PREDICTIVE SPECULATIVE COUPLING: DISTRIBUTION-PRESERVING ACCELERATION OF AUTOREGRESSIVE VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive visual generators produce images and videos one token at a time, making sequential target-model evaluations the dominant inference cost. Draft-free speculative Jacobi decoding evaluates a block in parallel, but it initializes each new round with distributions recycled from an earlier speculative context. This one-round lag lowers proposal–target overlap, and the loss compounds along the verified prefix as the window grows. We introduce our PSC method, a training-free framework that separates proposal optimization from sampling correctness. PSC extrapolates aligned proposal drift to form a transported distribution, couples the existing proposal sample to that distribution, and then couples the transported sample to the target through exact maximal correction. The composed decoder preserves the complete autoregressive joint distribution for any non-anticipating transport policy. Across Lumina-mGPT-7B, Janus-Pro-7B, and Cosmos-AR-4B, PSC consistently reduces target evaluations and end-to-end latency while maintaining comparable generation quality to exact autoregressive decoding. The proposed vectorized decoding path achieves up to $3.07\times$ NFE reduction and $2.51\times$ latency speedup over SJD. Extensive ablation studies verify the contribution of each component, confirming that the proposed transport and correction mechanisms jointly enable efficient acceleration without compromising sampling fidelity.

1 INTRODUCTION

Autoregressive (AR) visual models map a prefix to a categorical distribution and sample one token at a time. This interface supports image generation, unified multimodal modeling, and world simulation in systems such as Lumina-mGPT-7B, Janus-Pro-7B, and Cosmos-AR-4B (Liu et al., 2024; Chen et al., 2025; Agarwal et al., 2025); recent visual AR systems further explore next-scale prediction, improved visual tokenizers, and unified multimodal next-token modeling (Tian et al., 2024; Yu et al., 2023; Xie et al., 2024). However, a visual sample still contains hundreds or thousands of tokens, so ancestral sampling invokes the target model once per token. Reducing this sequential depth is the central inference problem.

Speculative decoding shortens this path by drafting a token block and verifying it in one target-model pass (Leviathan et al., 2023; Chen et al., 2023). For visual generation, Speculative Jacobi Decoding (SJD) removes the auxiliary drafter by recycling target distributions from one parallel Jacobi round as proposals for the next (Teng et al., 2024). Speculative Coupled Decoding (SCD) improves token reuse by coupling samples across adjacent distributions without changing either marginal (So et al., 2025). These methods remain exact, but their efficiency depends on proposal–target overlap. Figure 1 summarizes this efficiency gap and the PSC decoding cycle: predictive transport improves the recycled proposal, while exact correction and valid-prefix commitment preserve the target distribution.

The remaining failure is specific to recycled proposals: after the speculative context changes, the distribution reused by SJD describes the preceding round rather than the target being verified. If event i survives with conditional probability a_i, a prefix of length k survives with probability $\prod_{i=1}^{k} a_i$; early mismatch therefore suppresses every longer-prefix event.

PSC separates two mechanisms. Predictive proposal transport uses aligned distributions from completed Jacobi rounds to extrapolate a family $q_{t,\beta}$, with $\beta = 0$ recovering raw proposal reuse. Because the speculative token was sampled from q_t before $q_{t,\beta}$ was chosen, a first maximal coupling transports the sample to $q_{t,\beta}$; after one parallel target pass, a second coupling corrects it to p_t.

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.

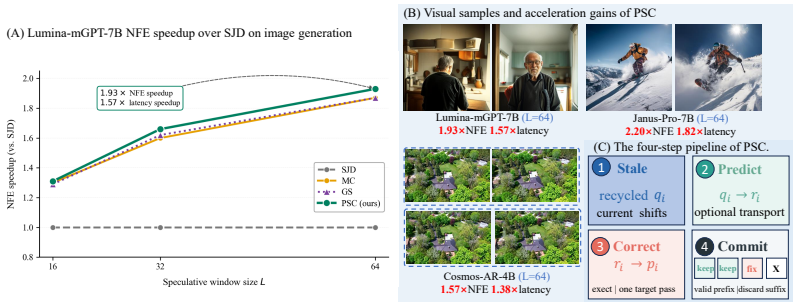


Figure 1: PSC overview for efficient autoregressive visual generation. (A) NFE speedup over SJD across speculative window sizes. (B) Representative samples and acceleration results on three visual generation models. (C) Four-step PSC decoding pipeline.

The construction is training-free, adds no target-model evaluation, and preserves the complete AR joint distribution for every non-anticipating transport choice. Thus prediction can change overlap while correction preserves the target law. The experiments evaluate the resulting decoder across two image models and one video model, with fixed-strength and selector ablations isolating transport-policy sensitivity.

Our contributions are as follows:

- We identify one-round proposal staleness in draft-free Jacobi decoding and formulate proposal–target overlap as the quantity connecting local mismatch to committed-prefix length.
- We introduce a training-free proposal-transport interface that can alter proposal geometry without invalidating an already drawn speculative sample.
- We compose exact sample transport with exact target correction and prove preservation of the complete autoregressive joint distribution for any non-anticipating transport choice.
- We evaluate the exact vectorized decoder across two image generators and one video generator, and analyze transport-policy sensitivity separately from end-to-end systems gains.

2 Background and Related Work

2.1 Autoregressive visual generation

Let $x = (x_1, \ldots, x_N)$ denote a sequence of discrete visual tokens from vocabulary $\mathcal{V}$. An AR model with parameters θ factorizes the generation distribution as

$$p_\theta(x) = \prod_{i=1}^{N} p_\theta(x_i \mid x_{<i}). \tag{1}$$

Exact ancestral sampling evaluates $p_\theta(\cdot \mid x_{<i})$ sequentially for every position. We call each sequential target-model invocation a neural function evaluation (NFE). Image and video tokenizers make N large, so NFE is a direct measure of the sequential generation path. Recent tokenizer and visual-AR designs, including MAGVIT-v2 and multi-scale autoregressive generation, make this sequential interface explicit (Yu et al., 2023; Tian et al., 2024).

2.2 Speculative decoding and SJD

Speculative decoding proposes a block $\tilde{x}_{i:i+L-1}$ from distributions $q_i, \ldots, q_{i+L-1}$ and evaluates the target distributions $p_i, \ldots, p_{i+L-1}$ in parallel (Leviathan et al., 2023; Chen et al., 2023; Miao et al., 2023). For one proposed token $X \sim q$, modified rejection sampling keeps X with probability

$$\alpha(X; p, q) = \min\left\{1, \frac{p(X)}{q(X)}\right\}. \tag{2}$$

On rejection, it samples from the normalized positive residual $[p - q]_+$. The resulting token follows p exactly. Verification proceeds from left to right and stops at the first rejection because every later target distribution conditions on an invalid speculative prefix.

SJD obtains the proposal block from the target model itself through parallel Jacobi iteration (Song et al., 2021; Teng et al., 2024). After each verification round, uncommitted target distributions initialize the next round. This design is training-free and avoids an extra draft network, but proposal quality depends on convergence between adjacent rounds. SCD changes the joint sampling rule between adjacent distributions and increases token collisions without changing either marginal. Maximal coupling attains the largest possible one-step collision probability, while Gumbel coupling shares random perturbations through the Gumbel-max construction (Gumbel, 1954) and provides a different multi-step stability profile. Recent speculative decoders also improve proposal quality with multiple decoding heads, feature-level drafting, dynamic draft trees, and self-speculative execution (Cai et al., 2024; Li et al., 2024a;b; Zhang et al., 2024; Elhoushi et al., 2024).

Adaptive verification methods control how much draft work reaches the target model. DSpark trains a semi-autoregressive language-model drafter and schedules verification from learned prefix-survival confidence and a serving-engine cost profile (Cheng et al., 2026). Other systems adapt the draft online, select candidate lengths, or account for hardware-dependent verification cost (Liu et al., 2023; Huang et al., 2024; Chen et al., 2024). PSC instead changes the proposal distribution before verification in draft-free visual Jacobi decoding; its coupling construction accounts for tokens sampled before that change.

2.3 Proposal staleness

At absolute position i and speculative round t, let $q_i^{(t)}$ denote the recycled proposal entering verification and let $p_i^{(t)}$ denote the newly evaluated target distribution. SJD treats $q_i^{(t)}$ as a static approximation of $p_i^{(t)}$. Changes from $q_i^{(t-1)}$ to $q_i^{(t)}$ reveal how probability mass moves as the Jacobi context converges; discarding this direction produces a one-round lag.

