Dynamic Causal Structure Discovery for Autoregressive Visual Generation

Songliang Guo[1*] **Qi Yan**[1*] **Jianzhou Wang**[3] **Xinfu Liu**[1]
Cheng Zhen[1] **Yirui Wu**[2†] **Lixin Yuan**[1] **Wenxiao Zhang**[1] **Jun Liu**[3]
[1] College of Computer Science and Software Engineering,
Hohai University, Nanjing, China
[2] College of Computing and Artificial Intelligence, United Arab Emirates University,
Al Ain, Abu Dhabi, United Arab Emirates
[3] School of Computing and Communications, Lancaster University, Lancaster, UK.
{guosongliang,yanqi}@hhu.edu.cn wu.yirui@uaeu.ac.ae

Abstract

Autoregressive (AR) models achieve strong performance in visual generation, yet their prefix-based factorization imposes a rigid one-dimensional dependency structure. We show that this structure is fundamentally suboptimal: any predefined ordering necessarily induces a trade-off between contextual sufficiency and necessity, resulting in both missing and redundant dependencies. To address this limitation, we propose **Dynamic Causal Structure Discovery**, an inference-time framework that replaces static conditioning with dynamic, content-dependent causal parent sets. By interpreting the Transformer as a causal graph, we reformulate generation as the problem of constructing a minimal parent set for each token. To make this objective tractable, we approximate do-interventions as vector-space ablations and derive a Fisher-based criterion that characterizes and corrects structural deviations during decoding. Our framework transforms AR factorization into an order-agnostic dependency structure, yielding improved generation quality and global coherence without retraining. Our code is available at `https://github.com/Songliang112/DCD`.

1 Introduction

Autoregressive (AR) models have emerged as a dominant paradigm for visual generation [1, 2, 3, 4], due to their architectural simplicity and the scalability of likelihood-based training (NLL) [5, 6, 7]. Beyond their probabilistic formulation, however, AR models implicitly impose a *structural prior*: token dependencies are constrained to a one-dimensional, strictly causal computational graph. This sequential factorization is not merely an implementation choice—it fundamentally governs how information is propagated during generation.

Unlike text, visual data exhibits dense, bidirectional, and content-dependent interactions[8, 9]. As a result, the prefix-based structure induced by AR factorization is inherently misaligned with the intrinsic dependency topology of images and video [5, 10]. In this work, we formalize this mismatch and show that any predefined ordering necessarily induces a structural trade-off. Specifically, it violates two complementary principles: *sufficiency*, by excluding relevant context outside the prefix, and *necessity*, by conditioning on redundant variables within the prefix [11]. We further prove that standard AR factorization cannot satisfy these properties simultaneously, leading to what we term

[*] Equal contribution.
[†] Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

structural suboptimality. Empirically, this manifests as degraded global coherence and accumulated generation errors in visual generation [6].

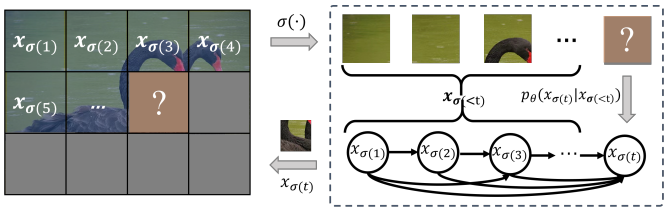
Recent efforts, including order randomization [2, 1], bidirectional attention adaptations [12, 13], and iterative refinement [7, 14, 15], improve performance in practice but largely retain the same prefix-based conditioning scheme. As a result, they mitigate empirical deficiencies without addressing the underlying structural constraint. To overcome this limitation, we propose **Dynamic Causal Structure Discovery**, a framework that replaces static, order-driven conditioning with dynamic, content-dependent dependency selection at inference time. Our approach builds on the causal framework of Halpern-Pearl[16, 17], reinterpreting the Transformer as a mechanistic causal graph and targeting a *minimal parent set* for each token. More details on motivation see Appendix(§B).

