO 2014 validation images (Lin et al., 2014) (ImageNet (Russakovsky et al., 2015) for DiT), a PickScore preference probability against the default (Kirstain et al., 2023), VBench (Huang et al., 2024b), perplexity and a judged win rate for language, and p99, the 99th percentile of $|\delta|$. **Protocol.** All variants of a backbone share the prompts and seeds, 10k samples per image configuration, VBench prompts, and 805 AlpacaEval instructions (Li et al., 2023b) with 12 seeds each, so every comparison is paired; timing uses a single NVIDIA H100 GPU (Appendix B).

5.1 Main Results

In Table 1, the three stochastic rows test Proposition 2: they make every quantizer stochastic and change only how the branches share random numbers, and on every backbone, shared randomness is the lowest of the three and antithetic randomness the highest, 48 to 67 times apart, with the default between shared and independent randomness. These rows are references, since they drop the deterministic rounding of 98.9% of the surveyed quantizers, which PairRound keeps. Sharing the scale lowers the distance of the default by 60% to 68% on each backbone and Agg. from 0.973 to 0.359, and improves the mean FID by 2.7 and the preference by 0.153. Adding shared randomness at the few stochastic quantizers lowers Agg. further, to 0.343, the lowest of all non-reference rows.

The published methods, which target the per-branch errors, lower the per-backbone distance of the default by 11% to 31%, but each keeps per-branch scales and stays above the shared scale on every backbone it covers; SVDQuant, with four-bit activations, lies above the default. Across the settings of Figure 3, the couplings coincide at $s = 0$ and separate as s grows, on SDXL and on the three language models, where the shared scale ends 2.6 to 3.1 times lower than the default at $s = 11$, and

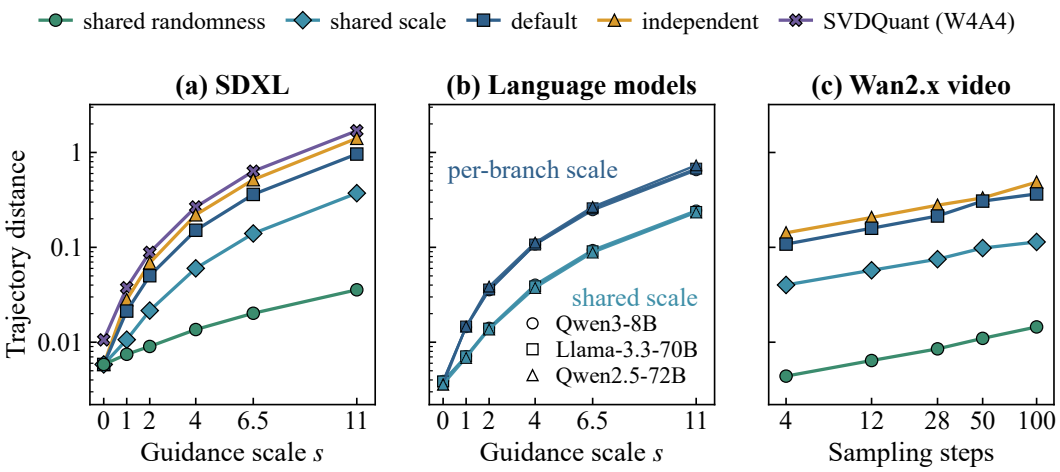


Figure 3: Trajectory distance across three settings. (a) SDXL against the guidance scale s for three couplings, the shared scale of PairRound and SVDQuant at W4A4. (b) Three language models with per-branch scales (the default) and with the shared scale of PairRound; marker shapes identify the models. (c) Wan2.x against the number of sampling steps at $s = 6.5$.

Table 2: Generality and composability. Left: SDXL under the shared scale, with the gain in trajectory distance over the default of the same setting; shaded rows are the reference setting. Right, top: eight published methods alone and with the shared scale underneath (†W4A4; ViDiT-Q has no FID). Right, bottom: trajectory distance of four stacks (Base), with their own fix (Host), with PairRound, and with both; Pred. adds the two single gains and Meas. is the joint gain.

Axis	Setting	Traj.↓	FID↓	Gain↑
Schedule	constant	0.140	18.80	0.221
	linear	0.066	17.05	0.103
	interval	0.052	16.55	0.079
	dynamic	0.089	17.53	0.144
Recipe	W8A8	0.026	15.90	0.041
	W4A8	0.140	18.80	0.221
	FP8	0.037	16.23	0.058
	per-channel W4	0.120	18.24	0.188
	SmoothQuant+W8A8	0.024	15.57	0.034
Steps	4	0.066	17.38	0.103
	12	0.106	18.27	0.158
	28	0.140	18.80	0.221
	50	0.177	19.53	0.237
	100	0.233	20.32	0.356

	Agg. trajectory distance			FID	
Method	alone	+PairRound	Red. (%)	alone	+PairRound
Q-Diffusion	0.881	0.335	62.0	18.22	15.59
PTQD	0.824	0.317	61.5	18.21	15.69
TFMQ-DM	0.784	0.300	61.7	18.02	15.49
MixDQ	0.739	0.291	60.6	17.75	15.40
Q-DiT	0.746	0.281	62.3	18.46	16.15
PTQ4DiT	0.699	0.276	60.5	18.33	16.31
ViDiT-Q	0.770	0.289	62.5	–	–
SVDQuant†	1.680	0.252	85.0	21.64	17.86

Stack	Base	Host	PairRound	Both	Pred.	Meas.
diffusers	0.363	0.281	0.140	0.112	0.305	0.251
TensorRT	0.277	0.211	0.088	0.071	0.255	0.206
vLLM	0.250	0.198	0.092	0.075	0.210	0.175
Vendor stack	0.199	0.149	0.059	0.049	0.190	0.150

they keep their order on the video model at every step count, with a gap that widens with the number of steps. The gain shows in every quality metric: FID drops by 2.3 to 3.5 on the seven image models, VBench rises by 2.1, perplexity drops by 0.26 to 0.30, and the preference or win rate against the default grows by 0.11 to 0.21, so that all 33 pairs of backbone and metric improve (Appendix C.2).

5.2 Generality and Composability

The gain of the shared scale persists in every setting we vary (Table 2, left). It is largest with four-bit weights and smallest at eight bits and with FP8, while the relative reduction stays between 57% and 62% in every setting. The gain grows with the number of steps, from 0.103 at 4 steps to 0.356 at 100, and schedules that apply strong guidance on fewer steps leave smaller gains. Under each of the eight published methods, the shared scale applied underneath removes 60.5% to 62.5% of the aggregate distance at W4A8 and 85% under SVDQuant at W4A4, and it lowers FID by 2.0 to 3.8 for the seven image methods. The gains combine, consistent with the methods targeting E_c and E_u and PairRound targeting V in Eq. (2). On four stacks that ship their own accuracy fix, PairRound gains 2.7 to 3.0 times as much as the fix, and the two combined reach 79% to 83% of the sum of their separate gains.

Table 3: Ablations. Left: SDXL reference setting, with synchronization (Sync) in μs per step and branch and the gain in trajectory distance over the default. Right: trajectory distance on three backbones, their mean, and Kept, the share of the gain of PairRound that a variant keeps, (default − variant)/(default − PairRound) in percent. Shaded rows are PairRound and its shared-randomness variant, and a dash marks a value not reported.

Variant	Traj.↓	FID↓	Pref.↑	p99↓	Sync	Gain↑
Default (per-branch)	0.361	22.30	0.500	0.198	0	0.000
PairRound	0.140	18.80	0.712	0.126	11.7	0.221
Shared scale + rand.	0.137	18.80	0.710	0.121	14.0	0.224
Output layer only	0.310	21.91	0.553	0.180	4.0	0.051
First third of steps	0.292	21.36	0.566	0.177	5.0	0.069
Stoch., independent	0.516	24.21	0.348	0.234	0	−0.155
Stoch., shared rand.	0.020	15.60	0.801	0.023	7.0	0.341
Default at W6A8	0.155	19.35	–	–	0	0.206
Default with 50 steps	0.414	23.56	–	–	0	−0.053
Det., shared per-tensor	0.187	19.69	0.659	0.139	0	0.174
One-branch correction	0.294	21.48	0.564	–	6.0	0.067

Variant	SDXL	DiT	SD3.5	Mean↓	Kept↑
PairRound (max)	0.140	0.089	0.064	0.097	100.0
Shared scale + rand.	0.137	0.086	0.061	0.094	101.7
Mean instead of max	0.147	0.094	0.064	0.102	97.6
Per-channel sharing	0.159	0.100	0.072	0.110	92.7
Per-tensor sharing	0.180	0.116	0.081	0.126	84.1
Output layer only	0.310	0.233	0.168	0.237	21.5
Replayed cached scales	0.143	0.090	0.064	0.099	99.1
Stoch., shared rand.	0.020	0.012	0.008	0.013	147.4
Stoch., independent	0.516	0.328	0.235	0.360	−47.6
Stoch., antithetic	1.011	0.645	0.457	0.704	−241.5

5.3 Ablation Studies

Table 3 changes one design choice at a time; Appendix D gives every sweep. **Sharing the scale supplies almost all of the gain.** Adding shared randomness at the stochastic quantizers keeps 101.7% of the gain of the shared scale alone, which supplies 98.3% of the joint gain, as few quantizers of the recipe round stochastically. **Each rule matters where its type of rounding is used.** When all quantizers round stochastically, the coupling of the random numbers sets the outcome: shared randomness gives a mean of 0.013, independent 0.360 and antithetic 0.704, in the order of Proposition 2, and antithetic random numbers roughly double the distance of the independent ones.

The maximum beats the mean. In deployed networks the mean of the two scales, which clips the larger branch, keeps 97.6% of the gain. **Finer sharing gains more, and coarser sharing costs less.** Sharing per channel keeps 92.7% of the gain and per tensor 84.1%, with 8 and 5 μs of synchronization instead of 11.7 μs. **The whole network and trajectory are needed.** Sharing at the output layer alone keeps about a fifth of the gain, sharing on the first third of the steps about three tenths, and sharing on three quarters of the layers about three quarters (Appendices D.1 and D.6), each with a smaller cost.

Cached scales keep the gain. Replaying shared scales cached from a first run keeps 99.1% of the gain, so the gain does not depend on computing the shared scale on the fly. **Other fixes fall short.** Two more weight bits, W6A8, bring the default to 0.155, above the 0.140 of PairRound at 1.5 times the weight memory; more sampling steps make the default worse; a consistent per-tensor scale gives 0.187; and a correction within a single branch reaches 0.294, since it does not see the other branch.

5.4 Mechanism Analysis

In Table 4, changing only the coupling of the random numbers moves E_c and E_u by at most 0.00016, while ρ is 0.93 to 0.97 under shared and below 0.1 under independent randomness and V changes by a factor of 13 to 20, in the order of Proposition 2. Sharing the scale raises the error of the unconditional branch, the coarsened one, by 39% to 50% but lifts ρ to about 0.78, so V falls by 61% to 67%. On all seven image backbones a_u/b lies between 0.47 and 0.69, well within the condition $a_u \leq 2b$ of Section 4.2, and on SDXL, DiT and SD3.5, where it is also measured, a_c lies between -0.00002 and 0.0003, well below b. With these coefficients, measured at $s = 6.5$, Proposition 4 predicts a gain on these three backbones at every scale in use. The measured change $\Delta(s)$ is negative at every scale, the share removed at $s = 11$ (60.4%, 67.0% and 67.4% on SDXL, DiT and SD3.5) is close to the predicted limits (61.3%, 64.8% and 67.4%), while the gain grows nearly in proportion to $s(1 + s)$.

Figure 4(a) splits the trajectory distance by source. In four single-source runs, each adding one recorded error source to the BF16 sampler, the branch difference accounts for 85% to 87% of the summed distance under the default; recorded under the shared scale, this run falls by 71% to 78% while the other three sources barely change. Through the network, the correlation of the two branch errors decays by a factor of about 0.95 per layer under shared randomness, the shared scale and the default, and the order of the couplings holds at every layer, in line with the first-order model of

Table 4: Interventions on the mechanism at the first layer. Left: per-branch errors, cross term, correlation $\rho = C/\sqrt{E_c E_u}$ and difference error V for four couplings on three backbones at $s = 6.5$ (errors in units of 10^{-3}); shaded rows are PairRound. Right, top: a_u and b of Proposition 4 (10^{-3}) and their ratio on seven image backbones. Right, bottom: the measured change $\Delta(s)$ of the guided error on SDXL and the share it removes (Appendix C.1).

Coupling	Model	E_c	E_u	C	ρ	V
Det., per-branch	SDXL	4.76	3.64	1.24	0.298	5.92
	DiT	4.02	3.22	0.69	0.192	5.86
	SD3.5	2.68	2.10	0.49	0.207	3.80
Stoch., independent	SDXL	4.76	3.65	0.38	0.091	7.65
	DiT	4.06	3.26	0.14	0.038	7.04
	SD3.5	2.57	2.03	0.11	0.048	4.38
Stoch., shared rand.	SDXL	4.92	3.68	4.11	0.966	0.38
	DiT	3.95	3.13	3.26	0.927	0.56
	SD3.5	2.68	2.15	2.26	0.942	0.31
Shared scale	SDXL	4.74	5.25	3.85	0.772	2.29
	DiT	4.32	4.82	3.54	0.776	2.06
	SD3.5	2.73	2.91	2.20	0.781	1.24

Model	a_u	b	a_u/b
SD1.5	0.97	2.02	0.48
SDXL	1.61	2.61	0.62
DiT	1.60	2.85	0.56
SD3.5	0.81	1.71	0.47
PixArt-Σ	1.45	2.11	0.69
SANA	0.77	1.50	0.51
Hunyuan-DiT	1.74	2.53	0.69

s	1	2	4	6.5	11
$\Delta(s)$	−0.008	−0.023	−0.077	−0.173	−0.465
Saved (%)	48.7	56.8	60.9	59.2	60.4

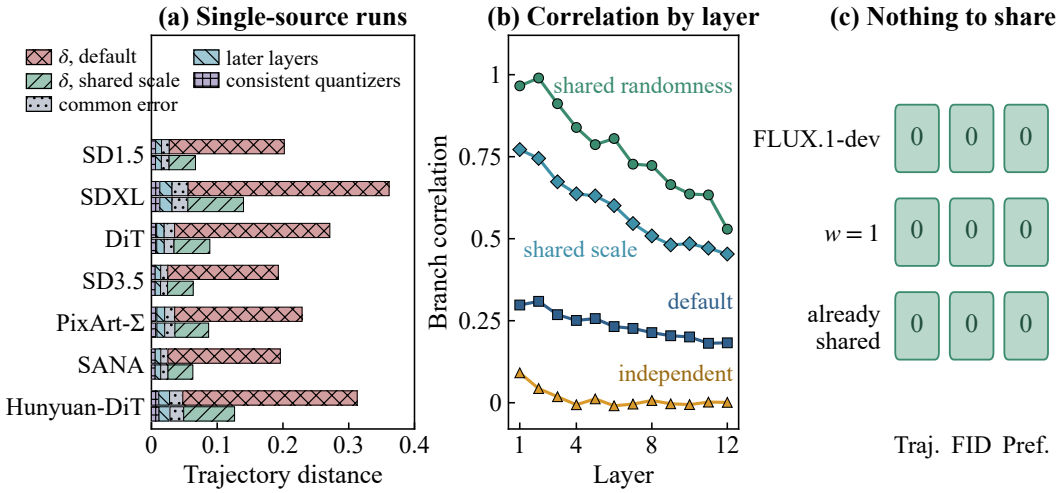


Figure 4: Error sources. (a) Trajectory distance of four single-source runs, stacked: errors of the consistent quantizers and later layers, the common error $(\boldsymbol{e}_c + \boldsymbol{e}_u)/2$, and $\boldsymbol{\delta}$, recorded under the default (upper bar) or the shared scale (lower bar). (b) Correlation of the branch errors after each layer of SDXL. (c) Gain with one branch (FLUX.1-dev or $w = 1$) or an already shared scale.

Appendix A.7. With only one branch, as in FLUX.1-dev or at $w = 1$, or with an already shared scale, PairRound leaves all three metrics unchanged (Figure 4c). Per sample, the shared scale does not raise the distance on 81% to 82% of the paired prompts of SDXL, DiT and SD3.5 (Appendix D.6), and under every coupling the worst tenth of the prompts carries 41% to 46% of the squared error, so a better coupling lowers the error without concentrating it on a small set of prompts (Appendix E.1).