The acceptance consequence is exact rather than heuristic. For categorical maximal correction, the expected keep probability is

$$a_i^{(t)} = \sum_{v \in \mathcal{V}} \min\{p_i^{(t)}(v), q_i^{(t)}(v)\} = 1 - \mathrm{TV}\left(p_i^{(t)}, q_i^{(t)}\right). \tag{3}$$

If G_L is the number of consecutive proposals that survive in a window of length L, then

$$\mathbb{E}[G_L] = \sum_{k=1}^{L} \prod_{i=1}^{k} a_i, \tag{4}$$

where each a_i is conditional on all earlier proposals surviving. The product makes long-window efficiency sensitive to early mismatch and motivates a predictor whose objective is overlap but whose errors cannot alter the final distribution.

2.4 Coupling as the correctness interface

A coupling of distributions r and s is any joint distribution $\pi(X, Y)$ whose marginals are r and s. Maximal coupling maximizes $\Pr[X = Y]$ and reaches $1 - \mathrm{TV}(r, s)$. This property supplies both pieces required by PSC: optimal token reuse and an exact marginal change when proposal prediction changes the distribution. Section 3 applies it first to proposal transport and then to target correction.

3 Predictive Speculative Coupling

3.1 Overview and decoding cycle

At each speculative position, let q denote the raw proposal recycled by SJD, let $r = q_\beta$ denote the transported proposal selected by PSC, and let p denote the target distribution evaluated on the transported prefix. One round changes proposal geometry from q to r, then uses two couplings to make the existing sample compatible with r and correct it to p. Figure 2 expands this $q \to r \to p$ construction across a speculative window: unchanged events form the retained prefix, the first corrected event is committed, and all later events are discarded because their target contexts are no longer valid. All predictor state belongs to one generation request and uses absolute token positions, preventing histories from crossing samples or becoming misaligned after variable-length commits.

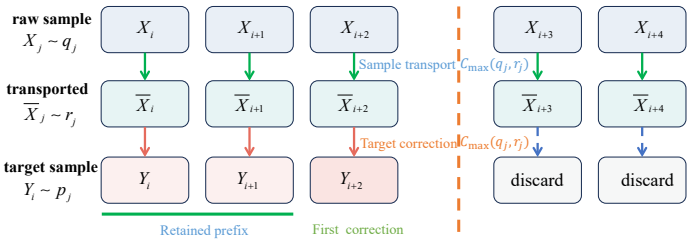


Figure 2: One PSC verification round. Tokens are transported from q_j to r_j and corrected to p_j; verification commits the retained prefix and first corrected token, then discards the invalid suffix.

3.2 Predictive Proposal Transport

Consider an event at absolute position i in round t. The raw SJD proposal is $q_i^{(t)}$, and an aligned history entry supplies $q_i^{(t-1)}$ when the same absolute position appears in the preceding round. We measure directional change in log-probability space,

$$\Delta_i^{(t)}(v) = \mathrm{clip}\Big(\log q_i^{(t)}(v) - \log q_i^{(t-1)}(v), -c, c\Big)\,. \tag{5}$$

Clipping limits extrapolation when either distribution assigns very small mass. For a small ordered set $\mathcal{B}$ with $0 \in \mathcal{B}$, predictive distribution transport constructs

$$q_{i,\beta}^{(t)}(v) = \frac{q_i^{(t)}(v)\exp\Big(\beta\Delta_i^{(t)}(v)\Big)}{\sum_{u\in\mathcal{V}} q_i^{(t)}(u)\exp\Big(\beta\Delta_i^{(t)}(u)\Big)}, \qquad \beta \in \mathcal{B}. \tag{6}$$

Equation (6) extrapolates the current direction for $\beta > 0$ and exactly recovers the raw proposal for $\beta = 0$. Events without aligned history always use $\beta = 0$.

3.3 Exact Coupling and Prefix Commit

The raw token $X_i^{(t)}$ already satisfies $X_i^{(t)} \sim q_i^{(t)}$. Replacing its associated density by $q_{i,\beta}^{(t)}$ without changing the token creates an invalid pair. We instead apply categorical maximal coupling from the raw proposal q to the selected proposal $r = q_\beta$:

$$\Pr[\widetilde{X} = X \mid X] = \min\left\{1, \frac{r(X)}{q(X)}\right\}, \tag{7}$$

$$\widetilde{X} \mid (\widetilde{X} \neq X) \sim \mathrm{Normalize}([r - q]_+)\,. \tag{8}$$

The output satisfies $\widetilde{X} \sim r$ exactly. This coupling maximizes reuse of the existing token under the selected prediction, while exactness holds uniformly across selector decisions. The token and the selected proposal travel together as one immutable event into target evaluation.

One parallel target forward pass evaluates $p_i^{(t)}$ along prefixes formed by the transported block $\widetilde{X}_{1:L}$. For each event in left-to-right order, a second maximal coupling maps $\widetilde{X}_i \sim q_{i,\beta}$ to $Y_i \sim p_i$:

$$\Pr[Y_i = \widetilde{X}_i \mid \widetilde{X}_i] = \min\left\{1, \frac{p_i(\widetilde{X}_i)}{q_{i,\beta}(\widetilde{X}_i)}\right\}, \tag{9}$$

$$Y_i \mid (Y_i \neq \widetilde{X}_i) \sim \mathrm{Normalize}([p_i - q_{i,\beta}]_+)\,. \tag{10}$$

The decoder commits retained events before the first rejection and the residual replacement at that rejection; later events are discarded because their target distributions condition on the rejected prefix. If all events survive, the decoder samples a bonus token from the final target distribution. KV-cache rollback follows the same committed-prefix length, and a stop token truncates the prefix before cache accounting.

3.4 Horizon-Aware Online Selection

Exactness permits any non-anticipating transported proposal; PSC therefore treats candidate selection as request-local online calibration rather than part of correctness. Proposal predictability varies

across a window: early positions receive more stable contexts, whereas late positions inherit more speculative uncertainty. We partition the window into horizon buckets h and maintain statistics separately for each (h, β) pair. After the target forward pass, we compute the counterfactual overlap

$$O^{(t)}_{h,\beta} = \operatorname*{mean}_{i \in h} \sum_{v \in \mathcal{V}} \min \left\{ p^{(t)}_i(v), q^{(t)}_{i,\beta}(v) \right\} \tag{11}$$

over events with valid history and record the gain relative to $\beta = 0$. Exponential moving estimates $(\mu_{h,\beta}, \sigma_{h,\beta})$ define the conservative score

$$S_{h,\beta} = \mu_{h,\beta} - \lambda \sigma_{h,\beta}. \tag{12}$$

The next round selects the candidate with the largest score after a minimum number of observations. Negative-gain feedback contracts the next decision toward the $\beta = 0$ reference, turning observed overlap into closed-loop transport calibration. Selection uses only completed rounds, so the current target distribution never leaks into the current proposal. Including $\beta = 0$ gives the controller a no-transport fallback without changing the guarantee.

Algorithm 1: Predictive Speculative Coupling

Input: Target AR model p_θ; prefix x; window L; candidates $\mathcal{B}$
Output: Exact sample from p_θ

Initialize request-local proposal history and overlap statistics
while x *is incomplete* **do**
  Obtain raw SJD events $(X_i, q_i, \mathrm{pos}_i)_{i=1}^{L}$
  Align proposal history by absolute position and form $q_{i,\beta}$ with Eq. (6)
  Select one β per horizon from statistics of completed rounds
  Couple $X_i \colon\ q_i \to q_{i,\beta}$ to obtain transported events $\widetilde{X}_i$
  Evaluate $(p_i)_{i=1}^{L}$ in one parallel target-model forward pass
  Couple $\widetilde{X}_i \colon\ q_{i,\beta} \to p_i$ from left to right
  Commit the valid prefix through the first correction, or append a target bonus after full acceptance
  Roll back the cache and retain only uncommitted events aligned with the new prefix
  Measure candidate overlap with p_i and update selector statistics
  Store raw proposals q_i as the next aligned history
return x

3.5 Computational properties

The method adds elementwise log-space arithmetic and two vectorized categorical maximal-coupling operations, but no neural network or extra target forward pass. Candidate overlaps share the already available target probabilities, and state updates stay on the accelerator to avoid per-round synchronization. For $K = |\mathcal{B}|$, window L, and vocabulary size $|\mathcal{V}|$, the additional dense work is $O(KL|\mathcal{V}|)$ and is parallel across events and vocabulary entries. The wall-clock benefit depends on whether a longer committed prefix amortizes this bandwidth cost; Section 5 reports NFE and latency together.