6 Conclusion

Classifier-free guidance multiplies the mean squared difference of the quantization errors of its two branches by $s(1 + s)$, a term that per-branch metrics cannot see and that deployed stacks set mostly by per-branch scales. PairRound shares the scale and the random numbers across the branches, so that their codes differ only where their values differ, and our analysis states when this pays. On eleven backbones its shared scale lowers the trajectory distance by 60% to 68% at matched bits, composes with published methods, and costs 12 μs per step and branch. Any computation that combines quantized evaluations with weights of opposite sign has this cross term, so its quantizers should be judged on the combination. **Limitations.** Our analysis of nearest rounding relies on the uniform model and covers one sampling step, while Section 5 measures whole trajectories; guidance with more than two branches follows the same identity (Appendix A.5) and is a natural next target.

Reproducibility Statement

Section 5 and Appendix B describe the backbones, quantization recipes, guidance scales, step counts, metrics and the paired protocol. Appendix A gives the proofs of all propositions, Appendix B.1 the survey of the six stacks, and Appendix B.2 the pairing rule, the implementation checks and Algorithm 1. Appendices C to F and Appendix I list the numbers behind all tables and figures, including the cost accounting. We will release our patches, survey tables and measurement scripts.

AI Use Statement

Generative AI tools were used only to edit the wording of text written by the authors, to search for related work and check bibliographic details, and to write code for plots, tables and typesetting. They were not used for any task that the ICLR 2027 AI policy requires authors to disclose: developing theory, formulating or proving mathematical claims, proposing or refining hypotheses, designing, implementing or running the method or experiments, generating, cleaning or analyzing data, translating text, or interpreting results. The models in our experiments are objects of study, not writing aids. The authors reviewed all AI-assisted work and take full responsibility for the content of this paper.

References

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

W. R. Bennett. Spectra of quantized signals. *Bell System Technical Journal*, 27(3):446–472, 1948. doi: 10.1002/j.1538-7305.1948.tb01340.x.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024. GitHub repository.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Lei Chen, Yuan Meng, Chen Tang, Xinzhu Ma, Jingyan Jiang, Xin Wang, Zhi Wang, and Wenwu Zhu. Q-DiT: Accurate post-training quantization for diffusion transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

comfyanonymous. ComfyUI: The most powerful and modular diffusion model GUI, API and backend with a graph/nodes interface. https://github.com/Comfy-Org/ComfyUI, 2023. GitHub repository.

Michael P. Connolly, Nicholas J. Higham, and Theo Mary. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing*, 43(1):A566–A585, 2021. doi: 10.1137/20M1334796.

Matteo Croci, Massimiliano Fasi, Nicholas J. Higham, Theo Mary, and Mantas Mikaitis. Stochastic rounding: Implementation, error analysis and applications. *Royal Society Open Science*, 9(3): 211631, 2022. doi: 10.1098/rsos.211631.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 30318–30332, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

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 *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon, Section A, Série 3*, 14:53–77, 1951.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Robert M. Gray and Thomas G. Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, 1993. doi: 10.1109/18.256489.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1737–1746, 2015.

J. M. Hammersley and K. W. Morton. A new Monte Carlo technique: Antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(3):449–475, 1956. doi: 10.1017/S0305004100031455.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. PTQD: Accurate post-training quantization for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. EfficientDM: Efficient quantization-aware fine-tuning of low-bit diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI. arXiv preprint arXiv:2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Wassily Hoeffding. Maßstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, 5(3):181–233, 1940.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. TFMQ-DM: Temporal feature maintenance quantization for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024b.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

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 (NeurIPS)*, volume 36, 2023.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 7, pp. 231–238, 1995.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. SVDQuant: Absorbing outliers by low-rank component for 4-bit diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-Diffusion: Quantizing diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17535–17545, 2023a.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023b. GitHub repository.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained Chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, 2nd edition, 2006. doi: 10.1007/0-387-28678-0.

Alexander 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 *International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804, 2022.

NVIDIA. NVIDIA TensorRT. https://developer.nvidia.com/tensorrt, 2024. High-performance deep learning inference SDK.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

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 (ICLR)*, 2024.

Lawrence G. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962. doi: 10.1109/TIT.1962.1057702.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

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

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, and Stella Biderman. Stay on topic with classifier-free guidance. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43197–43234, 2024.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Yang Sui, Yanyu Li, Anil Kag, Yerlan Idelbayev, Junli Cao, Ju Hu, Dhritiman Sagar, Bo Yuan, Sergey Tulyakov, and Jian Ren. BitsFusion: 1.99 bits weight quantization of diffusion model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. GitHub repository.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=SUMtDJqicd.

Bernard Widrow and István Kollár. *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-88671-0.

Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. PTQ4DiT: Post-training quantization for diffusion transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

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.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025.

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 (NeurIPS)*, volume 36, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27168–27183, 2022.

Tianchen Zhao, Xuefei Ning, Tongcheng Fang, Enshu Liu, Guyue Huang, Zinan Lin, Shengen Yan, Guohao Dai, and Yu Wang. MixDQ: Memory-efficient few-step text-to-image diffusion models with metric-decoupled mixed precision quantization. In *European Conference on Computer Vision (ECCV)*, 2024.

Tianchen Zhao, Tongcheng Fang, Haofeng Huang, Rui Wan, Widyadewi Soedarmadji, Enshu Liu, Shiyao Li, Zinan Lin, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. ViDiT-Q: Efficient and accurate quantization of diffusion transformers for image and video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

CONTENTS OF THE APPENDIX

A Proofs and Derivations **15**
- A.1 Proof of Proposition 1. . . . 16
- A.2 Proof of Proposition 2. . . . 17
- A.3 Proof of Proposition 3. . . . 17
- A.4 Proof of Proposition 4 and Corollary 1. . . . 18
- A.5 Guidance with More Than Two Branches 19
- A.6 One Quantizer in Numbers. . . . 20
- A.7 Propagation Through the Remaining Layers 20

B Experimental Details **21**
- B.1 Survey of Inference Stacks. . . . 21
- B.2 Implementation of PairRound 22
- B.3 Backbones, Data and Metrics 22
- B.4 Quantization Recipes and Published Methods 23

C Additional Results **23**
- C.1 Growth with the Guidance Scale 23
- C.2 Results on Every Backbone and Metric 25
- C.3 Sweeps on SDXL, the Language Models and the Video Model 27
- C.4 Generality on Every Axis. . . . 27
- C.5 Published Methods, Coverage and Paired Differences. . . . 28
- C.6 Four Metrics 28

D Additional Ablations **31**
- D.1 Fraction of the Layers Converted 31
- D.2 The Maximum Against a Narrower Shared Range 31
- D.3 Pieces of the Sampling Trajectory 32
- D.4 Interventions and Granularity 32
- D.5 One Third of the Network 33
- D.6 Observables Behind the Ablations 33

E Mechanism Details **34**
- E.1 Per-Sample Behavior 34
- E.2 The Cross Term on Every Backbone 34
- E.3 Cross Term of Six Couplings 36
- E.4 Distribution of the Branch Difference 36
- E.5 Single-Source Runs 36
- E.6 Branch Correlation by Layer. . . . 36
- E.7 Theory Against Measurement 36

F Cost **38**

G Additional Related Work **38**

H Notation **38**

I Numbers Behind the Figures **39**

A PROOFS AND DERIVATIONS

Throughout, $\boldsymbol{e}_c, \boldsymbol{e}_u \in \mathbb{R}^D$ are the quantization errors of the conditional and the unconditional branch at the point where they are compared, $E_c = \mathbb{E}\|\boldsymbol{e}_c\|^2$ and $E_u = \mathbb{E}\|\boldsymbol{e}_u\|^2$ are the per-branch errors,

$C = \mathbb{E}\langle \boldsymbol{e}_c, \boldsymbol{e}_u \rangle$ is their cross term, $\boldsymbol{\delta} = \boldsymbol{e}_c - \boldsymbol{e}_u$ is the branch difference with $V = \mathbb{E}\|\boldsymbol{\delta}\|^2$, and $\bar{\boldsymbol{e}} = (\boldsymbol{e}_c + \boldsymbol{e}_u)/2$ is the common error. The guidance scale is $s = w - 1$, expectations run over prompts, noise seeds and the rounding randomness, and all norms are Euclidean. Sections A.1 to A.4 prove the statements of the main text, Sections A.5 and A.7 add two results used in the discussion, on multiple branches and on later layers, and Section A.6 works through one quantizer with numbers.

A.1 Proof of Proposition 1

We derive the decomposition in three equivalent ways, each giving a different view of the guided error, and then its range when the per-branch errors are fixed. The first way expands the square and regroups the terms:

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 &= \mathbb{E}\|(1+s)\boldsymbol{e}_c - s\boldsymbol{e}_u\|^2 = (1+s)^2 E_c + s^2 E_u - 2s(1+s)\, C \\
&= E_c + (2s+s^2)\, E_c + s^2 E_u - 2s(1+s)\, C \\
&= E_c + s\,(E_c - E_u) + (s+s^2)\, E_c + (s+s^2)\, E_u - 2s(1+s)\, C \\
&= E_c + s\,(E_c - E_u) + s(1+s)\,\big(E_c + E_u - 2C\big) \\
&= E_c + s\,(E_c - E_u) + s(1+s)\, V.
\end{aligned} \tag{7}$$

The first line expands the square. The second writes $(1+s)^2 = 1 + 2s + s^2$. The third adds and subtracts sE_u, the fourth collects the terms that carry $s(1+s)$, and the last one uses $V = E_c + E_u - 2C$.

The second form writes the guided error around the conditional error, $\hat{\boldsymbol{e}} = (1+s)\boldsymbol{e}_c - s\boldsymbol{e}_u = \boldsymbol{e}_c + s\boldsymbol{\delta}$, so that

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 &= \mathbb{E}\|\boldsymbol{e}_c\|^2 + 2s\,\mathbb{E}\langle \boldsymbol{e}_c, \boldsymbol{\delta}\rangle + s^2\,\mathbb{E}\|\boldsymbol{\delta}\|^2 = E_c + 2s\,(E_c - C) + s^2 V, \\
2(E_c - C) &= (E_c - E_u) + (E_c + E_u - 2C) = (E_c - E_u) + V, \\
\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 &= E_c + s\,(E_c - E_u) + sV + s^2 V = E_c + s\,(E_c - E_u) + s(1+s)\, V.
\end{aligned} \tag{8}$$

Here the guidance adds s times the branch difference to the conditional error, and the cross term between the two turns $s^2 V$ into $s(1+s)V$. The third form splits the errors into their common part and their difference, $\boldsymbol{e}_c = \bar{\boldsymbol{e}} + \frac{1}{2}\boldsymbol{\delta}$ and $\boldsymbol{e}_u = \bar{\boldsymbol{e}} - \frac{1}{2}\boldsymbol{\delta}$, so that $\hat{\boldsymbol{e}} = \bar{\boldsymbol{e}} + (s + \frac{1}{2})\boldsymbol{\delta}$ and

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 &= \mathbb{E}\|\bar{\boldsymbol{e}}\|^2 + (2s+1)\,\mathbb{E}\langle \bar{\boldsymbol{e}}, \boldsymbol{\delta}\rangle + \big(s + \tfrac{1}{2}\big)^2 V \\
&= \tfrac{1}{4}\,(E_c + E_u + 2C) + \big(s + \tfrac{1}{2}\big)(E_c - E_u) + \big(s(1+s) + \tfrac{1}{4}\big)\, V \\
&= \tfrac{1}{2}\,(E_c + E_u) + \big(s + \tfrac{1}{2}\big)(E_c - E_u) + s(1+s)\, V \\
&= E_c + s\,(E_c - E_u) + s(1+s)\, V,
\end{aligned} \tag{9}$$

where the second line uses $\mathbb{E}\|\bar{\boldsymbol{e}}\|^2 = \frac{1}{4}(E_c + E_u + 2C)$ and $\mathbb{E}\langle \bar{\boldsymbol{e}}, \boldsymbol{\delta}\rangle = \frac{1}{2}(E_c - E_u)$, and the third uses $\frac{1}{4}(E_c + E_u + 2C) + \frac{1}{4}V = \frac{1}{2}(E_c + E_u)$. In this form the common error enters once, and the difference is scaled by $s + \frac{1}{2}$; at fixed per-branch errors a smaller difference lowers the guided error.

For the range, $\langle \mathbf{X}, \mathbf{Y}\rangle_{\mathbb{E}} = \mathbb{E}\langle \mathbf{X}, \mathbf{Y}\rangle$ is an inner product on square-integrable random vectors, so the Cauchy–Schwarz inequality gives $|C| \le \sqrt{E_c E_u}$ and

$$V = E_c + E_u - 2C \in \Big[\big(\sqrt{E_c} - \sqrt{E_u}\big)^2,\ \big(\sqrt{E_c} + \sqrt{E_u}\big)^2\Big], \tag{10}$$

an interval of length $4\sqrt{E_c E_u}$. At fixed E_c and E_u the guided error is affine in V with slope $s(1+s)$, so it fills an interval of length $4s(1+s)\sqrt{E_c E_u}$. Every point of the interval is reached. Take any $\rho \in [-1, 1]$ and a random vector $\mathbf{z}$ that is independent of $\boldsymbol{e}_c$, has mean zero and $\mathbb{E}\|\mathbf{z}\|^2 = E_c$, and set $\boldsymbol{e}_u = \sqrt{E_u/E_c}\,\big(\rho\,\boldsymbol{e}_c + \sqrt{1-\rho^2}\,\mathbf{z}\big)$; then $\mathbb{E}\|\boldsymbol{e}_u\|^2 = E_u$ and $C = \rho\sqrt{E_c E_u}$, and $\rho = \pm 1$ gives the two ends. □

Remark. The identity holds for any pair of errors and needs no assumption on their distribution. Per-branch metrics fix E_c and E_u, and with them only the first two terms and the width of the interval in which the guided error can lie. Where in this interval a quantizer lands is decided by C, which is invisible to both branches taken one at a time; the next three proofs show how the rounding sets C.

A.2 Proof of Proposition 2

At a stochastic quantizer the two branch errors depend on the random numbers only through their joint distribution. The classical bounds on this distribution give the largest and the smallest cross term that the marginals allow. The error of a value with fractional part f is $e = \sigma(\mathbf{1}[u < f] - f)$, so $\mathbb{E}[e] = \sigma(\Pr[u < f] - f) = 0$ for every value and every coupling. For two branches on grids of scales σ_c and σ_u,

$$\begin{aligned}
\mathbb{E}[e_c e_u] &= \sigma_c \sigma_u \, \mathbb{E}\big[(\mathbf{1}[u_c < f_c] - f_c)(\mathbf{1}[u_u < f_u] - f_u)\big] \\
&= \sigma_c \sigma_u \big(\Pr[u_c < f_c,\, u_u < f_u] - f_c \Pr[u_u < f_u] - f_u \Pr[u_c < f_c] + f_c f_u\big) \\
&= \sigma_c \sigma_u \big(K(f_c, f_u) - f_c f_u\big), \qquad K(a,b) = \Pr[u_c < a,\, u_u < b], \\
\max(f_c + f_u - 1,\, 0) \le K(f_c, f_u) &\le \min(f_c, f_u) \qquad \text{for every joint law of } (u_c, u_u), \\
u_u = u_c :\ K &= \min(f_c, f_u), \qquad u_u = 1 - u_c :\ K = \max(f_c + f_u - 1, 0), \\
u_u \perp u_c :\ K &= f_c f_u.
\end{aligned} \tag{11}$$

The first line writes the product of the two errors and the second expands it. The third uses $\Pr[u < f] = f$ for each uniform number, so two terms cancel, and names the joint distribution function K of the pair, its copula (Nelsen, 2006). The fourth line is the Fréchet–Hoeffding inequality (Fréchet, 1951; Hoeffding, 1940): the upper bound holds because the joint event lies in each marginal event, and the lower bound because $\Pr[A \cap B] \ge \Pr[A] + \Pr[B] - 1$. The last line evaluates K for the three couplings: with $u_u = u_c$ both events hold if and only if $u_c < \min(f_c, f_u)$, with $u_u = 1 - u_c$ if and only if $1 - f_u < u_c < f_c$, and independence factorizes. The marginals are untouched, since each branch still rounds up with probability f. For equal values on one grid, $f_c = f_u$ and a shared u give $\mathbf{1}[u_c < f] = \mathbf{1}[u_u < f]$ and hence $e_c = e_u$ on every draw. Summing over the coordinates of one quantizer gives its cross term C. □

Remark. This is the argument behind common random numbers (Glasserman & Yao, 1992): both errors are monotone in their random number, so their covariance is largest when the numbers coincide and smallest when they are antithetic (Hammersley & Morton, 1956). By Proposition 1, the shared coupling therefore minimizes V, and with it the guided error, among all couplings with the same marginals. The statement compares couplings of stochastic rounding only. Deterministic rounding makes the error a fixed function of the value, so the proposition does not place it among these couplings; Table 1 finds it between shared and independent randomness on each of our backbones.