4 Theoretical Guarantees

This section formalizes the separation between proposal quality and sampling correctness. We use the standard maximal-coupling identity and a standard tail-sum argument as tools; the PSC-specific claim is that composing $q \to q_\beta \to p$ with valid-prefix commit preserves the complete autoregressive law. All distributions share a finite vocabulary $\mathcal{V}$. We write $[z]_+ = \max(z, 0)$ elementwise and $\mathcal{A}(r, s) = \sum_v \min\{r(v), s(v)\}$.

Lemma 1 (Exact maximal transport)**.** *Let $X \sim r$ and let s be any categorical distribution on the same support. The keep-or-resample rule*

$$Y = \begin{cases} X, & U \le \min\{1, s(X)/r(X)\}, \\ Z, & \textit{otherwise}, \end{cases} \qquad Z \sim \mathrm{Normalize}([s - r]_+) \tag{13}$$

with $U \sim \mathcal{U}[0, 1]$ produces $Y \sim s$ and satisfies

$$\Pr[Y = X] = \mathcal{A}(r, s) = 1 - \mathrm{TV}(r, s). \tag{14}$$

No coupling of r and s has a larger equality probability.

Lemma 1 records the standard maximal-coupling identity used by both PSC stages: the positive residual restores the selected destination marginal for every normalized candidate. What is specific to PSC is the composition of this identity across proposal transport and target correction; prediction can target acceptance without entering the correctness argument.

Theorem 2 (Exactness of predictive speculative correction)**.** *Let $X \sim q$, let q_β be any normalized proposal selected without access to the current target-side random draw, and let p be the target distribution. If PSC first applies exact maximal transport $q \to q_\beta$ and then exact maximal correction $q_\beta \to p$, its output Y satisfies $Y \sim p$ for every admissible predictor output q_β.*

Proof. Lemma 1 gives $\widetilde{X} \sim q_\beta$ after the first coupling. Applying the same lemma with source q_β and destination p gives $Y \sim p$ after the second coupling. The composition therefore preserves the target marginal. □

Theorem 3 (Autoregressive joint-distribution preservation)**.** *Assume the target distributions in each verification window equal $p_\theta(\cdot \mid x_{<i})$ along the proposed prefixes, target correction follows Eq. (9)–(10), and the decoder discards every event after the first rejection. Then the complete sequence returned by Algorithm 1 follows*

$$\Pr[X_{1:N} = x_{1:N}] = \prod_{i=1}^{N} p_\theta(x_i \mid x_{<i}), \tag{15}$$

which is identical to ancestral AR sampling.

Proof sketch. Condition on the prefix at the start of a round. Every retained proposal equals a draw from its target distribution under maximal coupling. At the first rejection, the residual rule completes exactly the missing target mass, so the replacement also follows the correct conditional target distribution. All later parallel evaluations depend on an invalid prefix and never enter the output. Full acceptance adds a fresh sample from the bonus target. Induction over committed positions yields the AR factorization. Appendix A gives the measure-level calculation. □

Proposition 4 (Overlap controls speculative progress)**.** *Let $a_i = \mathcal{A}(p_i, q_i)$ be the conditional keep probability at event i given survival through event $i-1$. For G_L, the number of consecutive retained proposal events in a window of length L,*

$$\mathbb{E}[G_L] = \sum_{k=1}^{L} \prod_{i=1}^{k} a_i. \tag{16}$$

If another proposal sequence has conditional overlaps $a'_i \geq a_i$ for all i, then its expected retained prefix is no smaller.

Proof. The standard tail-sum argument gives $\mathbb{E}[G_L] = \sum_{k=1}^{L} \Pr[G_L \geq k]$. The event $G_L \geq k$ requires retention at every position through k, so the chain rule gives $\Pr[G_L \geq k] = \prod_{i=1}^{k} a_i$. Coordinate-wise monotonicity follows because every product contains nonnegative factors. □

Proposition 4 explains why the objective in Eq. (11) is operationally aligned with NFE. It also identifies the control structure behind horizon-aware selection: because early-event overlap appears in every subsequent product, gains near the front of the window have disproportionate leverage on expected progress. The per-horizon selector accounts for this asymmetry and includes a calibrated no-transport reference in the exact transport family.

5 Experiments

Our experiments answer four questions. **(Q1)** Does PSC reduce sequential target evaluations and wall-clock latency? **(Q2)** Does it preserve image and video quality within sampling variation? **(Q3)** How do transport strength and horizon-aware selection govern overlap and acceptance within the same exact decoder? **(Q4)** Does the method transfer across visual AR architectures, modalities, and window sizes? Within each table block, bold marks the best displayed value and underline the second distinct value; displayed ties share emphasis.

Table 1: Lumina-mGPT-7B on 5,000 MS-COCO prompts. MC is categorical maximal coupling, GS is shared-Gumbel coupling, and PSC is exact; speedups use the same-window SJD reference.

Config.	Eff.		Speedup		Quality		
	NFE ↓	Latency ↓	NFE ↑	Latency ↑	FID ↓	IS ↑	CLIP ↑
				$L = 16$			
SJD	1055.0490	58.9402	1.00×	1.00×	33.1930	31.1532±1.1099	25.9336
MC	814.5204	52.7612	1.30×	1.12×	**31.7289**	**31.6362±1.5397**	**25.9538**
GS	819.3954	50.0009	1.29×	1.18×	31.9619	31.5027±0.7890	25.8558
PSC	**802.7764**	**46.7576**	**1.31**×	**1.26**×	32.1388	31.0348±0.8241	25.9093
				$L = 32$			
SJD	1057.6252	58.2881	1.00×	1.00×	32.8442	30.1101±1.2750	25.8842
MC	661.3068	43.2528	1.60×	1.35×	**31.8459**	**32.0746±1.1492**	25.9355
GS	651.8996	41.7005	1.62×	1.40×	32.1956	30.9308±0.8013	**25.9656**
PSC	**636.4098**	**39.1320**	**1.66**×	**1.49**×	32.2799	30.9067±1.0273	25.9072
				$L = 64$			
SJD	1055.8714	56.7905	1.00×	1.00×	31.2753	32.7109±1.0229	25.8772
MC	563.6154	41.1820	1.87×	1.38×	**30.1587**	**33.5682±1.9150**	**25.9971**
GS	564.0340	43.8868	1.87×	1.29×	30.7891	32.8412±0.7680	25.9340
PSC	**547.5100**	**36.2669**	**1.93**×	**1.57**×	32.3252	30.7819±0.7434	25.9500

Table 2: Janus-Pro-7B on 5,000 MS-COCO prompts. MC is categorical maximal coupling, GS is shared-Gumbel coupling, and PSC is exact; every row passes the complete-sample gate, and speedups use same-window SJD.