A.3 Proof of Proposition 3

Write values in units of the grid, $y = x/\sigma$, and let the two branches hold y_c and y_u with $\tau = |y_c - y_u| = |\eta|/\sigma \le 1$. The rounding error of the code k is $e = \sigma k - x$, so the difference error is $\delta = \sigma(k_c - k_u) - \eta$ whenever the two branches share the grid, and the parts below compute $\mathbb{E}[\delta^2]$.

Part (a), nearest rounding. Nearest rounding changes its code at half-integers. The interval between y_u and y_c has length $\tau \le 1$, so it contains at most one half-integer, and under the uniform model, in which the position of y_u on the grid is uniform and independent of η, it contains one with probability τ; the codes then differ by one step in the direction of η. Hence, with $p = \tau$,

$$\begin{aligned}
\delta &= -\eta, & \delta^2 &= \sigma^2 p^2 \qquad \text{(same code, probability } 1 - p), \\
\delta &= \sigma \operatorname{sgn}(\eta) - \eta, & \delta^2 &= \sigma^2 (1 - p)^2 \quad \text{(codes differ, probability } p), \\
\mathbb{E}[\delta^2] &= \sigma^2 \big[(1 - p)\, p^2 + p\, (1 - p)^2\big] = \sigma^2 p\, (1 - p) \big[p + (1 - p)\big] \\
&= \sigma^2 p\, (1 - p).
\end{aligned} \tag{12}$$

If $\eta = 0$, then $p = 0$ and $\delta = 0$, with or without the uniform model, and equal values get one code.

Part (a), stochastic rounding with a shared random number. Stochastic rounding of y with the number u returns $\lfloor y \rfloor + \mathbf{1}[u < y - \lfloor y \rfloor]$, which equals $\lfloor y + v \rfloor$ with $v = 1 - u$, uniform on $[0, 1)$, up to a null event. The two branches therefore floor their values after one common random shift. For $y_c \ge y_u$ the codes differ, by exactly one, if and only if an integer n satisfies $y_u + v < n \le y_c + v$, that is,

$v \in [\, n - y_c,\ n - y_u)$ for some n:

$$\begin{aligned} \Pr[k_c \neq k_u] &= \Pr\Big[v \in \textstyle\bigcup_{n\in\mathbb{Z}} [\, n - y_c,\ n - y_u)\Big] \\ &= \big|[\, n^\star - y_c,\ n^\star - y_u) \cap [0,1)\big| + \big|[\, n^\star + 1 - y_c,\ n^\star + 1 - y_u) \cap [0,1)\big| \\ &= y_c - y_u = \tau \qquad (0 \le \tau \le 1), \\ \mathbb{E}[\delta^2] &= (1-\tau)\,\sigma^2\tau^2 + \tau\,\sigma^2(1-\tau)^2 = \sigma^2\tau\,(1-\tau), \end{aligned} \tag{13}$$

where $n^\star$ is the integer with $n^\star - y_c \le 0 < n^\star + 1 - y_c$. The intervals have length τ and are spaced by one, so their union meets $[0,1)$ in a set of length τ, which gives the first line; the second line then repeats the case analysis of Eq. (12). The result thus holds for every pair of values, because the shared random shift plays the same part that the uniform model plays for nearest rounding in the first case.

Part (b), two grids. Let both branches hold $x > 0$ (negative values are symmetric) with $\sigma_c = (1+\gamma)\sigma_u$, and put $y = x/\sigma_u$. The codes are $k_u = \mathrm{rnd}(y)$ and $k_c = \mathrm{rnd}(y/(1+\gamma))$. The interval $(y/(1+\gamma),\ y]$ has length $y\gamma/(1+\gamma)$; while this is below one it holds at most one half-integer, and under the uniform model it holds one with probability $p = y\gamma/(1+\gamma)$, in which case $k_c = k_u - 1$. Given the code k_u,

$$\begin{aligned} & & \delta &= (\sigma_c k_c - x) - (\sigma_u k_u - x) = \sigma_u\big((1+\gamma)\,k_c - k_u\big), \\ &\text{same code } k: & \delta &= \sigma_u\,\gamma\,k, \\ &\text{codes differ}: & \delta &= -\sigma_u\big(1 - \gamma\,(k_u - 1)\big), \\ & & \mathbb{E}[\delta^2 \mid k_u] &= \sigma_u^2\big[(1-p)\,\gamma^2 k_u^2 + p\,(1 - \gamma k_u + \gamma)^2\big] \\ & & &= \sigma_u^2\big[(1-p)\,p^2 + p\,(1-p)^2\big] + R = \sigma_u^2\,p\,(1-p) + R, \\ & & |R| &= \sigma_u^2\,O\big(\gamma\,(p+\gamma)\big). \end{aligned} \tag{14}$$

The first line writes the difference error, in which x cancels, and the second evaluates it in the two cases. For the third, $|k_u - y| \le \frac{1}{2}$ gives $\gamma k_u - p = \gamma(k_u - y) + \gamma^2 y/(1+\gamma) = O(\gamma)$ and $\gamma k_u + p = O(p+\gamma)$, so replacing γk_u by p changes the first term by $(1-p)(\gamma k_u - p)(\gamma k_u + p) = O(\gamma(p+\gamma))$ and the second by $p\,O(\gamma)$, which is the remainder R; it is small next to $\sigma_u^2 p\,(1-p)$ when $\gamma \ll \min(p, 1-p)$. □

Remark. All three cases give the variance of one Bernoulli event, that the two codes differ, in units of the squared grid step. What changes is where the probability p comes from. On one grid, with nearest rounding or a shared random number, it is $|\eta|/\sigma$ and vanishes when the branches agree. On two grids it is $|x|\gamma/((1+\gamma)\sigma_u)$ and does not vanish when they agree, and on one grid with independent random numbers two equal values differ with probability $2f(1-f)$, which gives $\mathbb{E}[\delta^2] = 2\sigma^2 f(1-f)$. Both rules of PairRound remove both sources and keep only the part of p that the values themselves create.

A.4 Proof of Proposition 4 and Corollary 1

Apply the quadratic form of Eq. (7) before and after sharing and subtract, writing $E_c' = E_c + a_c$, $E_u' = E_u + a_u$ and $C' = C + b$:

$$\begin{aligned} \Delta(s) &= \big[(1+s)^2 E_c' + s^2 E_u' - 2s(1+s)\,C'\big] \\ &\quad - \big[(1+s)^2 E_c + s^2 E_u - 2s(1+s)\,C\big] \\ &= (1+s)^2 a_c + s^2 a_u - 2s(1+s)\,b \\ &= a_c + 2s\,(a_c - b) + s^2\,(a_c + a_u - 2b) = a_c + 2s\,(a_c - b) + s^2\,\Delta V \\ &= a_c + s\,(a_c - a_u) + s(1+s)\,\Delta V, \\ \Delta'(s) &= 2\,(a_c - b) + 2s\,\Delta V. \end{aligned} \tag{15}$$

The second equality subtracts the two brackets, the third collects powers of s, the fourth reads the same quadratic through the decomposition, and the last line is its slope. If $\Delta V \le 0$ and $b > a_c$, the slope is negative for every $s \ge 0$, so Δ decreases on $[0,\infty)$. When $a_c \le 0$, $\Delta(0) = a_c \le 0$ and therefore $\Delta(s) < 0$ for all $s > 0$. When $a_c > 0$, the point $s' = a_c/(2(b - a_c))$ gives $\Delta(s') = a_c - a_c + s'^2 \Delta V \le 0$, so the single root s_0 lies in $(0, s']$. If $\Delta V > 0$, the leading coefficient is positive and $\Delta(s) \to +\infty$. Finally, with $G(s) = E_c + s(E_c - E_u) + s(1+s)V$ the guided error

before sharing,

$$\frac{-\Delta(s)}{G(s)} = \frac{-a_c - s\,(a_c - a_u) - s(1+s)\,\Delta V}{E_c + s\,(E_c - E_u) + s(1+s)\,V} \xrightarrow[s\to\infty]{} -\frac{\Delta V}{V} = 1 - \frac{V'}{V}, \tag{16}$$

since numerator and denominator are both quadratic in s, with leading coefficients $-\Delta V$ and V. □

For Corollary 1, absolute-maximum scaling sets $\sigma_b = \max_i |x_{b,i}|/q$ over a scaling group, where q is the largest code. A shared scale σ clips neither branch if and only if $\sigma q \geq \max_i |x_{b,i}|$ for both branches, that is, $\sigma \geq \max(\sigma_c, \sigma_u)$. On a shared grid the uniform model gives each branch the error $\sigma^2/12$ per coordinate, so the linear term of the decomposition vanishes, and part (a) of Proposition 3 gives $\sigma|\eta| - \eta^2$ for the difference error of a coordinate with $|\eta| \leq \sigma$. The guided error of such a coordinate and its derivative are

$$\begin{aligned} \phi(\sigma) &= \frac{\sigma^2}{12} + s \cdot 0 + s(1+s)\big(\sigma|\eta| - \eta^2\big), \\ \phi'(\sigma) &= \frac{\sigma}{6} + s(1+s)\,|\eta| \;\geq\; 0, \end{aligned} \tag{17}$$

so ϕ is smallest at the smallest admissible σ, which is $\max(\sigma_c, \sigma_u)$. □

The price of a shared grid. On a coordinate that both branches share, widening the grid of the unconditional branch from σ_u to $(1+\gamma)\sigma_u$ changes its error from $\sigma_u^2/12$ to $(1+\gamma)^2\sigma_u^2/12$, while sharing removes the difference error of part (b) of Proposition 3:

$$\begin{aligned} \text{added error} &= \frac{(1+\gamma)^2 - 1}{12}\,\sigma_u^2 = \frac{2\gamma + \gamma^2}{12}\,\sigma_u^2 \approx \frac{\gamma}{6}\,\sigma_u^2, \\ \text{removed error} &= \sigma_u^2\, p\,(1-p) + R, \qquad p = \frac{|x|\,\gamma}{(1+\gamma)\,\sigma_u} \approx \frac{|x|}{\sigma_u}\,\gamma, \\ \frac{\text{removed}}{\text{added}} &\approx \frac{6\,|x|}{\sigma_u}\,(1-p) \qquad (\gamma \text{ small}). \end{aligned} \tag{18}$$

While $p < \frac{1}{2}$, the removed part grows with the code magnitude $|x|/\sigma_u$ and the added part does not, so on shared coordinates away from zero the trade is favorable. Over a whole layer the net gain is smaller, since the coarser grid raises the error at every coordinate of that branch, while the removed mismatch depends on the values; Proposition 4 weighs both sides through the measured a_c, a_u and b.

A.5 Guidance with More Than Two Branches

The decomposition of Proposition 1 is a special case of one identity for any affine combination of the branch errors. Guidance with several conditions, or any sampler that extrapolates from one evaluation away from the others, combines more than two quantized evaluations with weights that sum to one, and the identity states which pairwise differences such a combination amplifies, and by how much.

Proposition 5 (Affine combinations of branch errors)**.** *Let $\hat{\boldsymbol{e}} = \sum_{k=1}^{m} w_k \boldsymbol{e}_k$ with $\sum_k w_k = 1$, and write $E_k = \mathbb{E}\|\boldsymbol{e}_k\|^2$ and $V_{jk} = \mathbb{E}\|\boldsymbol{e}_j - \boldsymbol{e}_k\|^2$. Then*

$$\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 = \sum_k w_k E_k - \frac{1}{2}\sum_{j,k} w_j w_k\, V_{jk}. \tag{19}$$

A pair of weights with opposite signs adds $|w_j w_k|\, V_{jk}$ to the guided error, and a pair with equal signs subtracts $w_j w_k V_{jk}$.

Proof. Expand the double sum of squared differences and use $\sum_k w_k = 1$:

$$\begin{aligned} \sum_{j,k} w_j w_k \|\boldsymbol{e}_j - \boldsymbol{e}_k\|^2 &= \sum_{j,k} w_j w_k \big(\|\boldsymbol{e}_j\|^2 + \|\boldsymbol{e}_k\|^2 - 2\langle \boldsymbol{e}_j, \boldsymbol{e}_k\rangle\big) \\ &= \Big(\sum_k w_k\Big)\sum_j w_j \|\boldsymbol{e}_j\|^2 + \Big(\sum_j w_j\Big)\sum_k w_k \|\boldsymbol{e}_k\|^2 - 2\Big\|\sum_k w_k \boldsymbol{e}_k\Big\|^2 \\ &= 2\sum_k w_k \|\boldsymbol{e}_k\|^2 - 2\|\hat{\boldsymbol{e}}\|^2. \end{aligned} \tag{20}$$

Rearranging and taking expectations gives Eq. (19). For two branches with $w_c = 1+s$ and $w_u = -s$, the first sum is $(1+s)E_c - sE_u = E_c + s(E_c - E_u)$ and the second is $-\frac{1}{2}\cdot 2\, w_c w_u\, V = s(1+s)V$, which is Proposition 1. □

Remark. With convex weights, Eq. (19) is the ambiguity decomposition of an ensemble (Krogh & Vedelsby, 1995): disagreement among members lowers the error of their average. Guidance extrapolates instead, one weight is negative, and the same disagreement now raises the error in proportion to $|w_j w_k|$. A shared scale over all branches removes the mismatch that per-branch grids add, for every pair at once, since part (a) of Proposition 3 applies to each pair on the common grid.

A.6 One Quantizer in Numbers

We follow the first layer of SDXL at $s = 6.5$ through the statements of Section 4, using the values of Table 4, so that each step can be checked by hand. Under the default, $E_c = 0.00476$, $E_u = 0.00364$ and $C = 0.00124$, and under the shared scale $E'_c = 0.00474$, $E'_u = 0.00525$ and $C' = 0.00385$:

$$\begin{aligned}
V &= 0.00476 + 0.00364 - 2\cdot 0.00124 = 0.00592,\\
V' &= 0.00474 + 0.00525 - 2\cdot 0.00385 = 0.00229,\\
\mathbb{E}\|\hat{\boldsymbol{e}}\|^2 &= 0.00476 + 6.5\cdot 0.00112 + 48.75\cdot 0.00592\\
&= 0.00476 + 0.00728 + 0.28860 = 0.30064,\\
\mathbb{E}\|\hat{\boldsymbol{e}}'\|^2 &= 0.00474 - 6.5\cdot 0.00051 + 48.75\cdot 0.00229 = 0.11306,\\
\Delta(6.5) &= a_c + 13\,(a_c - b) + 42.25\,\Delta V = -0.00002 - 0.03419 - 0.15337 = -0.18758,\\
1 - V'/V &= 1 - 0.00229/0.00592 = 0.613.
\end{aligned} \tag{21}$$

The difference term is 96% of the guided error under the default. The coefficients are $a_c = -0.00002$, $a_u = 0.00161$ and $b = 0.00261$, so $a_u < 2b$ and $\Delta V = -0.00363$; the quadratic removes 62.4% of the guided error at $s = 6.5$, and its limit at large s is 61.3%. With the two per-branch errors of the default held fixed, Eq. (10) gives $V \in [0.00008,\ 0.01673]$, so the guided error could lie anywhere between 0.0157 and 0.8274, a factor of 53 at identical per-branch errors. For a single coordinate with $|x|/\sigma_u = 40$ and a scale gap of $\gamma = 1\%$, part (b) of Proposition 3 then gives $p = 0.396$ and $\mathbb{E}[\delta^2] \approx 0.24\,\sigma_u^2$ when both branches hold the same value, which a shared grid removes completely at an added cost of about $0.0017\,\sigma_u^2$ for the coarsened branch, about 140 times less than it removes.

A.7 Propagation Through the Remaining Layers

Between a quantizer and the output lie the remaining layers, whose Jacobians differ between the branches because the branches see different inputs. A first-order model with two stated assumptions shows how the cross term of a quantizer reaches the output and why the couplings keep their order.