Config.	Eff.		Speedup		Quality		
	NFE ↓	Latency ↓	NFE ↑	Latency ↑	FID ↓	IS ↑	CLIP ↑
				$L = 16$			
SJD	318.6516	11.1109	1.00×	1.00×	**37.0203**	22.5954±1.0942	**23.3929**
MC	**188.4736**	8.7864	**1.69**×	1.26×	37.2567	**22.6936±1.1418**	23.3586
GS	190.9358	**7.6564**	1.67×	**1.45**×	37.1228	22.2324±1.2145	23.2681
PSC	189.4092	7.9814	1.68×	1.39×	37.6869	22.4196±0.8645	23.3001
				$L = 32$			
SJD	313.9960	11.5263	1.00×	1.00×	38.0198	22.2496±0.8734	23.2219
MC	241.4704	9.9538	1.30×	1.16×	**37.1930**	22.3747±0.9996	23.2278
GS	244.0648	10.3663	1.29×	1.11×	38.6733	**22.4388±0.9948**	23.0838
PSC	**152.7910**	**7.0612**	**2.06**×	**1.63**×	37.6683	22.3614±1.0032	**23.3255**
				$L = 64$			
SJD	309.2518	12.1655	1.00×	1.00×	38.2089	22.1830±1.4717	23.2661
MC	141.4574	13.1413	2.19×	0.93×	38.0468	22.0594±0.7278	**23.3322**
GS	144.3756	7.4495	2.14×	1.63×	37.6134	22.8512±1.1319	23.2665
PSC	**140.8402**	**6.6746**	**2.20**×	**1.82**×	**36.9151**	**23.2846±1.1563**	23.2992

5.1 Models, data, and metrics

We evaluate Lumina-mGPT-7B (Liu et al., 2024) and Janus-Pro-7B (Chen et al., 2025) on the same 5,000 MS-COCO validation prompts, ordering, and sampling configuration (Lin et al., 2014). Image quality uses FID, IS, and CLIP score (Heusel et al., 2017; Barratt & Sharma, 2018; Radford et al., 2021). We evaluate Cosmos-AR-4B (Cosmos-1.0-Autoregressive-4B) (Agarwal et al., 2025) on RealEstate10K (Zhou et al., 2018) using FVD, frame FID, and frame IS (Unterthiner et al., 2018). Table 3 reports the available quantitative values by window and coupling rule; because completion counts differ slightly across runs, these video quality values are unpaired. The qualitative appendix uses a separate shared-identity subset. Efficiency uses sequential target-model evaluations (NFE) and end-to-end latency, including proposal construction, correction, cache updates, and required synchronization.

5.2 Baselines and protocol

We compare SJD (Teng et al., 2024), SCD with categorical maximal coupling (MC) (Kumar et al., 2026), and SCD with shared-Gumbel coupling (GS) at $L \in \{16, 32, 64\}$; Cosmos also uses $L = 128$.

Table 3: Cosmos-AR-4B on RealEstate10K. MC is categorical maximal coupling and GS is shared-Gumbel coupling. We report NFE, end-to-end latency, and unpaired frame FID (F-FID). PSC uses full-vocabulary target correction at every reported window.

Metric	$L = 16$				$L = 32$			
	SJD	MC	GS	PSC	SJD	MC	GS	PSC
NFE ↓	1848.51	**1739.52**	2052.15	1740.29	1281.90	**1068.08**	1338.67	1069.16
Latency (s) ↓	67.74	**64.05**	73.27	69.64	49.06	**40.50**	52.06	44.71
F-FID ↓	15.24	15.38	15.36	**15.13**	15.68	**15.35**	**15.35**	15.46
Metric	$L = 64$				$L = 128$			
	SJD	MC	GS	PSC	SJD	MC	GS	PSC
NFE ↓	1015.79	650.20	863.70	**648.79**	1727.60	**562.79**	599.38	563.11
Latency (s) ↓	42.87	**27.91**	39.59	31.00	74.57	25.78	**25.29**	29.67
F-FID ↓	15.18	15.51	15.15	**15.09**	15.19	15.44	**14.97**	15.67

Table 4: Janus-Pro-7B mechanism ablation on MS-COCO at $L = 64$. Vec., PDT, and Adapt. denote vectorized execution, predictive transport, and adaptive calibration. All rows satisfy the 5,000-prompt completion gate and use exact target correction.

M	Variant	Vec.	PDT	Adapt.	NFE ↓	Latency ↓	FID ↓	IS ↑	CLIP ↑
1	Scalar MC	–	–	–	141.4574	13.8091	38.0468	22.0594±0.7278	23.3322
2	Vectorized MC	✓	–	–	**140.8352**	6.8262	36.9073	**23.2990±1.1741**	23.3015
	Fixed $\beta = 0.25$	✓	✓	–	161.7700	7.1203	37.5359	22.6362±0.9863	23.3003
3	Fixed $\beta = 0.5$	✓	✓	–	176.4456	7.3576	36.8809	22.7813±1.2486	**23.3479**
	Fixed $\beta = 1$	✓	✓	–	284.3516	11.8181	**36.7803**	22.8533±0.6603	23.2959
4	PSC	✓	✓	✓	140.8402	**6.2766**	36.9151	23.2846±1.1563	23.2992

The default predictor uses $\mathcal{B} = \{0, 0.25, 0.5, 1.0\}$, drift clip $c = 2$, momentum 0.8, risk weight $\lambda = 1$, and two warm-up observations, without per-prompt tuning. The paired no-transport control shares the PSC implementation, random-number policy, and cache logic, enabling separate attribution of vectorized decoder effects and transport-policy effects. Runs record candidate overlap, selected-β counts, transport and target acceptance, NFE, and latency. For Cosmos, every PSC row uses full-vocabulary target correction and therefore falls under the distribution-preservation guarantee.

5.3 Main results

Tables 1–3 report all evaluated windows. The image results use 5,000 aligned prompts, whereas Cosmos quality uses the completed videos from each run and is therefore unpaired.

Efficiency across windows. On Lumina-mGPT-7B, the exact PSC path lowers NFE and latency relative to SJD at every evaluated window; at $L = 64$, NFE falls from 1055.87 to 547.51 and latency from 56.79 to 36.27 seconds, reductions of 48.1% and 36.1%. On Janus-Pro-7B, the corresponding endpoint reduces NFE from 309.25 to 140.84 and latency from 12.17 to 6.67 seconds, reductions of 54.5% and 45.1%. Across both image models, the FID, IS, and CLIP differences are descriptive rather than evidence of non-inferiority. On Cosmos-AR-4B, PSC improves over SJD through $L = 64$, where it reaches 648.79 NFE and 31.00 seconds, reductions of 36.1% and 27.7%.

Qualitative comparison. Figure 3 places one aligned prompt per image model for Janus-Pro-7B and Lumina-mGPT-7B in the upper strip and two time points from Cosmos-AR-4B in the lower strip. The examples cover object and outdoor scenes while exposing per-sample NFE; they provide qualitative evidence only. Additional aligned prompts and full video timelines are provided in Appendix E.

Exact operating points. Exactness follows from full-vocabulary target correction and valid-prefix commit, not from empirical quality agreement. The image and Cosmos PSC rows use this exact path. Every reported PSC row uses full-vocabulary correction and is distribution preserving, while the video-quality measurements are unpaired. At the longest reported window, Cosmos PSC reduces NFE and latency by 67.4% and 60.2% relative to SJD, corresponding to $3.07\times$ and $2.51\times$ speedups.

Table 5: Janus-Pro-7B selector ablation on MS-COCO at $L = 64$. Horizon, Risk, and Gain denote horizon conditioning, confidence-risk calibration, and realized-gain contraction. All rows satisfy the 5,000-prompt completion gate; PSC reuses the result in Table 4.

S	Variant	Horizon	Risk	Gain	NFE ↓	Latency ↓	FID ↓	IS ↑	CLIP ↑
					$L = 64$				
1	Global statistics	–	✓	✓	**140.8356**	6.4773	**36.9083**	23.2954±1.1708	**23.3012**
2	Risk-neutral	✓	–	✓	141.0596	7.8692	37.0714	**23.4927±1.0560**	23.2828
3	No gain contraction	✓	✓	–	140.8504	6.6130	36.9212	23.2807±1.1597	23.2988
4	PSC	✓	✓	✓	140.8402	**6.2766**	36.9151	23.2846±1.1563	23.2992

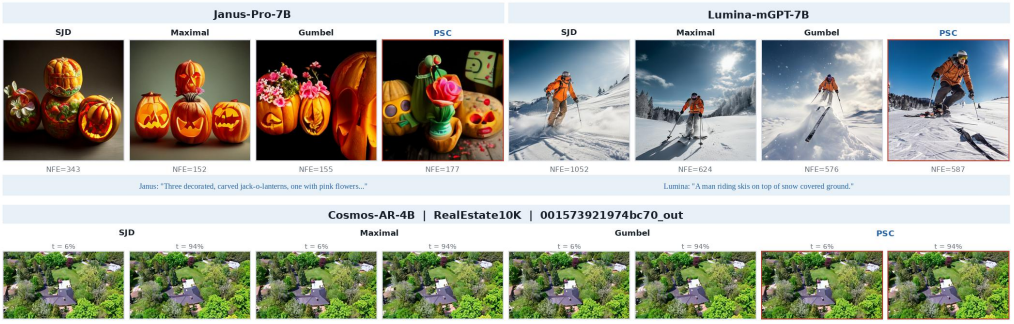


Figure 3: Compact cross-model qualitative comparison at $L = 64$. The upper row shows one aligned COCO prompt per image model under SJD, maximal coupling, Gumbel coupling, and PSC for Janus-Pro-7B (left) and Lumina-mGPT-7B (right); the lower row shows two normalized time points for one RealEstate10K identity under Cosmos-AR-4B. NFE labels report per-sample target evaluations. These examples illustrate qualitative behavior and are not used to make an aggregate quality claim.

Where the measured gain comes from. Tables 4 and 5 report the $L = 64$ ablations that separate decoder execution, transport strength, and online selection. Vectorization changes NFE by only 0.4%, from 141.4574 to 140.8352, while reducing latency from 13.8091 s to 6.8262 s. The fixed-β sweep identifies transport strength as a high-leverage variable: NFE rises to 284.3516 at $\beta = 1$, whereas the adaptive PSC endpoint remains near the vectorized reference. The corresponding $L = 32$ measurements are provided in the Appendix tables 6 and 7. Together with Proposition 4, this sensitivity motivates selecting transport from horizon-level evidence.

Component sensitivity. The global-statistics, risk-neutral, no-gain-contraction, and PSC selector variants remain tightly clustered at $L = 64$. These comparisons characterize controller sensitivity within a shared exact-correction framework. Absolute latency and small differences between controller paths remain sensitive to hardware and run-to-run load, while exactness follows from the correction construction and is independent of selector choice.

6 Conclusion

Conclusion. PSC reframes speculative visual decoding as exact proposal transport: prediction optimizes the intermediate distribution, and maximal residual correction preserves the target AR law independently of the selected transport. Across image and video models, PSC preserves the target distribution and reduces sequential target evaluations and end-to-end latency relative to SJD. The method is training-free, and all reported PSC rows use the same exact correction interface.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, et al. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024. doi: 10.48550/arXiv.2401.10774.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Zhuoming Chen, Avner May, Ruslan Svirschevski, Yuhsun Huang, Max Ryabinin, et al. Sequoia: Scalable, robust, and hardware-aware speculative decoding. *arXiv preprint arXiv:2402.12374*, 2024. doi: 10.48550/arXiv.2402.12374.

Xin Cheng, Xingkai Yu, Chenze Shao, Jiashi Li, Yunfan Xiong, Yi Qian, Jiaqi Zhu, Shirong Ma, Xiaokang Zhang, Jiasheng Ye, et al. Dspark: Confidence-scheduled speculative decoding with semi-autoregressive generation. *arXiv preprint arXiv:2607.05147*, 2026.

Mohamed Elhoushi, Akshat Shrivastava, Faris Ramadan, et al. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.681.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

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. *Advances in neural information processing systems*, 30, 2017.

Kaixuan Huang, Xudong Guo, and Mengdi Wang. Specdec++: Boosting speculative decoding via adaptive candidate lengths. *arXiv preprint arXiv:2405.19715*, 2024.

Tanishq Kumar, Tri Dao, and Avner May. Speculative speculative decoding. *arXiv preprint arXiv:2603.03251*, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International conference on machine learning*, pp. 19274–19286. PMLR, 2023.

Yuhui Li, Fangyun Wei, Chao Zhang, Hongyang Zhang, et al. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024a. doi: 10.48550/arXiv.2401.15077.

Yuhui Li, Fangyun Wei, Chao Zhang, Hongyang Zhang, et al. Eagle-2: Faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024b. doi: 10.18653/v1/2024.emnlp-main.422.

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*, pp. 740–755. Springer, 2014.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yi Xin, Xinyue Li, Qi Qin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.

Xiaoxuan Liu, Lanxiang Hu, Peter Bailis, Alvin Cheung, Zhijie Deng, Ion Stoica, and Hao Zhang. Online speculative decoding. *arXiv preprint arXiv:2310.07177*, 2023.

Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Zhengxin Zhang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, et al. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. *arXiv preprint arXiv:2305.09781*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Junhyuk So, Hyunho Kook, Chaeyeon Jang, and Eunhyeok Park. Speculative coupled decoding for training-free lossless acceleration of autoregressive visual generation. *arXiv preprint arXiv:2510.24211*, 2025.

Yang Song, Chenlin Meng, Renjie Liao, and Stefano Ermon. Accelerating feedforward computation via parallel nonlinear equation solving. In *International Conference on Machine Learning*, pp. 9791–9800. PMLR, 2021.

Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*, 2024.

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

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Jinheng Xie, Weijia Mao, et al. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. doi: 10.48550/arXiv.2408.12528.

Lijun Yu, José Luis Lezama, Nitesh B. Gundavarapu, Luca Versari, Kihyuk Sohn, et al. Language model beats diffusion – tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. doi: 10.48550/arXiv.2310.05737.

Yuxin Zhang, Yijia Gao, et al. Draft&verify: Lossless large language model acceleration via self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.acl-long.607.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

A Proof Details

A.1 Proof of Lemma 1

Fix $y \in \mathcal{V}$. The keep branch contributes

$$\Pr[Y = y,\, Y = X] = r(y)\min\{1, s(y)/r(y)\} = \min\{r(y), s(y)\}. \tag{17}$$

The total rejection mass is

$$1 - \sum_v \min\{r(v), s(v)\} = \sum_v [s(v) - r(v)]_+. \tag{18}$$

Therefore, the residual branch contributes $[s(y) - r(y)]_+$ and

$$\Pr[Y = y] = \min\{r(y), s(y)\} + [s(y) - r(y)]_+ = s(y). \tag{19}$$

Summing the keep contribution over y gives $\Pr[Y = X] = \sum_y \min\{r(y), s(y)\} = 1 - \mathrm{TV}(r, s)$. The coupling inequality gives $\Pr[X \neq Y] \geq \mathrm{TV}(r, s)$ for every coupling, so this equality probability is maximal (Lindvall, 2002).

A.2 Proof of Theorem 3

Let C denote the committed prefix before a verification round. For the first speculative event, Theorem 2 gives the target conditional $p_\theta(\cdot \mid C)$. If the event is retained, its realized value equals the value used to form the next parallel target distribution; hence the second event uses the valid conditional $p_\theta(\cdot \mid C, X_1)$. This argument repeats through every retained event. At the first rejected event, the positive-residual branch supplies the missing target mass and produces the same conditional target distribution. No event after this correction is committed. If every event is retained, the bonus draw uses the final valid prefix and follows its target conditional.

Suppose the decoder produces the correct AR law through position $m-1$. Conditioning on any prefix $x_{<m}$ with nonzero probability, the round argument above gives

$$\Pr[X_m = x_m \mid X_{<m} = x_{<m}] = p_\theta(x_m \mid x_{<m}). \tag{20}$$

Induction and the chain rule give the complete factorization in Theorem 3. The predictor state and selected β can depend on prior completed rounds because they only change the intermediate proposal; Theorem 2 holds conditionally on any such state.

B Implementation Details

Numerical support. The implementation normalizes nonnegative probabilities in float precision and clips log drift only on support shared by current and historical proposals. When history is absent or support is not shared, the drift is zero. The maximal-coupling residual is $[s-r]_+$, and a sampled rejection with zero residual mass raises an error rather than substituting an inexact token.

Alignment. History uses absolute generated-token positions instead of tensor offsets. Each speculative commit changes the offset of the remaining block, while absolute positions remain stable. The protected anchor never enters proposal prediction or target correction.

Controller. The default candidate set is $(0, 0.25, 0.5, 1.0)$. Statistics use exponential momentum 0.8, and selection maximizes mean gain minus one standard deviation after two observations. One negative observation moves the selected index toward zero; a repeated negative streak returns it to zero.

C Ablation Details

The main paper reports the $L = 64$ ablations for a single-window comparison. The corresponding $L = 32$ measurements are provided here with the same metric ordering and emphasis convention.

Table 6: Janus-Pro-7B mechanism ablation at $L = 32$. Vec., PDT, and Adapt. denote vectorized execution, predictive transport, and adaptive calibration. All rows satisfy the 5,000-prompt completion gate and use exact target correction.