Proposition 6 (Propagation of the cross term, first order)**.** *Let the errors $\boldsymbol{e}_c, \boldsymbol{e}_u$ injected at a quantizer reach the output through the Jacobians $\boldsymbol{J}_c$ and $\boldsymbol{J}_u$, and let later layers add errors $\boldsymbol{n}_c, \boldsymbol{n}_u$ that are not correlated with the injected errors. Assume that on the error directions the Jacobians share a gain g and align to a factor $\lambda \in [-1, 1]$, both the same for every coupling, so that $\mathbb{E}\langle \boldsymbol{J}_c\boldsymbol{e}_c, \boldsymbol{J}_u\boldsymbol{e}_u\rangle = \lambda g^2 C$ and $\mathbb{E}\|\boldsymbol{J}_b\boldsymbol{e}_b\|^2 = g^2 E_b$ for $b \in \{c, u\}$. Then the output cross term is $\lambda g^2 C + \mathbb{E}\langle \boldsymbol{n}_c, \boldsymbol{n}_u\rangle$ and the output per-branch errors are $g^2 E_b + \mathbb{E}\|\boldsymbol{n}_b\|^2$, for every coupling of the two branches and every s.*

Proof. To first order the output errors are $\boldsymbol{e}_c^{\text{out}} = \boldsymbol{J}_c\boldsymbol{e}_c + \boldsymbol{n}_c$ and $\boldsymbol{e}_u^{\text{out}} = \boldsymbol{J}_u\boldsymbol{e}_u + \boldsymbol{n}_u$, so

$$\begin{aligned}
C^{\text{out}} &= \mathbb{E}\langle \boldsymbol{J}_c\boldsymbol{e}_c, \boldsymbol{J}_u\boldsymbol{e}_u\rangle + \mathbb{E}\langle \boldsymbol{n}_c, \boldsymbol{n}_u\rangle + \mathbb{E}\langle \boldsymbol{J}_c\boldsymbol{e}_c, \boldsymbol{n}_u\rangle + \mathbb{E}\langle \boldsymbol{n}_c, \boldsymbol{J}_u\boldsymbol{e}_u\rangle\\
&= \mathbb{E}\big[\boldsymbol{e}_c^\top \boldsymbol{J}_c^\top \boldsymbol{J}_u\boldsymbol{e}_u\big] + C_{\boldsymbol{n}}\\
&= \lambda\, g^2\, \mathbb{E}\langle \boldsymbol{e}_c, \boldsymbol{e}_u\rangle + C_{\boldsymbol{n}} = \lambda g^2 C + C_{\boldsymbol{n}},\\
E_b^{\text{out}} &= \mathbb{E}\|\boldsymbol{J}_b\boldsymbol{e}_b\|^2 + \mathbb{E}\|\boldsymbol{n}_b\|^2 = g^2 E_b + \mathbb{E}\|\boldsymbol{n}_b\|^2, \qquad b \in \{c, u\}.
\end{aligned} \tag{22}$$

The first line expands the output cross term, the second drops the two mixed terms, which vanish because the later errors are uncorrelated with the injected ones, and names $C_{\boldsymbol{n}} = \mathbb{E}\langle \boldsymbol{n}_c, \boldsymbol{n}_u\rangle$. The third line is the alignment assumption, and the last applies the common gain to the per-branch errors. Since λ and g are the same for every coupling by assumption, the cross term of every coupling is scaled by the same factor λg^2. □

Remark. Without the alignment assumption the output cross term is $\mathbb{E}[\boldsymbol{e}_c^\top \boldsymbol{J}_c^\top \boldsymbol{J}_u \boldsymbol{e}_u]$, which depends on the error directions, so the proposition is a first-order model. Without later-layer errors it makes the correlation fall by the factor λ at each layer, the same factor for every coupling. Figure 4(b) agrees with this assumption: the correlation falls by a factor of 0.950, 0.951 and 0.953 per layer under shared randomness, the shared scale and the default, and the order of the couplings is the same at every layer (Table 27). How the sampler carries the error from one step to the next one is not a part of the model, and we measure it in Section 5.2, where the gain grows as the number of steps grows.

B Experimental Details

B.1 Survey of Inference Stacks

For each of the six stacks we followed the path of one guided sampling step and recorded every activation quantizer that it executes, together with its rounding rule and whether its scale is fixed, shared by the two branches, or computed per branch at run time. Counts come from the executed path rather than from configuration fields, as a per-tensor kernel that is called once per branch computes two scales. Table 5 lists the four types per stack; the counts add up to the total in every row. The last column is the share of coordinates that carry the same value in both branches and receive different codes on the inconsistent quantizers, and Table 6 tabulates it by quantizer type, bit width and layer.

Table 5: Activation quantizers of six inference stacks by rounding rule and by whether the scale is shared by the two branches or computed per branch. Consistent is the share of deterministic quantizers with a shared scale, Per-branch the share of deterministic quantizers with per-branch scales, Stoch. the share of stochastic quantizers, and Mism. the share of same-value coordinates that receive different codes on the inconsistent quantizers. The last row pools the counts of the six stacks and averages Mism. over them.

	Deterministic		Stochastic			Share (%)			
Stack	Shared	Per-branch	Shared	Per-branch	Total	Consistent	Per-branch	Stoch.	Mism. (%)
diffusers	42	169	2	0	213	19.7	79.3	0.9	19.8
TensorRT	64	150	3	0	217	29.5	69.1	1.4	26.5
vLLM	83	197	2	1	283	29.3	69.6	1.1	27.9
ComfyUI	78	324	2	3	407	19.2	79.6	1.2	22.8
SANA kernel	96	279	1	3	379	25.3	73.6	1.1	19.2
Vendor stack	62	190	1	1	254	24.4	74.8	0.8	23.2
All six	425	1,309	11	8	1,753	24.2	74.7	1.1	23.2

Table 6: Share (%) of coordinates that carry the same value in both branches and receive different codes, by quantizer type, bit width and layer. Quantizers with a shared scale and deterministic rounding never produce a mismatch.

Type	Bits	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
Det./shared	W4A8	0.0	0.0	0.0	0.0	0.0	0.0
	W8A8	0.0	0.0	0.0	0.0	0.0	0.0
Det./per-branch	W4A8	34.2	17.4	14.5	26.5	30.5	25.6
	W8A8	20.1	20.5	15.6	21.6	12.2	8.4
Stoch./shared	W4A8	12.8	13.1	16.4	15.7	27.0	30.8
	W8A8	17.7	21.7	10.4	10.0	19.4	16.4
Stoch./per-branch	W4A8	19.4	19.9	24.1	34.1	21.8	30.4
	W8A8	22.1	18.6	17.6	9.9	22.1	10.6

Deterministic rounding with per-branch scales holds between 69.1% and 79.6% of the quantizers of every stack and 74.7% overall, and stochastic rounding accounts for 19 of $1{,}753$ quantizers. Across stacks, 19.2% to 27.9% of the same-value coordinates receive different codes on the inconsistent quantizers, and the rate varies from 8.4% to 34.2% across layers and bit widths without falling to zero anywhere, in contrast to the consistent quantizers, whose rate is zero at every layer and bit width.

B.2 Implementation of PairRound

Algorithm 1 gives PairRound at one activation quantizer and one step. The first test leaves a quantizer alone when there is no second branch or the tensors do not correspond. The random stream is keyed by quantizer and step, so the branches draw the same number, and the exchange takes the larger of the two per-branch scales at the chosen granularity. Nothing else in the sampler or model changes.

Algorithm 1 PairRound at one activation quantizer, one guided step.

Require: branch tensors $\boldsymbol{X}_c, \boldsymbol{X}_u$; rounding rule rnd; granularity (coordinate, channel or tensor)
1: **if** only one branch runs at this step, or $\boldsymbol{X}_c$ and $\boldsymbol{X}_u$ do not correspond **then**
2: quantize each evaluated tensor with its own scale and **return**
3: **end if**
4: if rnd is stochastic, draw u from the stream keyed by (quantizer, step) for both branches
5: compute the per-branch scales σ_c, σ_u at the chosen granularity
6: $\bar{\sigma} \leftarrow \max(\sigma_c, \sigma_u)$ ▷ one exchange between the branches
7: $\hat{\boldsymbol{X}}_c \leftarrow \bar{\sigma}\ \mathrm{rnd}(\boldsymbol{X}_c/\bar{\sigma}; u), \quad \hat{\boldsymbol{X}}_u \leftarrow \bar{\sigma}\ \mathrm{rnd}(\boldsymbol{X}_u/\bar{\sigma}; u)$ ▷ u unused if rnd is deterministic
Ensure: $\hat{\boldsymbol{X}}_c, \hat{\boldsymbol{X}}_u$

Two tensors correspond when they have the same shape and their coordinates carry the same meaning in both branches. All activations in the latent space satisfy this, since both branches act on the same state $\boldsymbol{x}_t$, and so do the activations of each decoded token of the language models. The conditional and the unconditional prompts do not, and pairing them by index would couple unrelated values, so their quantizers keep their own scales, as do the other quantizers whose tensors do not correspond, 16 to 18 in total on SDXL, DiT and SD3.5. For the ablation with cached scales, the shared scales recorded in a first run are read back in a second run instead of being exchanged between the branches. Table 7 reports the implementation checks of Algorithm 1, all measured on 10,000 paired samples.

Table 7: Implementation checks on 10,000 samples. Rows 1 to 4 measure zero, and row 5 is the largest change of E_c or E_u between independent and shared randomness.

#	Check	Measured	Samples
1	Noise seeds differ between couplings (share)	0	10,000
2	Stochastic couplings differ at $s = 0$ (distance)	0	10,000
3	Weight codes differ between branches (share)	0	10,000
4	Paired tensors that do not correspond (share)	0	10,000
5	Change of E_c or E_u under recoupling (max.)	0.00016	10,000

B.3 Backbones, Data and Metrics

The text-to-image models run on 10k captions of the COCO 2014 validation set (Lin et al., 2014), with one caption per image identifier drawn once and fixed. DiT, which is class-conditional, runs on ten labels per ImageNet class (Russakovsky et al., 2015). Wan2.x runs on the VBench prompts (Huang et al., 2024b), and the language models run on the 805 AlpacaEval instructions (Li et al., 2023b), each decoded with 12 seeds, for 9,660 generations per configuration. For the language models the conditional branch sees the instruction and the unconditional branch the prompt without it (Sanchez et al., 2024), and the next token is sampled from the guided logits at every decoding step.

FID uses Inception features against the COCO 2014 validation images, or the ImageNet validation images for DiT, and each model is scored against its own reference set. The preference score passes the PickScore values (Kirstain et al., 2023) of the image of a row and of the image of the default for the same prompt through a two-way softmax, so that 0.5 means parity, and averages the result over prompts. The language-model win rate against the default is judged by a fixed instruction-tuned model that sees both answers anonymously and in both orders, with ties allowed. Agg. is a normalized score that combines the trajectory distances of the backbones that a row covers, so rows with a different coverage can be compared, and p99 means the 99th percentile of $|\delta|$ over all coordinates and samples.

Timing uses one NVIDIA H100 GPU with 80 GB of memory in the SDXL reference setting, and the per-row kernel times of Table 12 are measured on the backbone named in its last column. The BF16

references of the two largest language models run on four GPUs, since their weights alone need about 140 GB, and they are not timed.

B.4 Quantization Recipes and Published Methods

Weights are quantized once, identically for both branches, so they cannot create a cross-branch inconsistency. The reference recipe is W4A8 with per-token dynamic activation scales, which is where per-branch scales arise, and the other recipes are W8A8, FP8 (E4M3), per-channel W4 and SmoothQuant+W8A8 (Xiao et al., 2023). The eight published methods keep their own weight quantization and calibration and run at W4A8 on the reference recipe, except SVDQuant, which runs at W4A4 as designed (Li et al., 2025); applied underneath a method, the shared scale changes only its per-branch activation scales, which it replaces by their maximum. The guidance schedules of Table 2 keep the weight constant, raise it linearly over the steps (Wang et al., 2024), apply it on a limited interval (Kynkäänniemi et al., 2024), or change it dynamically during sampling. Each published row of Table 1 covers the backbones for which its method was designed, one to seven out of the eleven.

C Additional Results

C.1 Growth with the Guidance Scale

Table 8 follows the guided error before and after the shared scale at six guidance scales on three backbones, and Figure 5 draws it. The removed error grows nearly in proportion to $s(1+s)$, with R^2 between 0.9996 and 0.9997, and its share of the guided error rises from about half at $s = 1$ to between 59% and 68% for $s \geq 4$, approaching the limit $1 - V'/V$ of Proposition 4. Normalized by $s(1+s)$, the guided error after sharing levels off from $s = 4$ onward on all three backbones, and the spread of the growth across backbones shrinks with sharing by a factor of 2.2 to 4.9 at every scale.

Table 8: Guided error at the first layer before and after the shared scale at each guidance scale in the guidance sweep, with the coefficient $s(1+s)$ of V, the error removed and its share (%).

Model	s	$s(1+s)$	Before	After	Removed	Share (%)
SDXL	0	0	0.00485	0.00485	0.00000	0
	1	2	0.01734	0.00890	0.00844	48.7
	2	6	0.04078	0.01763	0.02315	56.8
	4	20	0.12598	0.04929	0.07669	60.9
	6.5	48.75	0.29212	0.11916	0.17296	59.2
	11	132	0.77008	0.30488	0.46520	60.4
DiT	0	0	0.00399	0.00399	0.00000	0
	1	2	0.01717	0.00763	0.00954	55.6
	2	6	0.04173	0.01523	0.02650	63.5
	4	20	0.13058	0.04196	0.08862	67.9
	6.5	48.75	0.29600	0.10097	0.19503	65.9
	11	132	0.79247	0.26144	0.53103	67.0
SD3.5	0	0	0.00250	0.00250	0.00000	0
	1	2	0.01093	0.00499	0.00594	54.3
	2	6	0.02715	0.01017	0.01698	62.5
	4	20	0.07984	0.02660	0.05324	66.7
	6.5	48.75	0.18978	0.06667	0.12311	64.9
	11	132	0.51568	0.16818	0.34750	67.4

Table 9 tests the decomposition as a predictor. It holds E_c, E_u and C at the first-layer values of Table 4, measured at $s = 6.5$, predicts the guided error at every scale, and compares the prediction with the runs of the guidance sweep in Table 8. The predictions stay within 5% of the measurements with per-branch scales and within 7% with the shared scale, over five scales and three backbones, so the three statistics of one quantizer, measured at one scale, predict the guided errors at every scale.

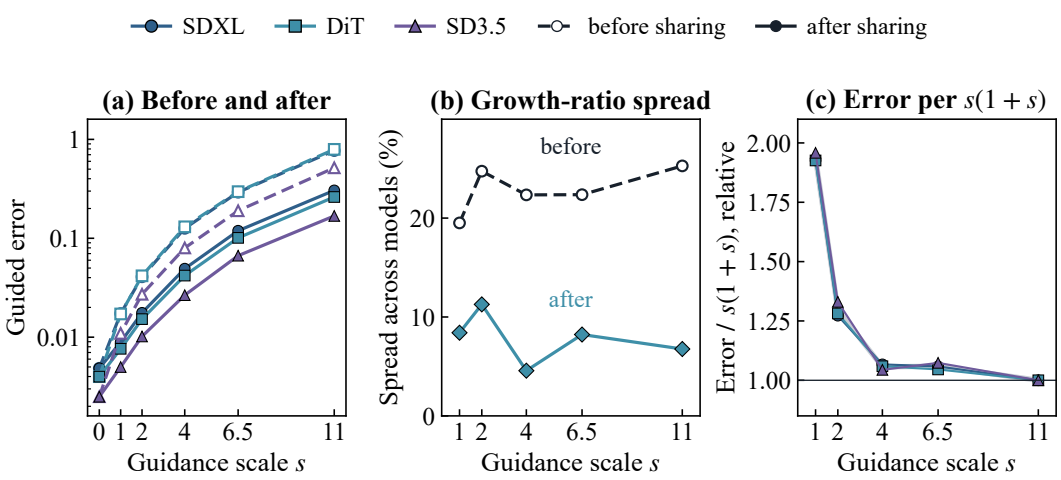


Figure 5: Growth with the guidance scale on SDXL, DiT and SD3.5. (a) Guided error, on a log scale, before (dashed, open) and after (solid, filled) the shared scale. (b) Spread of the growth ratio $G(s)/G(0)$ of the guided error $G(s)$ across the three backbones, $(\max - \min)/\text{mean}$, before and after. (c) Guided error after sharing divided by $s(1+s)$, normalized to one at $s = 11$; the three curves overlap their gray mean.

Table 9: The decomposition as a predictor across guidance scales: guided error measured in the guidance sweep and predicted from E_c, E_u and C measured at the first layer at $s = 6.5$ (Table 4), with the relative difference, predicted minus measured over measured, in percent.

		Per-branch scale			Shared scale		
Model	s	Measured	Predicted	Diff. (%)	Measured	Predicted	Diff. (%)
SDXL	1	0.0173	0.0177	2.2	0.0089	0.0088	−1.0
	2	0.0408	0.0425	4.3	0.0176	0.0175	−1.0
	4	0.1260	0.1276	1.3	0.0493	0.0485	−1.6
	6.5	0.2921	0.3006	2.9	0.1192	0.1131	−5.1
	11	0.7701	0.7985	3.7	0.3049	0.3014	−1.1
DiT	1	0.0172	0.0165	−3.7	0.0076	0.0079	4.1
	2	0.0417	0.0408	−2.3	0.0152	0.0157	3.0
	4	0.1306	0.1244	−4.7	0.0420	0.0435	3.7
	6.5	0.2960	0.2949	−0.4	0.1010	0.1015	0.5
	11	0.7925	0.7863	−0.8	0.2614	0.2707	3.6
SD3.5	1	0.0109	0.0109	−0.6	0.0050	0.0050	0.8
	2	0.0272	0.0266	−1.9	0.0102	0.0098	−3.5
	4	0.0798	0.0810	1.5	0.0266	0.0268	0.8
	6.5	0.1898	0.1917	1.0	0.0667	0.0620	−7.0
	11	0.5157	0.5107	−1.0	0.1682	0.1644	−2.2

C.2 Results on Every Backbone and Metric

Tables 10 and 11 give FID and the preference score on each of the seven image models for all twenty rows of Table 1, and Table 12 gives the video and language columns together with the kernel time. Figure 6 summarizes the change that the shared scale brings on each backbone. The trajectory distance, the quality metric and the preference all move in the favorable direction on every one of the eleven backbones, and the couplings keep their order when only a third of the network is converted.

Table 10: FID (↓) on the seven image models for the twenty rows of Table 1; shaded rows are PairRound and its shared-randomness variant, and a dash marks a backbone that a method does not cover.

Coupling or method	Bits	SD1.5	SDXL	DiT	SD3.5	PixArt-Σ	SANA	Hunyuan-DiT
BF16 (reference)	BF16	9.85	13.82	9.21	16.78	11.36	16.67	10.50
Det., shared per-tensor	W4A8	13.29	19.69	12.99	20.99	15.63	21.12	15.42
Default (per-branch)	W4A8	15.08	22.30	15.36	23.31	17.44	23.04	17.59
Stoch., independent	W4A8	16.07	24.21	16.04	23.94	18.90	24.27	19.30
Stoch., shared rand.	W4A8	10.77	15.60	10.42	18.00	12.62	18.16	11.89
Stoch., antithetic	W4A8	19.20	29.27	19.35	27.82	22.62	27.74	23.58
Oracle, one grid	W4A8	10.94	15.69	10.41	18.03	12.64	18.35	11.98
Shared scale, per coord.	W4A8	12.75	18.80	12.55	20.46	14.98	20.73	14.92
Shared scale, per channel	W4A8	13.09	19.54	12.73	20.53	15.45	20.88	15.28
Shared scale, per tensor	W4A8	13.28	19.62	13.35	20.73	15.79	20.87	15.38
Shared scale, output layer	W4A8	14.78	21.91	14.84	22.88	16.99	22.62	16.99
Shared scale + rand.	W4A8	12.86	18.80	12.35	20.50	15.09	20.77	14.62
Q-Diffusion	W4A8	14.85	21.59	–	–	–	–	–
PTQD	W4A8	14.73	21.70	–	–	–	–	–
TFMQ-DM	W4A8	14.69	21.36	–	–	–	–	–
MixDQ	W4A8	14.62	20.88	–	–	–	–	–
Q-DiT	W4A8	–	–	14.45	22.73	16.65	22.15	16.31
PTQ4DiT	W4A8	–	–	14.34	22.25	16.42	22.22	16.42
ViDiT-Q	W4A8	–	–	–	–	–	–	–
SVDQuant	W4A4	17.20	25.53	17.68	25.62	19.56	25.66	20.25

Table 11: Preference probability against the default (↑, 0.5 is parity) on the seven image models for the twenty rows of Table 1; shading and dashes as in Table 10.

Coupling or method	Bits	SD1.5	SDXL	DiT	SD3.5	PixArt-Σ	SANA	Hunyuan-DiT
BF16 (reference)	BF16	0.692	0.809	0.745	0.682	0.711	0.676	0.779
Det., shared per-tensor	W4A8	0.612	0.659	0.658	0.605	0.604	0.605	0.647
Default (per-branch)	W4A8	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Stoch., independent	W4A8	0.435	0.348	0.441	0.457	0.389	0.440	0.362
Stoch., shared rand.	W4A8	0.681	0.801	0.734	0.678	0.707	0.675	0.768
Stoch., antithetic	W4A8	0.216	0.071	0.182	0.255	0.150	0.238	0.088
Oracle, one grid	W4A8	0.691	0.802	0.742	0.676	0.698	0.672	0.772
Shared scale, per coord.	W4A8	0.633	0.712	0.675	0.622	0.631	0.614	0.687
Shared scale, per channel	W4A8	0.622	0.693	0.659	0.618	0.613	0.605	0.672
Shared scale, per tensor	W4A8	0.607	0.663	0.653	0.615	0.603	0.608	0.654
Shared scale, output layer	W4A8	0.528	0.553	0.533	0.523	0.537	0.521	0.549
Shared scale + rand.	W4A8	0.634	0.710	0.688	0.630	0.642	0.622	0.689
Q-Diffusion	W4A8	0.528	0.553	–	–	–	–	–
PTQD	W4A8	0.538	0.564	–	–	–	–	–
TFMQ-DM	W4A8	0.549	0.574	–	–	–	–	–
MixDQ	W4A8	0.551	0.590	–	–	–	–	–
Q-DiT	W4A8	–	–	0.569	0.543	0.551	0.550	0.582
PTQ4DiT	W4A8	–	–	0.575	0.550	0.558	0.546	0.590
ViDiT-Q	W4A8	–	–	–	–	–	–	–
SVDQuant	W4A4	0.361	0.266	0.316	0.364	0.346	0.375	0.290

Table 12: Video and language results for the rows of Table 1: trajectory distance and VBench on Wan2.x, trajectory distance and win rate against the default on the three language models, kernel time in ms per step and branch on the backbone named in the last column, and the number n of backbones a row covers; shading and dashes as in Table 10.

	Wan2.x		Qwen3-8B		Llama-3.3-70B		Qwen2.5-72B				
Coupling or method	Traj.	VBench	Traj.	Win	Traj.	Win	Traj.	Win	ms	n	Timed on
BF16 (reference)	0	83.89	0	0.720	0	0.736	0	0.717	13.92	11	SD1.5
Det., shared per-tensor	0.096	79.85	0.123	0.651	0.122	0.637	0.117	0.642	12.45	11	SD1.5
Default (per-branch)	0.214	80.33	0.246	0.500	0.258	0.500	0.267	0.500	12.54	11	SD1.5
Stoch., independent	0.278	79.03	0.348	0.411	0.337	0.415	0.332	0.439	12.65	11	SD1.5
Stoch., shared rand.	0.009	82.85	0.012	0.713	0.014	0.729	0.013	0.709	12.62	11	SD1.5
Stoch., antithetic	0.566	76.48	0.681	0.160	0.691	0.164	0.684	0.182	12.56	11	SD1.5
Oracle, one grid	0.008	83.30	0.012	0.698	0.014	0.696	0.013	0.719	12.60	11	SD1.5
Shared scale, per coord.	0.075	82.46	0.094	0.645	0.091	0.649	0.088	0.665	12.62	11	SD1.5
Shared scale, per channel	0.085	83.13	0.111	0.644	0.106	0.654	0.101	0.668	12.70	11	SD1.5
Shared scale, per tensor	0.096	80.72	0.117	0.615	0.118	0.619	0.113	0.644	12.54	11	SD1.5
Shared scale, output layer	0.191	81.49	0.207	0.533	0.223	0.545	0.236	0.539	12.58	11	SD1.5
Shared scale + rand.	0.071	83.20	0.091	0.649	0.088	0.673	0.087	0.681	12.50	11	SD1.5
Q-Diffusion	–	–	–	–	–	–	–	–	12.78	2	SD1.5
PTQD	–	–	–	–	–	–	–	–	12.82	2	SD1.5
TFMQ-DM	–	–	–	–	–	–	–	–	12.70	2	SD1.5
MixDQ	–	–	–	–	–	–	–	–	12.85	2	SD1.5
Q-DiT	–	–	–	–	–	–	–	–	8.13	5	DiT
PTQ4DiT	–	–	–	–	–	–	–	–	8.14	5	DiT
ViDiT-Q	0.175	82.60	–	–	–	–	–	–	14.11	1	Wan2.x
SVDQuant	–	–	–	–	–	–	–	–	12.88	7	SD1.5

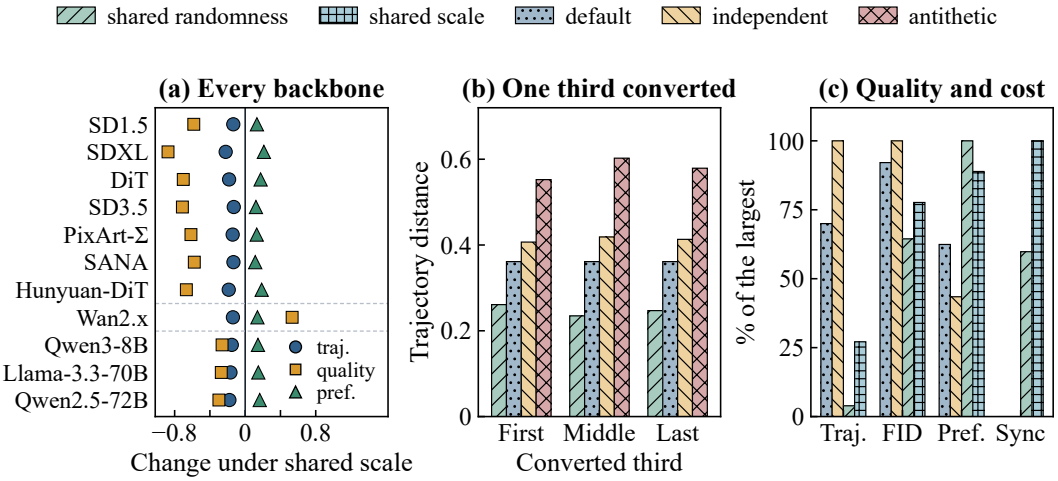


Figure 6: Effect sizes, converted third and cost. (a) Change under the shared scale on each backbone: trajectory distance, quality (FID and VBench divided by four, perplexity as is; VBench rises when quality improves) and preference or win rate. (b) Trajectory distance when only the first, middle or last third of the network is converted, for four couplings. (c) Trajectory distance, FID, preference and synchronization time of four couplings on SDXL, each as a percentage of the largest of the four.

C.3 Sweeps on SDXL, the Language Models and the Video Model

Tables 13 to 15 list the values behind Figure 3 and the bit-width axis. On SDXL, the four couplings coincide at $s = 0$ and separate as s grows, and the shared scale stays below the default at every positive scale and every step count. SVDQuant runs with four-bit activations in the guidance sweep and with eight-bit activations in the step sweep, so it lies above the default in the first and below it in the second. With more weight bits all couplings improve, and the shared scale keeps its place between shared randomness and the default at every bit width. The tails of $|\delta|$ keep the same order.

Table 13: SDXL: trajectory distance against the guidance scale s at 28 steps (left; SVDQuant at W4A4) and against the number of steps at $s = 6.5$ (right; all rows at W4A8). Indep. and Shared are stochastic rounding with independent and shared random numbers, and Sh. scale is PairRound.

s	Default	Indep.	Shared	Sh. scale	SVDQuant
0	0.0058	0.0058	0.0058	0.0058	0.0106
1	0.0213	0.0281	0.0075	0.0106	0.0375
2	0.0502	0.0685	0.0090	0.0216	0.0884
4	0.1514	0.2202	0.0136	0.0601	0.2664
6.5	0.3613	0.5162	0.0202	0.1400	0.6340
11	0.9623	1.4093	0.0357	0.3724	1.6939

Steps	Default	Indep.	Shared	Sh. scale	SVDQuant
4	0.1690	0.2633	0.0098	0.0660	0.1153
12	0.2645	0.3913	0.0151	0.1060	0.1790
28	0.3613	0.5162	0.0202	0.1400	0.2328
50	0.4142	0.6261	0.0253	0.1770	0.2929
100	0.5881	0.8326	0.0312	0.2325	0.3730

Table 14: SDXL: FID against the bit width (top) and the tail of $|\delta|$ for the four W4A8 couplings and SVDQuant, the value exceeded with the stated probability (bottom). A dash marks a bit width without an SVDQuant value.

Bits	Default	Indep.	Shared	Sh. scale	SVDQuant
W4A8	22.30	24.21	15.60	18.80	–
W6A8	19.35	20.50	14.85	16.76	17.90
W8A8	16.95	18.20	14.27	15.90	16.43
FP8	17.64	18.65	14.60	16.23	17.16

Tail	Default	Indep.	Shared	Sh. scale	SVDQuant
0.99	0.0009	0.0012	0.0001	0.0006	0.0008
0.9	0.0093	0.0111	0.0011	0.0061	0.0078
0.75	0.0236	0.0293	0.0028	0.0154	0.0196
0.5	0.0472	0.0617	0.0063	0.0310	0.0426
0.25	0.0862	0.1038	0.0106	0.0562	0.0699
0.1	0.1215	0.1505	0.0152	0.0810	0.1047
0.01	0.1980	0.2340	0.0230	0.1260	0.1585
0.001	0.2484	0.3013	0.0297	0.1555	0.2022
0.0001	0.2991	0.3567	0.0362	0.1902	0.2391

Table 15: Trajectory distance of the three language models against the guidance scale with per-branch scales (the default) and with the shared scale (left), and of Wan2.x against the number of steps at $s = 6.5$ under four couplings (right).

	Qwen3-8B		Llama-3.3-70B		Qwen2.5-72B	
s	Per-br.	Shared	Per-br.	Shared	Per-br.	Shared
0	0.0039	0.0039	0.0039	0.0039	0.0036	0.0036
1	0.0144	0.0072	0.0146	0.0071	0.0146	0.0068
2	0.0352	0.0142	0.0360	0.0139	0.0386	0.0136
4	0.1070	0.0406	0.1071	0.0389	0.1122	0.0373
6.5	0.2460	0.0940	0.2580	0.0910	0.2670	0.0880
11	0.6500	0.2451	0.6749	0.2361	0.7328	0.2327

Steps	Default	Indep.	Shared	Sh. scale
4	0.1082	0.1422	0.0044	0.0401
12	0.1590	0.2064	0.0064	0.0573
28	0.2140	0.2780	0.0085	0.0750
50	0.3073	0.3320	0.0110	0.0983
100	0.3656	0.4864	0.0145	0.1140

C.4 Generality on Every Axis

Table 16 extends Table 2 to every axis we vary: the eleven backbones, the guidance schedules, the quantization recipes, three published methods with the shared scale underneath, three settings in which PairRound changes nothing, and the step counts. The last column is the gain in trajectory

distance over the default of the same setting, and it is zero only in these three settings, with one branch or a scale already shared, where our rule leaves every quantizer and the outputs unchanged.

Table 16: Generality on every axis: trajectory distance under the shared scale, the quality metric of the domain (FID for images, VBench for video, perplexity for language), preference or win rate against the default of each backbone on the backbone axis and of the reference setting elsewhere, p99 of $|\delta|$ and the gain over the default of the same setting.

Axis	Setting	Traj.↓	Quality	Pref.↑	p99↓	Gain↑
Backbone	SD1.5	0.067	12.75	0.633	0.107	0.135
	SDXL	0.140	18.80	0.712	0.126	0.221
	DiT	0.089	12.55	0.675	0.117	0.183
	SD3.5	0.064	20.46	0.622	0.094	0.129
	PixArt-Σ	0.087	14.98	0.631	0.116	0.142
	SANA	0.063	20.73	0.614	0.089	0.133
	Hunyuan-DiT	0.126	14.92	0.687	0.122	0.186
	Wan2.x	0.075	82.46	0.639	0.113	0.139
	Qwen3-8B	0.094	8.37	0.645	0.114	0.152
	Llama-3.3-70B	0.091	8.36	0.649	0.100	0.167
	Qwen2.5-72B	0.088	8.32	0.665	0.114	0.179
Schedule	constant	0.140	18.80	0.712	0.126	0.221
	linear	0.066	17.05	0.763	0.119	0.103
	interval	0.052	16.55	0.779	0.116	0.079
	dynamic	0.089	17.53	0.756	0.122	0.144
Recipe	W8A8	0.026	15.90	0.791	0.055	0.041
	W4A8	0.140	18.80	0.712	0.126	0.221
	FP8	0.037	16.23	0.768	0.064	0.058
	per-channel W4	0.120	18.24	0.723	0.114	0.188
	SmoothQuant+W8A8	0.024	15.57	0.809	0.051	0.034
Method (Agg.)	Q-Diffusion	0.335	15.59	0.730	0.121	0.546
	PTQ4DiT	0.276	16.31	0.693	0.100	0.423
	SVDQuant	0.252	17.86	0.583	0.148	1.428
No change	FLUX.1-dev (distilled)	0.005	14.70	0.799	0.012	0
	$w = 1$	0.006	14.63	0.812	0.015	0
	shared per-tensor scale	0.187	19.69	0.659	0.139	0
Steps	4	0.066	17.38	0.771	0.118	0.103
	12	0.106	18.27	0.737	0.121	0.158
	28	0.140	18.80	0.712	0.126	0.221
	50	0.177	19.53	0.667	0.131	0.237
	100	0.233	20.32	0.642	0.136	0.356

C.5 Published Methods, Coverage and Paired Differences

Table 17 gives the published methods with the shared scale applied underneath, and Table 18 the paired differences that the text quotes, where negative distance and FID and positive preference favor the first entry. Figure 7 shows how the twenty rows of Table 1 separate as the guidance scale grows and how many backbones each row covers. The worst of the nineteen quantized rows is 2.6 times the best at $s = 0$ and 88 times at $s = 11$, so the choice of coupling matters more with strong guidance.