M	Variant	Vec.	PDT	Adapt.	NFE ↓	Latency ↓	FID ↓	IS ↑	CLIP ↑
1	Scalar MC	–	–	–	153.3814	9.5976	37.5607	22.8628±1.0373	23.2891
2	Vectorized MC	✓	–	–	152.8016	**6.1779**	37.6608	22.3619±0.9844	23.3274
	Fixed $\beta = 0.25$	✓	✓	–	169.6120	6.8039	**36.7606**	**23.1017±1.2663**	23.3198
3	Fixed $\beta = 0.5$	✓	✓	–	185.3300	7.3416	37.4139	22.5130±1.2587	**23.3927**
	Fixed $\beta = 1$	✓	✓	–	273.7844	10.6121	37.5516	22.9243±1.0821	23.2846
4	PSC	✓	✓	✓	**152.7910**	6.2568	37.6683	22.3614±1.0032	23.3255

Table 7: Janus-Pro-7B selector ablation at $L = 32$. Horizon, Risk, and Gain denote horizon conditioning, confidence-risk calibration, and realized-gain contraction. All rows satisfy the 5,000-prompt completion gate.

S	Variant	Horizon	Risk	Gain	NFE ↓	Latency ↓	FID ↓	IS ↑	CLIP ↑
1	Global statistics	–	✓	✓	152.7996	6.4173	**37.6650**	**22.3681±0.9816**	**23.3267**
2	Risk-neutral	✓	–	✓	152.9436	6.3360	37.6736	22.3035±0.8055	23.3209
3	No gain contraction	✓	✓	–	152.7916	6.3198	37.6742	22.3566±1.0044	23.3248
4	PSC	✓	✓	✓	**152.7910**	**6.2568**	37.6683	22.3614±1.0032	23.3255

D PARTIPROMPTS GENERALIZATION

We additionally evaluate Janus-Pro-7B on the complete 1,632-prompt PartiPrompts set under the same $L = 64$ decoding configuration for all four methods. Table 8 reports aggregate efficiency and quality measurements, and Figure 4 shows a fixed aligned subset of the generated outputs.

Table 8: Janus-Pro-7B on all 1,632 PartiPrompts prompts at $L = 64$. Speedups use SJD as the reference. PartiPrompts provides no real reference images, so FID is unavailable and is not reported. IS and CLIP differences are descriptive.

Method	NFE ↓	Latency (s) ↓	NFE speedup ↑	Latency speedup ↑	IS ↑	CLIP ↑
SJD	302.77	11.63	1.00×	1.00×	14.93±1.23	<u>23.48</u>
MC	<u>136.77</u>	13.20	<u>2.21×</u>	0.88×	**15.11±0.58**	**23.54**
GS	138.76	<u>6.95</u>	2.18×	<u>1.67×</u>	<u>15.02±1.45</u>	<u>23.48</u>
PSC	**136.02**	**5.94**	**2.23×**	**1.96×**	14.70±1.06	23.32

PSC reaches $2.23\times$ NFE and $1.96\times$ latency speedups relative to SJD on this prompt suite. The IS and CLIP values are reported as descriptive checks across independently coupled samples rather than as evidence of a statistically significant quality difference.

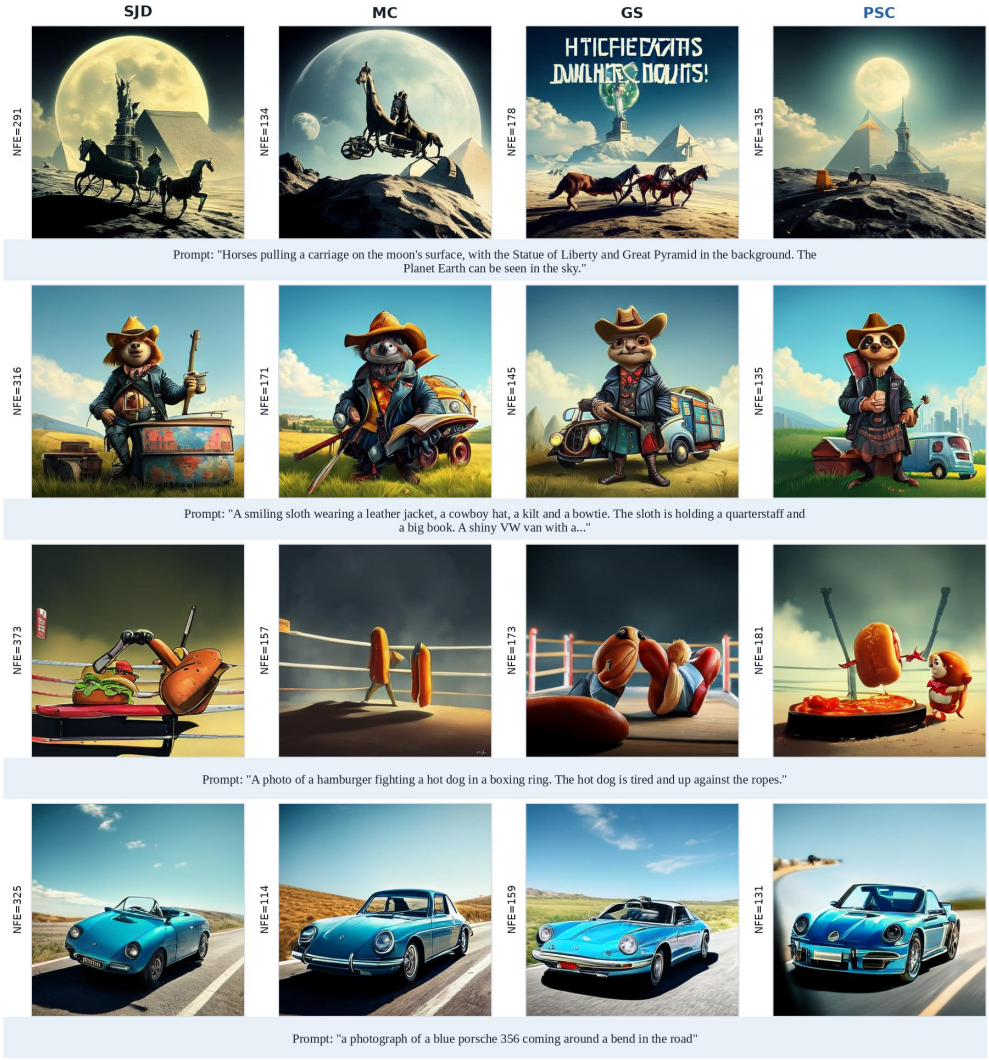


Figure 4: Janus-Pro-7B comparison on four fixed aligned PartiPrompts prompts (IDs 275, 301, 332, and 423) at $L = 64$. Each row keeps the prompt fixed across SJD, maximal coupling (MC), shared-Gumbel coupling (GS), and PSC; vertical labels report per-sample NFE. These examples illustrate output diversity and fidelity but do not establish an aggregate quality claim.

E QUALITATIVE COMPARISONS

All qualitative panels below use $L = 64$, with the same prompt or video identity shared across the four methods in each row. All method panels use neutral borders under aligned source conditions, while the blue "PSC" label identifies our method; vertical labels report per-sample NFE in the image grids. The image prompts form a fixed aligned subset of the common four-method output intersection. The figures provide visual evidence of output fidelity and are not used to make a statistical quality claim.