C.6 Four Metrics

Table 19 evaluates six couplings on three backbones under the trajectory distance, FID, a normalized kernel maximum mean discrepancy between image features, and the preference score. Trajectory distance, FID and preference rank the six couplings in the same order, up to near ties between the shared scale and its combination with shared randomness. The kernel discrepancy separates the couplings less and orders some differently, and it puts the default further from the best coupling than the other three metrics do, and on SDXL it ranks the shared scale behind the default (see Figure 8).

Table 17: Published methods with the shared scale applied underneath: aggregate trajectory distance, FID and preference over the image models that a method covers, p99, the absolute reduction Red. of the aggregate distance, and the number n of backbones covered; a dash marks a metric without a covered model. SVDQuant runs at W4A4, the others at W4A8.

Method	Bits	Traj.↓	FID↓	Pref.↑	p99↓	Red.	n
Q-Diffusion + PairRound	W4A8	0.335	15.59	0.730	0.121	0.546	2
PTQD + PairRound	W4A8	0.317	15.69	0.676	0.106	0.507	2
TFMQ-DM + PairRound	W4A8	0.300	15.49	0.684	0.105	0.484	2
MixDQ + PairRound	W4A8	0.291	15.40	0.691	0.104	0.448	2
Q-DiT + PairRound	W4A8	0.281	16.15	0.661	0.094	0.465	5
PTQ4DiT + PairRound	W4A8	0.276	16.31	0.693	0.100	0.423	5
ViDiT-Q + PairRound	W4A8	0.289	–	–	0.104	0.481	1
SVDQuant + PairRound	W4A4	0.252	17.86	0.583	0.148	1.428	7

Table 18: Paired differences in aggregate trajectory distance, mean FID and mean preference over the image models.

Comparison	Traj.	FID	Pref.
Shared scale − default	−0.614	−2.704	0.153
Shared scale − output layer only	−0.492	−2.260	0.119
Stoch.: shared − independent randomness	−1.284	−6.467	0.310
Shared scale (W4A8) − SVDQuant (W4A4)	−1.321	−5.187	0.322
SVDQuant + PairRound − SVDQuant (both W4A4)	−1.428	−3.780	0.252
Shared scale + randomness − shared scale	−0.016	−0.029	0.006

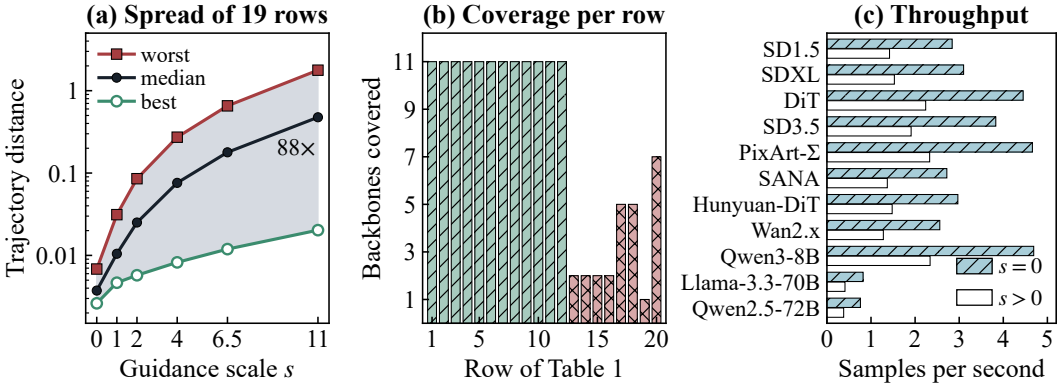


Figure 7: The rows of Table 1. (a) Trajectory distance, on a log scale, of the best, median and worst of the nineteen quantized rows at each guidance scale, with the range shaded. (b) Number of backbones that each row covers; green bars cover all eleven. (c) Samples per second on each backbone at $s = 0$, where one branch runs, and at $s > 0$, where two run.

Table 19: Six couplings under four metrics on three backbones. MMD^2 is a normalized squared kernel maximum mean discrepancy between the image features of the quantized and the BF16 sampler.

Model	Coupling	Traj.↓	FID↓	MMD^2 ↓	Pref.↑
SDXL	Default (per-branch)	0.361	22.30	1.111	0.500
	Stoch., independent	0.516	24.21	1.162	0.348
	Stoch., shared randomness	0.020	15.60	1.019	0.801
	Stoch., antithetic	1.011	29.27	1.141	0.071
	Shared scale	0.140	18.80	1.133	0.712
	Shared scale + randomness	0.137	18.80	1.089	0.710
DiT	Default (per-branch)	0.271	15.36	1.089	0.500
	Stoch., independent	0.328	16.04	1.054	0.441
	Stoch., shared randomness	0.012	10.42	0.948	0.734
	Stoch., antithetic	0.645	19.35	1.110	0.182
	Shared scale	0.089	12.55	1.003	0.675
	Shared scale + randomness	0.086	12.35	1.016	0.688
SD3.5	Default (per-branch)	0.193	23.31	1.139	0.500
	Stoch., independent	0.235	23.94	1.162	0.457
	Stoch., shared randomness	0.008	18.00	1.086	0.678
	Stoch., antithetic	0.457	27.82	1.140	0.255
	Shared scale	0.064	20.46	1.117	0.622
	Shared scale + randomness	0.061	20.50	1.080	0.630

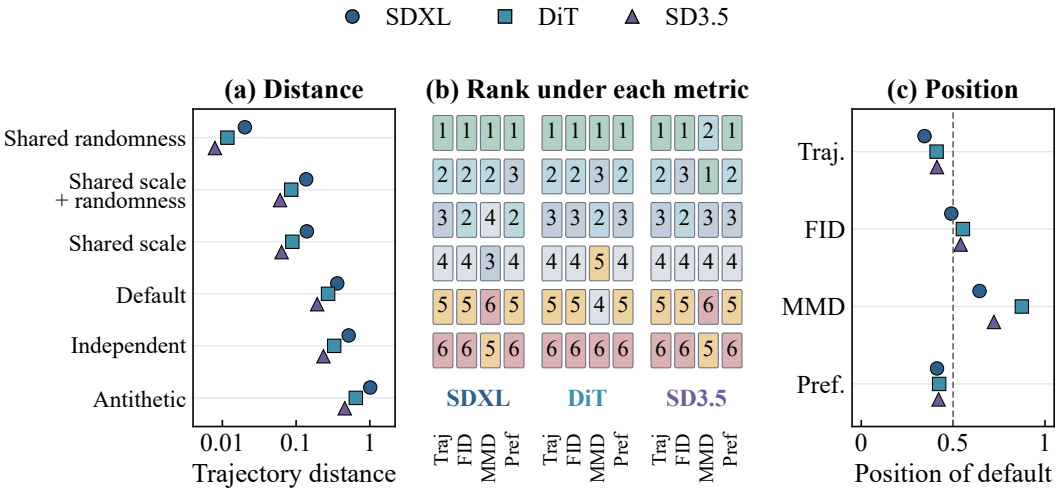


Figure 8: Four metrics on SDXL, DiT and SD3.5. (a) Trajectory distance of the six couplings on a logarithmic axis. (b) Rank of each coupling under each metric, rows in the order of (a). (c) Position of the default between the best coupling (0) and the worst (1) under each metric.

D Additional Ablations

D.1 Fraction of the Layers Converted

Table 20 and Figure 9 convert a fraction of the layers to the shared scale, from $1/32$ of the layers to the whole network. On all three backbones the gain rises close to linearly with the fraction converted, with departures from the linear trend of at most 0.012 in trajectory distance. Every fraction we test gains, and each tested fraction keeps a roughly proportional share of the gain of the whole network.

Table 20: Trajectory distance, gain over the default and nominal synchronization (the full-network value times the fraction, μs per step and branch) when a fraction of the layers is converted to the shared scale.

Model	Fraction	Traj.↓	Gain↑	Sync
SDXL	0	0.3613	0	0
	1/32	0.3568	0.0045	0.37
	1/4	0.2992	0.0621	2.94
	1/2	0.2513	0.1100	5.87
	3/4	0.1979	0.1634	8.81
	1	0.1400	0.2213	11.74
DiT	0	0.2713	0	0
	1/32	0.2642	0.0071	0.38
	1/4	0.2137	0.0577	3.01
	1/2	0.1844	0.0869	6.01
	3/4	0.1360	0.1353	9.02
	1	0.0888	0.1825	12.02
SD3.5	0	0.1930	0	0
	1/32	0.1865	0.0065	0.37
	1/4	0.1571	0.0359	2.97
	1/2	0.1259	0.0670	5.94
	3/4	0.0957	0.0973	8.92
	1	0.0635	0.1295	11.89

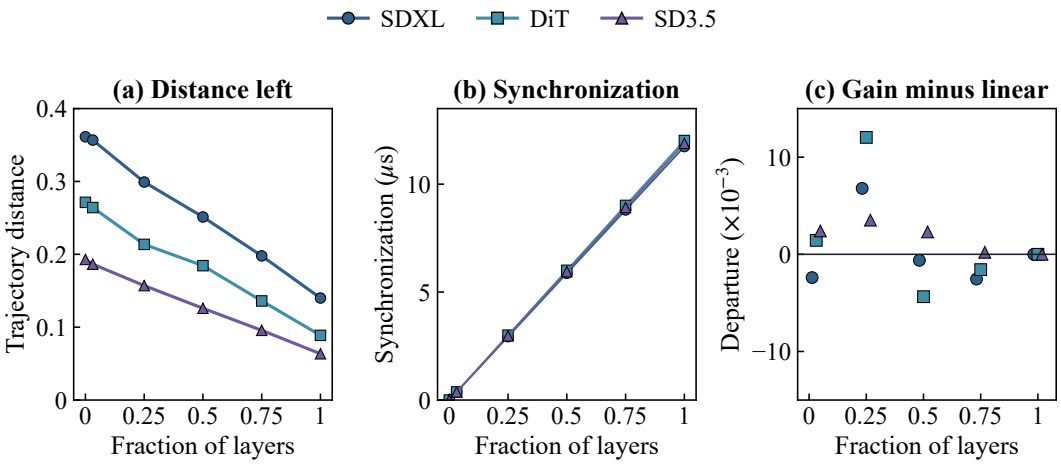


Figure 9: Fraction of the layers converted on SDXL, DiT and SD3.5. (a) Trajectory distance that remains. (b) Nominal synchronization, the full-network value times the fraction, in μs per step and branch. (c) Measured gain minus the gain that grows linearly with the fraction, in units of 10^{-3}.

D.2 The Maximum Against a Narrower Shared Range

Corollary 1 makes the maximum the finest shared grid that clips neither branch, and a narrower range trades clipping for a finer grid. Table 21 and Figure 10 test this trade on a scalar stress test: pairs of activations from each backbone are quantized to a low-bit grid over a shared half-range, values outside the range are clipped, and the guided error is compared with the error at the maximum scale, whose half-range is 1.18. On this test every narrower range lowers the guided error, by up to 0.081 at

a half-range of 0.9, where 8.4% to 8.6% of the coordinates are clipped, and a wider range raises it. In the deployed networks, the mean of the two scales, which clips the larger branch, keeps 97.6% of the gain of the maximum (Table 3), so we keep the maximum, which needs no search of the range.

Table 21: Scalar stress test of the shared range: guided error under each half-range, the error at the maximum scale (half-range 1.18, shaded rows), the saving over it, and the share of coordinates clipped (%).

Model	Half-range	Guided error	At maximum	Saving	Clipped (%)
SDXL	0.9	0.1279	0.2089	0.0810	8.55
	1	0.1507	0.2089	0.0581	1.22
	1.09	0.1689	0.2089	0.0400	0.23
	1.18	0.2089	0.2089	0	0
	1.3	0.2462	0.2089	−0.0373	0
DiT	0.9	0.1309	0.2041	0.0732	8.57
	1	0.1494	0.2041	0.0548	1.27
	1.09	0.1718	0.2041	0.0323	0.22
	1.18	0.2041	0.2041	0	0
	1.3	0.2461	0.2041	−0.0420	0
SD3.5	0.9	0.1336	0.2028	0.0693	8.42
	1	0.1522	0.2028	0.0507	1.16
	1.09	0.1736	0.2028	0.0292	0.05
	1.18	0.2028	0.2028	0	0
	1.3	0.2362	0.2028	−0.0334	0

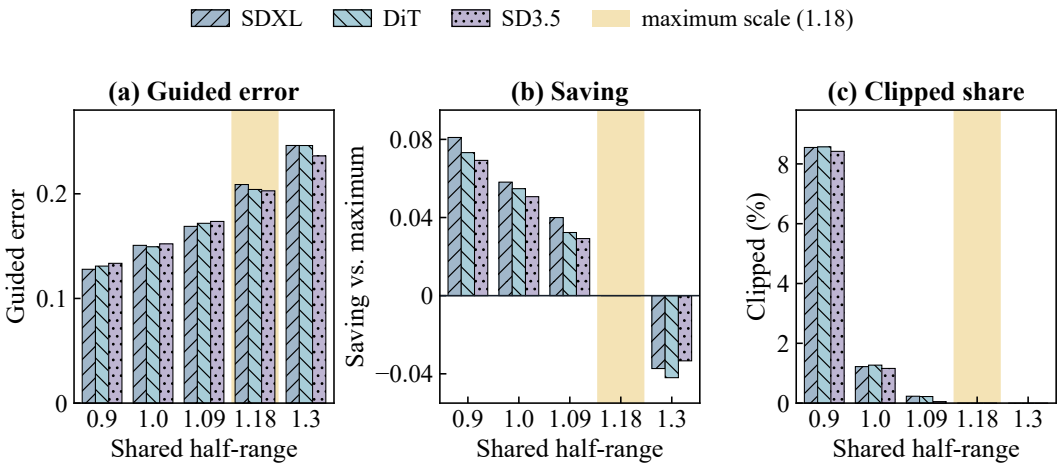


Figure 10: Shared range on the stress test for SDXL, DiT and SD3.5. (a) Guided error under each shared half-range; the shaded column is the maximum scale. (b) Saving over the error at the maximum scale. (c) Share of coordinates clipped.

D.3 Pieces of the Sampling Trajectory

Figure 11 applies the shared scale on pieces of the sampling trajectory. The sampling steps are split into three consecutive pieces with weights 0.31, 0.36 and 0.33, their shares of the trajectory, and the shared scale is applied on a union of pieces. The gain of each union we test stays close to the sum of the gains of its pieces. The step-wise picture matches the layer-wise one of Appendix D.1: each piece contributes a share of the gain close to its own weight, as Figure 11(b) shows piece by piece.

D.4 Interventions and Granularity

Figure 12 draws the shared-scale results of Table 4 together with the granularity sweep. On SDXL the shared scale removes between 49% and 61% of the guided error at every positive scale, and at $s = 0$ the two errors coincide with that of a single branch. At every positive guidance scale, sharing per coordinate gives a lower distance than per channel, per channel lower than per tensor, and all three stay well below sharing at the output layer alone. On all seven image backbones, a_u stays below b.

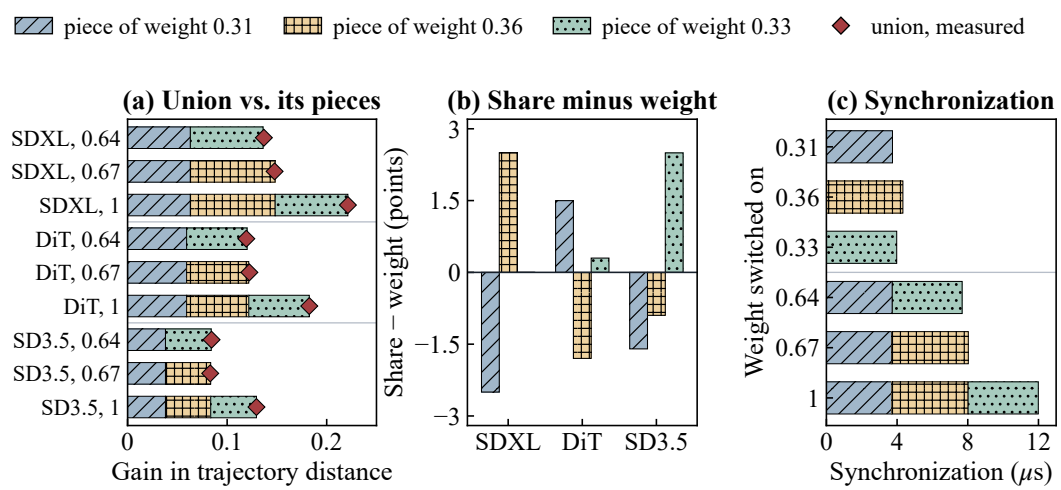


Figure 11: Pieces of the sampling trajectory on SDXL, DiT and SD3.5. (a) Gain in trajectory distance of three unions of pieces, labeled by their total weight and stacked by piece, with the measured gain of each union as a diamond. (b) Share of the whole gain contributed by each piece minus its weight, in percentage points. (c) Nominal synchronization of each piece and union, the full-network value split by weight, in μs per step and branch.

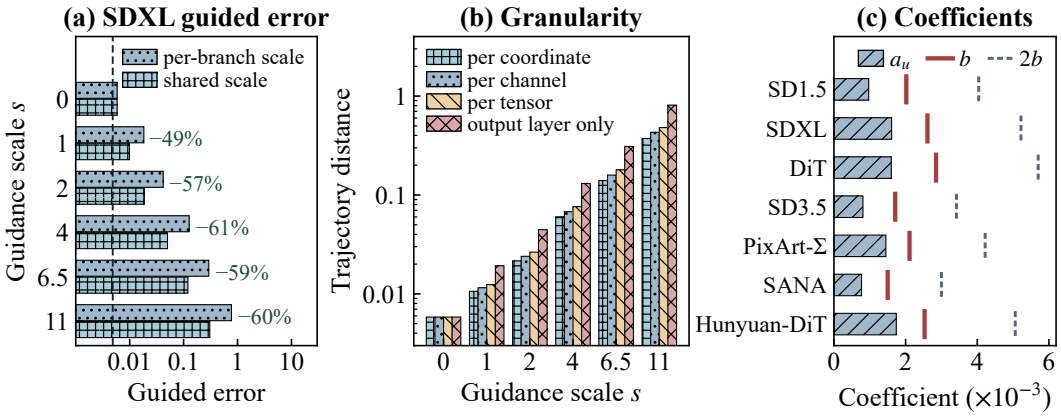


Figure 12: Interventions and granularity. (a) Guided error on SDXL, on a log scale, with per-branch scales and with the shared scale at six guidance scales, with the share removed; the dashed line is the error of one branch alone. (b) Trajectory distance on SDXL, on a log scale, for four granularities of sharing at six guidance scales. (c) The coefficients a_u, b and $2b$ of Proposition 4 on seven image backbones, in units of 10^{-3}.

D.5 One Third of the Network

Table 22 converts only the first, the middle or the last third of the network to a stochastic coupling while the rest keeps the default. The order of the couplings is the same in each third, with shared randomness lowest and antithetic randomness highest, and no third alone recovers the gain of converting the whole network, in line with the nearly proportional growth shown in Appendix D.1.

D.6 Observables Behind the Ablations

Table 23 lists the observables behind the ablations of Section 5.3 and the checks of Section 5.4 on the three backbones, with their mean. The shared scale supplies about 98% of the joint gain on every backbone, the output layer alone keeps about a fifth, the first third of the steps about three tenths, the gain grows nearly in proportion to $s(1+s)$, and under a fifth of the paired prompts become worse.

Table 22: SDXL: trajectory distance and FID when only one third of the network is converted to each coupling; the default row is the same in all three thirds.

Third	Coupling	Traj.↓	FID↓
First	Default (per-branch)	0.3613	22.30
	Stoch., independent	0.4068	22.87
	Stoch., shared randomness	0.2610	20.37
	Stoch., antithetic	0.5524	24.40
Middle	Default (per-branch)	0.3613	22.30
	Stoch., independent	0.4188	23.01
	Stoch., shared randomness	0.2348	19.85
	Stoch., antithetic	0.6025	24.94
Last	Default (per-branch)	0.3613	22.30
	Stoch., independent	0.4132	22.94
	Stoch., shared randomness	0.2470	20.10
	Stoch., antithetic	0.5791	24.69

Table 23: The observables behind the ablations and checks on SDXL, DiT and SD3.5, and their mean. Shares are fractions and times are in μs.

#	Observable	SDXL	DiT	SD3.5	Mean
1	Share of the joint gain from the shared scale	0.987	0.984	0.978	0.983
2	Mean minus maximum scale, stress-test error	-0.041	-0.032	-0.031	-0.035
3	Guided error, antithetic minus independent	0.403	0.357	0.231	0.330
4	Share of the gain kept, output layer only	0.232	0.210	0.193	0.212
5	Share of the gain kept, first third of steps	0.313	0.304	0.305	0.307
6	Distance, per channel minus per coordinate	0.019	0.011	0.009	0.013
7	Quantizers left unpaired (count)	17	18	16	17
8	Shared random stream (μs per step and branch)	7.02	7.02	6.77	6.94
9	Scale exchange (μs per step and branch)	11.74	12.02	11.89	11.88
10	FID gain of the shared scale	3.50	2.81	2.85	3.05
11	R^2 of the gain against $s(1+s)$	0.9997	0.9996	0.9997	0.9997
12	Gain with a single branch	0	0	0	0
13	Share of prompts where sharing is worse	0.187	0.189	0.176	0.184

E Mechanism Details

E.1 Per-Sample Behavior

Figure 13 follows the errors sample by sample on 512 paired inputs per backbone. The propagated error over the error at the quantizer depends on the model, between 0.77 and 1.24 across the eleven backbones, and plays the role of the gain g^2 of Proposition 6. The sample correlation of the two branch errors follows the order of Proposition 2 on every backbone, close to one under shared randomness, close to zero under independent randomness and close to minus one under antithetic randomness, with the default between 0.19 and 0.35. Under every coupling the worst tenth of the samples carries 41% to 46% of the squared error and the worst hundredth 7% to 8%, although the mean error changes by two orders of magnitude; the spread over samples is alike for all couplings.

E.2 The Cross Term on Every Backbone

Figure 14 reports, for each of the eleven backbones, the branch errors, with E_u before and after the scale is shared, and the guided error under four couplings, averaged over six layers, together with the cross term relative to its bound $\sqrt{E_c E_u}$. Sharing the scale raises E_u by 32% to 49%, while the correlation rises from between 0.19 and 0.35 under the default to between 0.76 and 0.78, and the guided error with the shared scale falls between shared randomness and the default on each backbone.

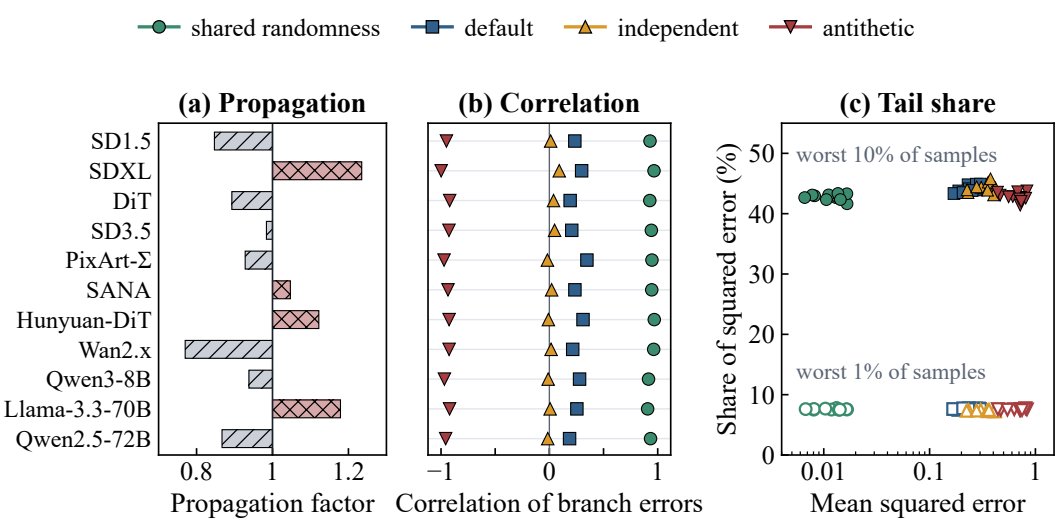


Figure 13: Per-sample view on eleven backbones. (a) Propagated error over the error at the quantizer; bars start at one, so bars to the left damp and bars to the right amplify. (b) Sample correlation of the two branch errors under four couplings. (c) Share of the squared error carried by the worst 51 (filled) and the worst 5 (open) of 512 samples against the mean squared error on a log axis.

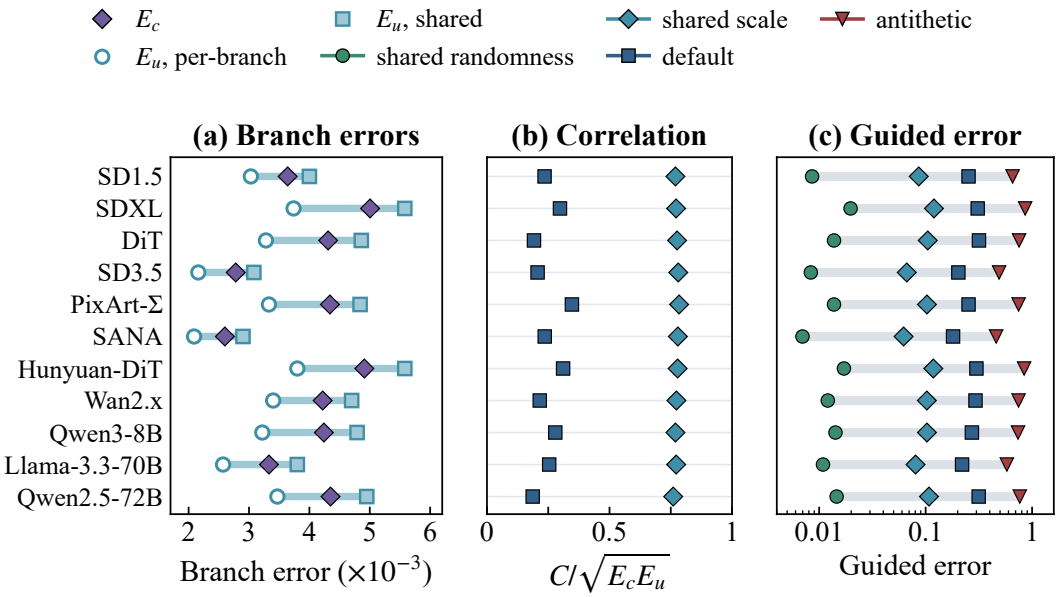


Figure 14: The cross term on every backbone at $s = 6.5$. (a) Branch errors averaged over six layers: E_c, and E_u with per-branch scales and with the shared scale. (b) Cross term relative to its bound, $C/\sqrt{E_c E_u}$, for the default and the shared scale. (c) Guided error under four couplings averaged over six layers, on a log axis, with the range shaded.

E.3 Cross Term of Six Couplings

Table 24 places the cross term of six couplings at the first layer of SDXL inside its Cauchy–Schwarz interval. Shared randomness sits near the upper end of the interval and antithetic randomness at its lower end, independent randomness stays close to zero, and the deterministic couplings lie in between, the shared per-tensor scale higher than the per-branch default. The shared scale reaches 77% of its bound, with a larger bound because it raises one per-branch error by coarsening its grid.

Table 24: Cross term C at the first layer of SDXL for six couplings, the ends of its interval $\mp\sqrt{E_c E_u}$ at the per-branch errors of the same row, and the correlation $\rho = C/\sqrt{E_c E_u}$ (errors in units of 10^{-3}).

Coupling	C	$-\sqrt{E_c E_u}$	$+\sqrt{E_c E_u}$	ρ
Det., shared per-tensor scale	3.34	−4.89	4.89	0.683
Det., per-branch scale (default)	1.24	−4.16	4.16	0.298
Stoch., independent	0.38	−4.17	4.17	0.091
Stoch., shared randomness	4.11	−4.26	4.26	0.966
Stoch., antithetic	−4.16	−4.17	4.17	−0.999
Shared scale (PairRound)	3.85	−4.99	4.99	0.772

E.4 Distribution of the Branch Difference

Table 25 gives the quantiles of the branch difference δ behind Figure 2(b). The median is zero and the distribution is close to symmetric under every coupling. Between the 1st and 99th percentile, the default is 8.6 times as wide as under shared randomness, independent 10.2 and antithetic 14.8 times.

Table 25: Quantiles of the branch difference δ over coordinates in the SDXL reference setting.

Coupling	p1	p10	p25	p50	p75	p90	p99	p99.9	p99.99
Default (per-branch)	−0.1814	−0.0966	−0.0502	0	0.0506	0.0991	0.1769	0.2381	0.2736
Stoch., independent	−0.2083	−0.1177	−0.0616	0	0.0611	0.1272	0.2166	0.2827	0.3338
Stoch., shared randomness	−0.0206	−0.0118	−0.0060	0	0.0065	0.0115	0.0211	0.0286	0.0339
Stoch., antithetic	−0.3056	−0.1663	−0.0888	0	0.0878	0.1620	0.3126	0.3989	0.4560

E.5 Single-Source Runs

The single-source runs of Figure 4(a) separate the sources of trajectory deviation. In a run of the default at $s = 6.5$, the local quantization error of each branch is recorded at every paired quantizer as the difference between the quantized and the unquantized version of the same input, so that errors carried from earlier layers are not counted twice, and the two records give the common error $\bar{\boldsymbol{e}}$ and the difference $\boldsymbol{\delta}$. The errors of the later, unpaired layers and of the quantizers whose scale is already shared are recorded in the same way at their own positions. Four runs of the BF16 sampler with the same prompts and noise then add one source each: $\bar{\boldsymbol{e}}$ to both branches, $+\boldsymbol{\delta}/2$ and $-\boldsymbol{\delta}/2$ to the two branches, the later-layer errors, or the errors of the consistent quantizers. The four runs are repeated with each source recorded in a run with the shared scale, which gives the lower bars of Figure 4(a).

Table 26 lists them. Under the default, the difference-only run accounts for 84.6% to 87.2% of their sum on every model. Recorded under the shared scale, the same run falls by 70.8% to 77.9%, while the other three sources change by less than 0.0013, so the repair acts on the difference source alone.

E.6 Branch Correlation by Layer

Table 27 gives the correlation of the two branch errors after each of twelve layers of SDXL, behind Figure 4(b). The three couplings with a correlation lose it at about the same rate, and under independent randomness the correlation starts at 0.09 and stays within 0.02 of zero after layer two.

E.7 Theory Against Measurement

Table 28 collects the quantitative predictions of Sections 3 and 4 and the measurements that test them, including those settings for which the analysis predicts no effect and the measurements show none.

Table 26: Trajectory distance from the BF16 sampler of the four single-source runs on seven image models, with the sources recorded under the default or under the shared scale, and the share (%) of the difference-only run in their sum.

Model	Recorded under	Common	Difference	Later layers	Consistent quant.	Share (%)
SD1.5	default	0.0119	0.1748	0.0090	0.0063	86.5
	shared scale	0.0116	0.0402	0.0090	0.0062	60.0
SDXL	default	0.0245	0.3055	0.0185	0.0128	84.6
	shared scale	0.0240	0.0847	0.0185	0.0128	60.5
DiT	default	0.0159	0.2361	0.0114	0.0079	87.0
	shared scale	0.0148	0.0545	0.0115	0.0080	61.4
SD3.5	default	0.0108	0.1681	0.0083	0.0058	87.1
	shared scale	0.0107	0.0390	0.0082	0.0057	61.4
PixArt-Σ	default	0.0153	0.1936	0.0118	0.0083	84.6
	shared scale	0.0155	0.0513	0.0119	0.0083	59.0
SANA	default	0.0109	0.1709	0.0084	0.0058	87.2
	shared scale	0.0111	0.0378	0.0084	0.0057	60.0
Hunyuan-DiT	default	0.0202	0.2652	0.0164	0.0112	84.8
	shared scale	0.0209	0.0775	0.0166	0.0116	61.3

Table 27: Correlation of the two branch errors after each layer of SDXL under four couplings.