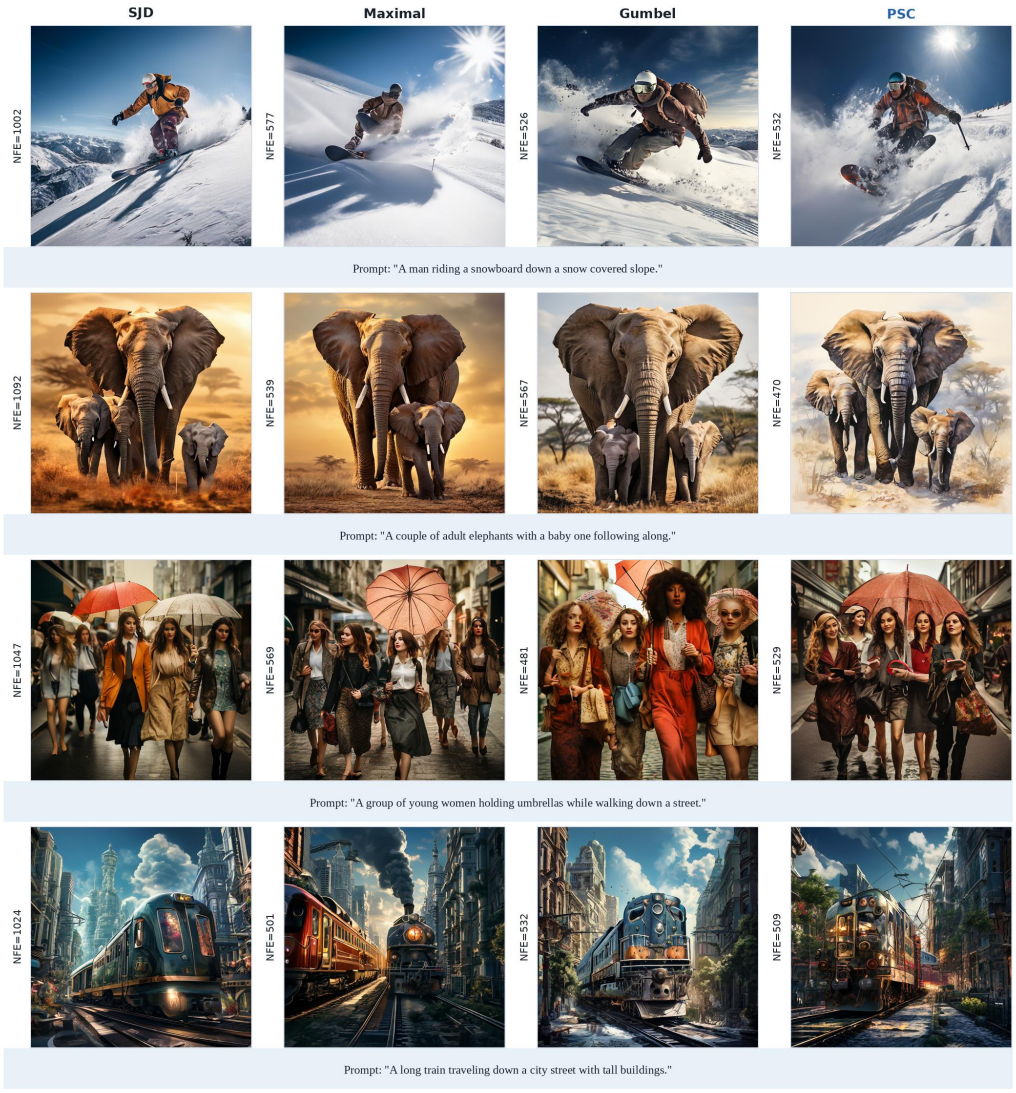


Figure 5: Lumina-mGPT-7B comparison on four fixed COCO prompts; each row changes only the coupling rule.

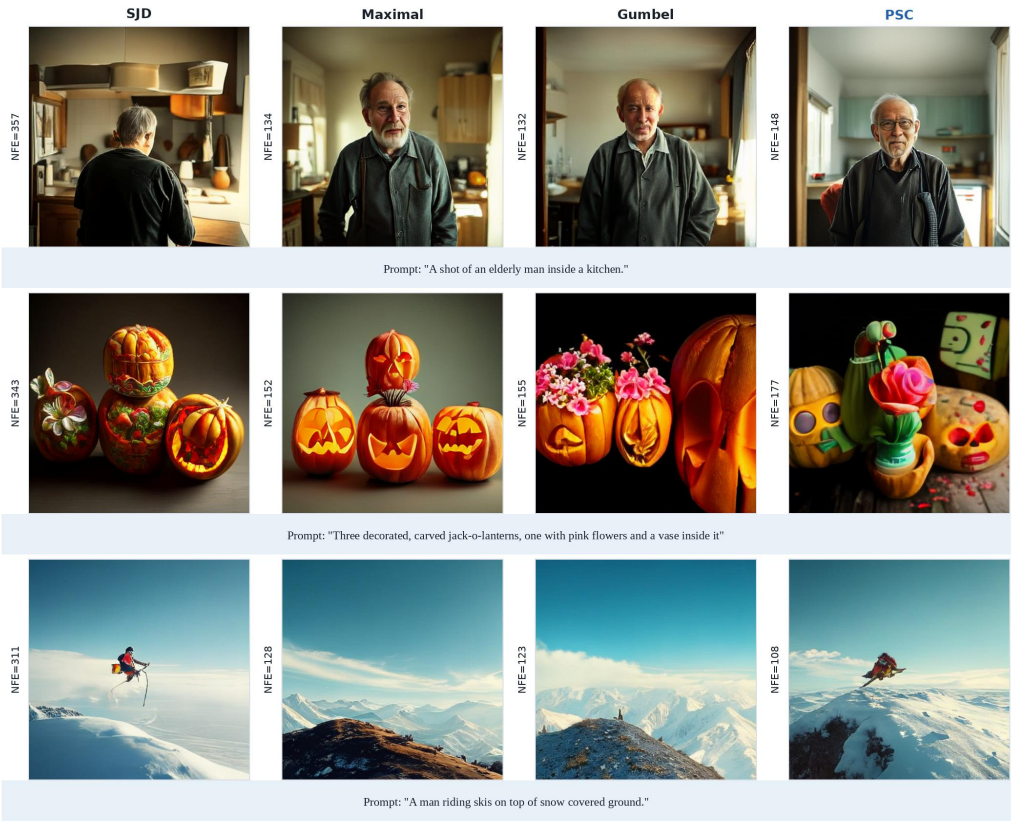


Figure 6: Additional Janus-Pro-7B comparison on three aligned COCO prompts.

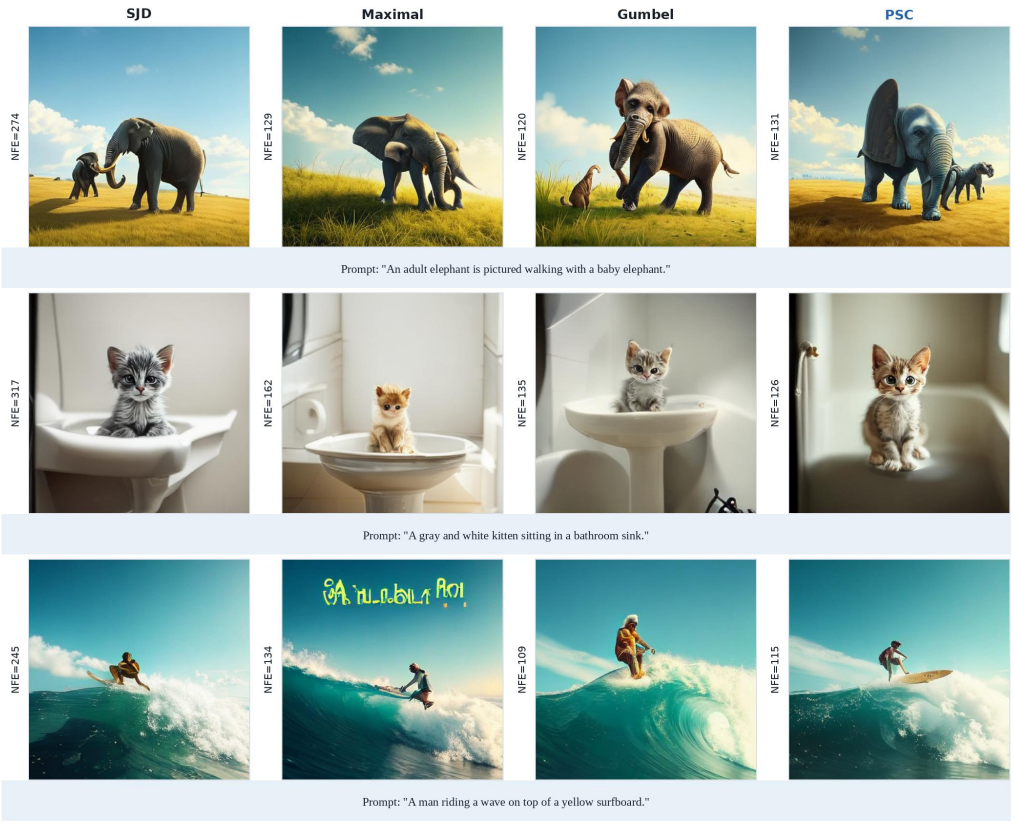


Figure 7: Additional Janus-Pro-7B comparison on three aligned COCO prompts.