Layer	Shared randomness	Shared scale	Default	Independent
1	0.966	0.772	0.298	0.091
2	0.990	0.745	0.309	0.043
3	0.912	0.674	0.268	0.018
4	0.839	0.637	0.251	−0.006
5	0.787	0.631	0.257	0.012
6	0.805	0.601	0.232	−0.009
7	0.727	0.546	0.227	−0.004
8	0.723	0.509	0.214	0.007
9	0.665	0.481	0.204	−0.003
10	0.636	0.485	0.201	−0.006
11	0.634	0.471	0.181	0.002
12	0.529	0.453	0.183	0.001

Table 28: Predictions of the analysis and the measurements that test them.

Quantity	Source	Predicted	Measured
Change of E_c, E_u under recoupling	Prop. 2	0	at most 0.00016
Order shared > independent > antithetic of ρ	Prop. 2	all backbones	11 of 11
Mismatch on deterministic shared-scale quantizers	Prop. 3	0%	0%
V after sharing the scale	Prop. 3	lower	-61% to -67%
Condition $a_u \le 2b$	Prop. 4	holds	$a_u/b \in [0.47, 0.69]$
Share removed at large s (SDXL, DiT, SD3.5)	Prop. 4	61.3, 64.8, 67.4%	60.4, 67.0, 67.4%
Guided error across scales from one scale	Prop. 1	agreement	within 7.0%
Gain against $s(1+s)$	Prop. 1	near proportional	$R^2 \ge 0.9996$
Per-layer decay of ρ (shared rand., scale, default)	Prop. 6	equal (assumed)	0.950, 0.951, 0.953
Gain with one branch or an already shared scale	Eq. (3)	0	0 in 9 of 9

F Cost

Table 29 lists, for each granularity of the shared scale, the synchronization per step and branch, the kernel time, the peak memory, the throughput and the two quality columns of the reference setting. The synchronization is the only compute cost that PairRound adds. It falls with the granularity, from $11.7\,\mu$s per step and branch when the scale is shared per coordinate to $4\,\mu$s when only the output layer shares it, and at every granularity it stays near 0.1% of the kernel time or below. The measured throughput of every granularity lies between 1.538 and 1.560 samples per second, within 1.5% of the default, and the peak memory lies within 0.35 GB of it. PairRound leaves the kernels unchanged, so these differences come from run-to-run timing variation, which exceeds the synchronization it adds. At $s = 0$, only one branch runs and nothing is synchronized, like a guidance-distilled model.

Table 29: Cost in the SDXL reference setting: synchronization in μs per step and branch, kernel time in ms per step and branch, peak memory in GB, throughput in samples per second, and the SDXL trajectory distance and FID (Tables 1 and 10). The shaded row is the default.

Granularity	Sync	Kernel	Memory	Samples/s	Traj.↓	FID↓
Shared scale, per coordinate	11.74	11.377	9.66	1.560	0.140	18.80
Shared scale, per channel	8.00	11.543	9.75	1.538	0.159	19.54
Shared scale, per tensor	5.00	11.417	9.55	1.555	0.180	19.62
Shared scale, output layer only	4.00	11.526	10.05	1.541	0.310	21.91
Default (per-branch scale)	0	11.551	9.71	1.538	0.361	22.30

G Additional Related Work

Diffusion models and samplers. Diffusion models learn to reverse a noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b), and deterministic and higher-order samplers integrate the learned field in few steps (Song et al., 2021a; Karras et al., 2022). The sampler carries the error made at each step into the next one, in line with the growth of the gain with the number of steps in Table 2. Large text-to-image systems rely on strong guidance (Nichol et al., 2022; Saharia et al., 2022), where $s(1 + s)$ multiplies the difference error by 48.75 to 132 for w ranging from 7.5 to 12.

Quantization of language models. Weight quantization (Frantar et al., 2023; Lin et al., 2024; Shao et al., 2024) and activation smoothing (Xiao et al., 2023) lower the per-branch error, and dynamic per-token activation scales (Dettmers et al., 2022; Yao et al., 2022) make the scales of the two branches differ. Quantization-aware training passes gradients through the rounding with a straight-through estimator (Bengio et al., 2013); like post-training methods, it targets the error of each forward pass.

Evaluation. FID compares Inception features (Heusel et al., 2017), and kernel discrepancies on image features are a known alternative (Jayasumana et al., 2024). Agreement between text and image and human preference are modeled by CLIPScore (Hessel et al., 2021) and by learned preference models (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023), and our preference score is a pairwise comparison against the default in the spirit of Bradley & Terry (1952), paired per prompt and seed.

H Notation

Table 30 groups the symbols of this paper by topic, with the section or appendix of each definition.

Table 30: Notation.

Symbol	Meaning	Defined in
w, $s = w - 1$	guidance weight and guidance scale	Section 1
$\boldsymbol{x}_t$, c	sampler state at step t and the condition	Section 3
$\boldsymbol{\epsilon}_c, \boldsymbol{\epsilon}_u \in \mathbb{R}^D$, $\hat{\boldsymbol{\epsilon}}_t$	conditional, unconditional and guided predictions	Section 3
σ, rnd, y, f, u	scale, rounding rule, value in grid units x/σ, fractional part, uniform random number	Section 3
$\boldsymbol{e}_c, \boldsymbol{e}_u, \hat{\boldsymbol{e}}$	branch errors and the error of the guided prediction	Section 3
E_c, E_u, C	per-branch errors and their cross term	Section 3
$\boldsymbol{\delta}$, V, $\bar{\boldsymbol{e}}$	branch difference, its mean square, and the common error	Section 3, App. A
$\boldsymbol{X}_c, \boldsymbol{X}_u, \sigma_c, \sigma_u, \bar{\sigma}$	paired tensors, per-branch scales, shared scale	Section 4
f_c, f_u, u_c, u_u, K	fractional parts, random numbers and their copula	Section 4.1, App. A.2
x_c, x_u, p	values of the two branches at a coordinate, probability that their codes differ	Section 1
η, δ, γ	value difference $x_c - x_u$, difference error at a coordinate, relative scale gap	Section 4.2
q	largest code of absolute-maximum scaling	App. A.4
E_c', E_u', C', V'	the same quantities after sharing the scale	Section 4.2
a_c, a_u, b, ΔV, $\Delta(s)$, s_0	changes caused by sharing and their effect on the guided error	Section 4.2
ρ	correlation $C/\sqrt{E_c E_u}$ of the branch errors	Section 5.4
w_k, E_k, V_{jk}	weights, errors and pairwise differences of m branches	App. A.5
$\boldsymbol{J}_c, \boldsymbol{J}_u$, g, λ, $\boldsymbol{n}_c, \boldsymbol{n}_u$	branch Jacobians, gain, alignment, later-layer errors	App. A.7

I Numbers Behind the Figures

The tables of this section list the values behind the figures that no earlier table reports. Table 31 belongs to Figure 1 and Table 32 to Figure 2(a), and each other table names its figure in the caption.

Table 31: Numbers behind Figure 1: per-branch errors (10^{-3}) of four couplings at W4A8, and the trajectory distance to the BF16 sampler at $s = 6.5$ and at $s = 0$ (the latter in units of 10^{-3}).

Model	Coupling	E_c	E_u	Traj., $s = 6.5$	Traj., $s = 0$
SDXL	Stoch., shared rand.	4.92	3.68	0.0202	5.83
	Default (per-branch)	4.76	3.64	0.3613	5.83
	Stoch., independent	4.76	3.65	0.5162	5.83
	Stoch., antithetic	4.82	3.60	1.0113	5.83
DiT	Stoch., shared rand.	3.95	3.13	0.0117	3.65
	Default (per-branch)	4.02	3.22	0.2713	3.65
	Stoch., independent	4.06	3.26	0.3285	3.65
	Stoch., antithetic	4.11	3.22	0.6450	3.65
SD3.5	Stoch., shared rand.	2.68	2.15	0.0079	2.63
	Default (per-branch)	2.68	2.10	0.1930	2.63
	Stoch., independent	2.57	2.03	0.2348	2.63
	Stoch., antithetic	2.60	2.00	0.4570	2.63

Table 32: Numbers behind Figure 2(a): per-branch errors E_c and E_u, the same under the three couplings, and the guided error $\mathbb{E}\|\hat{\boldsymbol{e}}\|^2$ under each coupling at six layers of SDXL (10^{-3}), with the ratio of the guided error under independent to that under shared randomness.

			Guided error			
Layer	E_c	E_u	Shared rand.	Default	Independent	Indep./shared
1	4.76	3.64	31.5	300.6	384.9	12.2
2	4.90	3.63	16.6	300.9	424.6	25.6
3	5.03	3.70	17.0	305.9	433.8	25.4
4	5.07	3.78	17.4	301.9	441.5	25.4
5	5.06	3.83	17.7	323.0	442.2	24.9
6	5.21	3.85	18.1	317.4	453.6	25.1

Table 33: Numbers behind Figure 6(a): change of the trajectory distance, of the quality metric of the domain and of the preference or win rate when the default is replaced by the shared scale. FID and perplexity (PPL) fall and VBench rises when quality improves.

Backbone	ΔTraj.	Metric	ΔQuality	ΔPref.
SD1.5	-0.135	FID	-2.33	0.133
SDXL	-0.221	FID	-3.50	0.212
DiT	-0.183	FID	-2.81	0.175
SD3.5	-0.129	FID	-2.85	0.122
PixArt-Σ	-0.142	FID	-2.46	0.131
SANA	-0.133	FID	-2.31	0.114
Hunyuan-DiT	-0.186	FID	-2.67	0.187
Wan2.x	-0.139	VBench	2.13	0.139
Qwen3-8B	-0.152	PPL	-0.26	0.145
Llama-3.3-70B	-0.167	PPL	-0.27	0.149
Qwen2.5-72B	-0.179	PPL	-0.30	0.165

Table 34: Numbers behind Figure 7(a, c): trajectory distance of the best, median and worst of the nineteen quantized rows of Table 1 at each guidance scale, with the ratio of worst to best (left), and throughput in samples per second with one branch ($s = 0$) and with two branches ($s > 0$) on each backbone (right).

s	Best	Median	Worst	Worst/best
0	0.0026	0.0037	0.0068	2.6
1	0.0047	0.0105	0.0313	6.7
2	0.0057	0.0251	0.0854	14.9
4	0.0082	0.0760	0.2726	33.1
6.5	0.0119	0.1786	0.6545	54.9
11	0.0203	0.4757	1.7745	87.6

Backbone	$s = 0$	$s > 0$
SD1.5	2.84	1.42
SDXL	3.10	1.53
DiT	4.45	2.24
SD3.5	3.83	1.91
PixArt-Σ	4.66	2.33
SANA	2.72	1.37
Hunyuan-DiT	2.97	1.48
Wan2.x	2.56	1.28
Qwen3-8B	4.69	2.34
Llama-3.3-70B	0.82	0.41
Qwen2.5-72B	0.76	0.38

Table 35: Numbers behind Figure 11(a, b): gain in trajectory distance of three unions of pieces of the trajectory, as the sum of the gains of its pieces and as measured (left), and the share of the whole gain contributed by each piece minus its weight, in percentage points (right). Pieces 1, 2 and 3 have weights 0.31, 0.36 and 0.33, their shares of the trajectory.

Model	Pieces	Weight	Sum of pieces	Measured
SDXL	1, 3	0.64	0.1361	0.1369
	1, 2	0.67	0.1483	0.1477
	1, 2, 3	1.00	0.2213	0.2213
DiT	1, 3	0.64	0.1201	0.1192
	1, 2	0.67	0.1217	0.1223
	1, 2, 3	1.00	0.1825	0.1825
SD3.5	1, 3	0.64	0.0840	0.0844
	1, 2	0.67	0.0835	0.0830
	1, 2, 3	1.00	0.1295	0.1295

Model	Piece 1	Piece 2	Piece 3
SDXL	−2.50	2.50	0.00
DiT	1.50	−1.80	0.30
SD3.5	−1.60	−0.90	2.50

Table 36: Numbers behind Figure 13(a, b): propagated error over the error at the quantizer, and sample correlation of the two branch errors under four couplings, on each backbone.

		Correlation of the branch errors			
Backbone	Propagation	Shared rand.	Default	Independent	Antithetic
SD1.5	0.847	0.929	0.235	0.012	−0.950
SDXL	1.235	0.966	0.298	0.091	−0.999
DiT	0.893	0.927	0.192	0.038	−0.920
SD3.5	0.984	0.942	0.207	0.048	−0.927
PixArt-Σ	0.928	0.947	0.347	−0.018	−0.972
SANA	1.047	0.945	0.236	0.021	−0.936
Hunyuan-DiT	1.122	0.968	0.311	−0.009	−0.925
Wan2.x	0.770	0.963	0.215	0.016	−0.926
Qwen3-8B	0.938	0.915	0.279	−0.011	−0.969
Llama-3.3-70B	1.179	0.907	0.254	0.007	−0.921
Qwen2.5-72B	0.867	0.934	0.187	−0.015	−0.957

Table 37: Numbers behind Figure 13(c): smallest and largest value over the eleven backbones of the mean squared error and of the share (%) of the squared error carried by the worst 51 and the worst 5 of 512 samples, for each coupling.

	Mean sq. error		Worst 51 (%)		Worst 5 (%)	
Coupling	Min	Max	Min	Max	Min	Max
Stoch., shared rand.	0.007	0.017	41.6	43.4	7.5	7.8
Default (per-branch)	0.167	0.302	43.3	44.9	7.5	7.7
Stoch., independent	0.224	0.419	43.1	45.8	7.2	7.5
Stoch., antithetic	0.438	0.832	41.4	43.8	7.5	7.8

Table 38: Numbers behind Figure 14: branch errors (10^{-3}), with E_u under per-branch and under shared scales, and the guided error under four couplings, averaged over six layers, and the correlation $\rho = C/\sqrt{E_c E_u}$ of the two branch errors under both scales, as in Figure 13(b), all at $s = 6.5$.

		E_u		ρ		Guided error			
Backbone	E_c	Per-branch	Shared	Per-branch	Shared	Shared rand.	Shared scale	Default	Antithetic
SD1.5	3.64	3.03	4.00	0.235	0.769	0.0086	0.0862	0.2530	0.6560
SDXL	5.01	3.74	5.58	0.298	0.772	0.0197	0.1198	0.3083	0.8610
DiT	4.31	3.28	4.86	0.192	0.776	0.0138	0.1047	0.3170	0.7550
SD3.5	2.78	2.16	3.08	0.207	0.781	0.0084	0.0665	0.2030	0.4900
PixArt-Σ	4.34	3.33	4.84	0.347	0.784	0.0138	0.1029	0.2530	0.7470
SANA	2.60	2.09	2.90	0.236	0.779	0.0070	0.0621	0.1810	0.4600
Hunyuan-DiT	4.91	3.80	5.58	0.311	0.778	0.0171	0.1181	0.3000	0.8420
Wan2.x	4.22	3.40	4.70	0.215	0.773	0.0121	0.1027	0.2940	0.7460
Qwen3-8B	4.24	3.22	4.79	0.279	0.769	0.0142	0.1029	0.2720	0.7380
Llama-3.3-70B	3.33	2.57	3.80	0.254	0.772	0.0109	0.0804	0.2200	0.5790
Qwen2.5-72B	4.35	3.47	4.95	0.187	0.760	0.0146	0.1075	0.3150	0.7660

Table 39: Numbers behind Figures 5(b, c) and 12(b): spread of the growth ratio $G(s)/G(0)$ across SDXL, DiT and SD3.5 before and after sharing (%), their ratio, and the guided error after sharing divided by $s(1+s)$ and normalized to one at $s = 11$ (left); trajectory distance on SDXL for four granularities of sharing (right).

	Spread (%)			Normalized error		
s	Before	After	Ratio	SDXL	DiT	SD3.5
1	19.5	8.4	2.3	1.927	1.926	1.958
2	24.7	11.3	2.2	1.272	1.282	1.330
4	22.3	4.6	4.9	1.067	1.059	1.044
6.5	22.4	8.2	2.7	1.058	1.046	1.073
11	25.3	6.8	3.7	1.000	1.000	1.000

s	Coord.	Channel	Tensor	Output
0	0.0058	0.0058	0.0058	0.0058
1	0.0106	0.0115	0.0124	0.0193
2	0.0216	0.0240	0.0264	0.0446
4	0.0601	0.0677	0.0763	0.1309
6.5	0.1400	0.1590	0.1800	0.3099
11	0.3724	0.4306	0.4804	0.8113