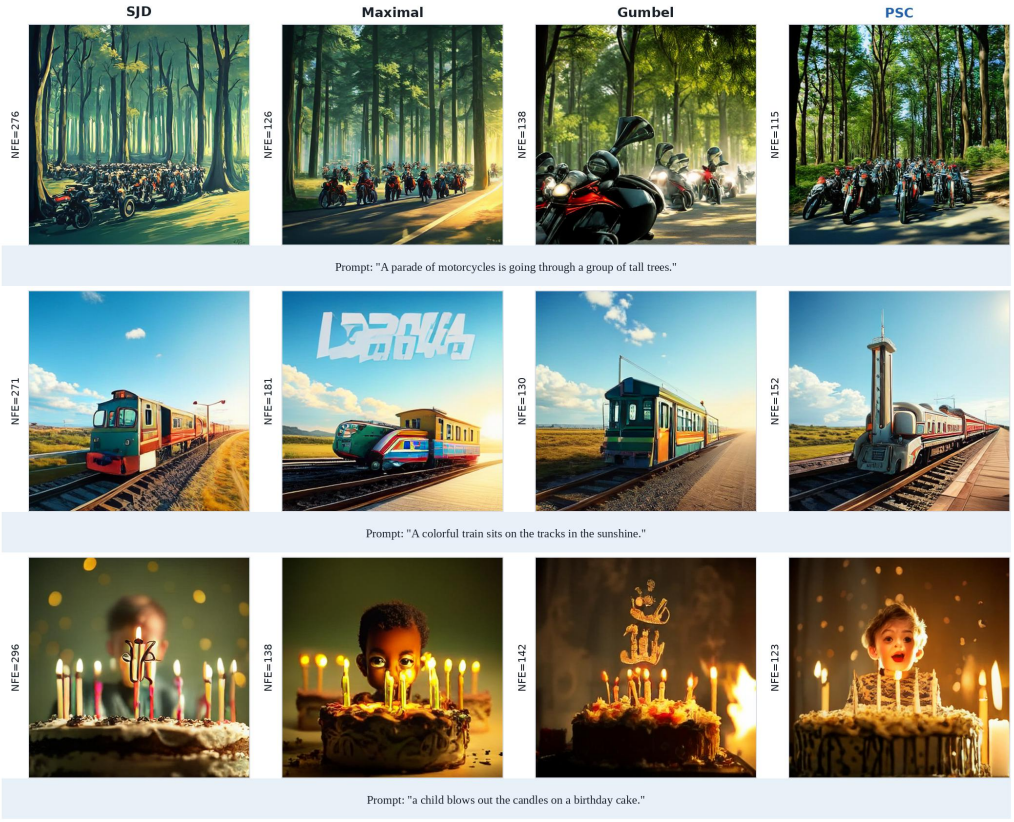


Figure 8: Additional Janus-Pro-7B comparison on three aligned COCO prompts.

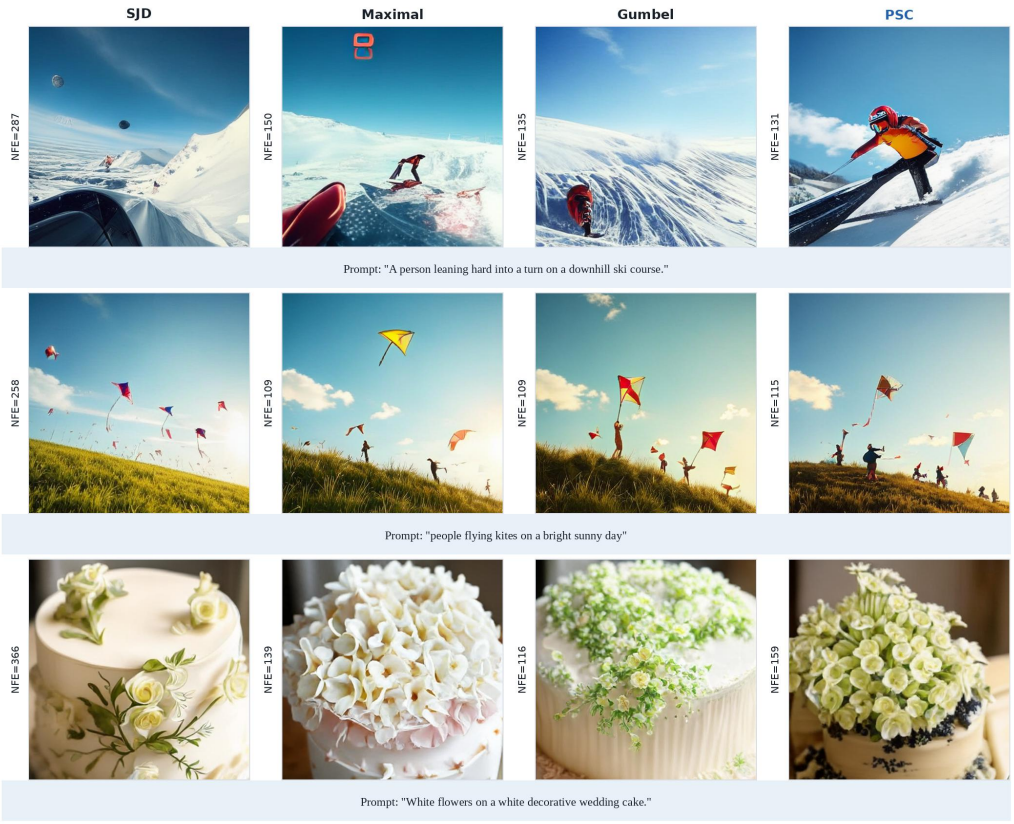


Figure 9: Additional Janus-Pro-7B comparison on three aligned COCO prompts.

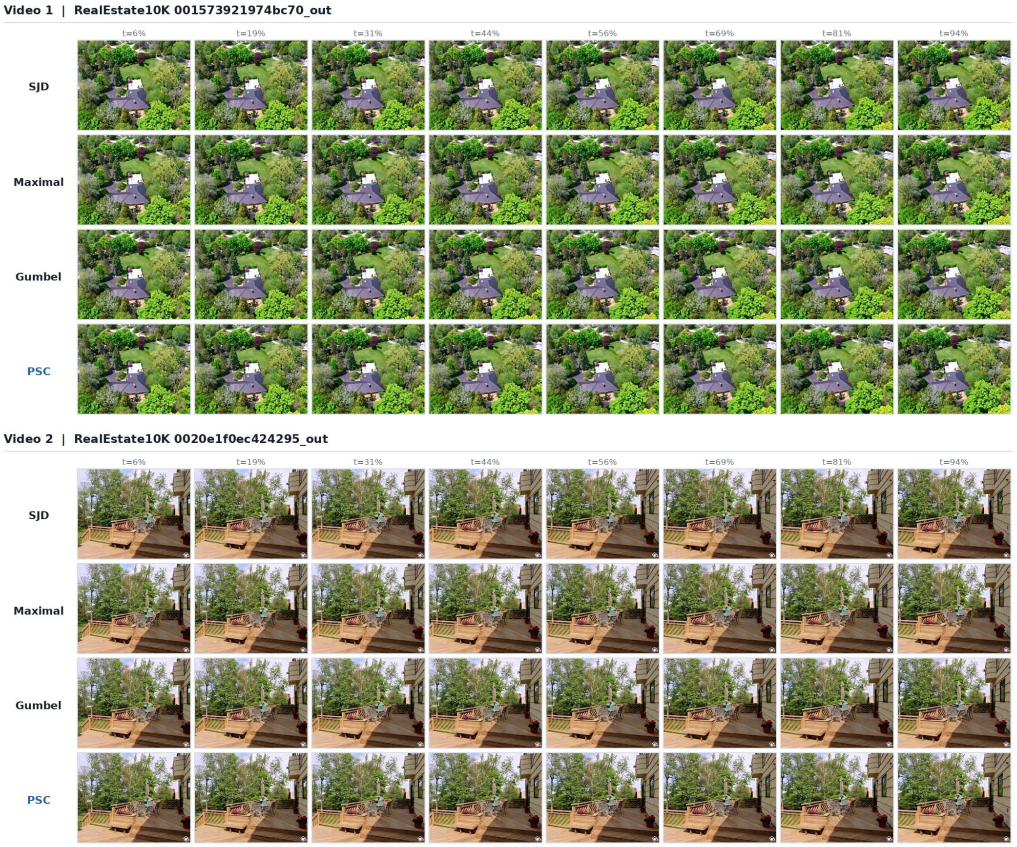


Figure 10: Cosmos-AR-4B comparison for two RealEstate10K conditioning identities at eight normalized time points.