To make this objective tractable, we establish a correspondence between causal interventions and model computation. Specifically, we show that do-interventions can be approximated as vector-space ablations in the Transformer residual stream[18], whose effects admit a second-order characterization via the Fisher information matrix [19, 20]. This yields an efficient procedure that detects missing dependencies, prunes redundant context, and refines predictions during decoding. Overall, our framework transforms the fixed AR factorization into a dynamically constructed, order-agnostic dependency structure, bringing the model's computation into alignment with the data's minimal dependency structure. Our main contributions are three-fold:

- **Theoretical Formalization:** We show that prefix-based factorization cannot satisfy sufficiency and necessity simultaneously, implying structural suboptimality.
- **Causal Mapping:** We approximate do-interventions via a Fisher-based criterion for efficient causal evaluation in Transformers.
- **Dynamic Discovery:** We develop an inference-time algorithm that constructs minimal parent sets without retraining.

2 Causal Foundations of Autoregressive Modeling

As shown in Figure 1, we formalize autoregressive (AR) visual generation as a **structural causal process**, in which sequential factorization induces an explicit dependency graph over tokens. Under this view, AR is not merely a statistical decomposition, but a **mechanistic computational graph** where each prediction is functionally determined by a fixed parent set.



Visual Generation at Step t AR Generation as a Causal Graph

Figure 1: AR visual generation as a causal graph. Under a generation order σ, each token is predicted from previously generated tokens, inducing an explicit dependency structure.

Definition 1 (AR Visual Generation)**.** *Given a sequence of discrete visual tokens* $\mathbf{x} = \{x_{\sigma(1)}, \ldots, x_{\sigma(T)}\}$*, an AR model factorizes the joint distribution as:*

$$p_\theta(\mathbf{x}) \;=\; \prod_{t=1}^{T} p_\theta\big(x_{\sigma(t)} \;\big|\; x_{\sigma(<t)}\big), \tag{1}$$

where σ *is a permutation over* $\{1, \ldots, T\}$ *defining the generation order. The order* σ *may be* ***fixed*** *(e.g., raster scan [3]) or* ***randomized*** *per sample [21]. For brevity, we use* x_t*,* $x_{<t}$ *and* $x_{>t}$ *in place of* $x_{\sigma(t)}$*,* $x_{\sigma(<t)}$ *and* $x_{\sigma(>t)}$ *under the order* σ*.*

Definition 2 (AR Generation as a Causal Graph)**.** *Fix a generation order* σ*, autoregressive generation induces a directed graph* $\mathcal{G}_\sigma = (\mathcal{V}, \mathcal{E}_\sigma)$ *with node set* $\mathcal{V} = \{x_1, \ldots, x_T\}$ *and edge set* $\mathcal{E}_\sigma$*, uniquely determined by the following properties:*

(2.a) Order-Induced Acyclicity. $\mathcal{G}_\sigma$ *is a directed acyclic graph consistent with* σ*, i.e.,*

$$(x_j \to x_k) \in \mathcal{E}_\sigma \implies j < k. \tag{2}$$

(2.b) Parent Set Determinism. *For each node $x_{\sigma(t)}$, the parent set is strictly fixed as*

$$P(x_t) = x_{<t}. \tag{3}$$

Consequently, under p_θ, $x_{\sigma(t)}$ is conditionally independent of all non-descendants given $P(x_{\sigma(t)})$.

(2.c) Mechanistic Realizability. *Every edge $x_j \to x_k \in \mathcal{E}_\sigma$ corresponds to an explicit computational pathway in the model. In Transformer architectures, this constraint is enforced by the attention mask, which strictly restricts information flow to the parent set $P(x_t)$ through the residual stream.*

While $\mathcal{G}_\sigma$ is a valid causal graph by construction [22], its parent sets are entirely determined by the imposed sequentiality σ, rather than the underlying dependency structure of the data. As a result, the AR factorization enforces a **structural prior** over the dependency topology that is inherently misaligned with intrinsic dependencies. This mismatch necessarily induces inefficiencies in context utilization, which we formalize next.

3 Beyond Sequentiality: A Causal Perspective on AR Generation

Since sequential AR factorization imposes a one-dimensional dependency structure that is often misaligned with the bidirectional nature of visual data [23], its conditioning structure is generally suboptimal for prediction. We thus analyze this mismatch from a causal perspective, identify its key limitations (§3.1), and formalize the target dependency structure that resolves them (§3.2).

3.1 Structural Limitations of Sequential Factorization

Given Definition 2, the parent set $P(x_t)$ is entirely determined by the generation order σ, independent of the intrinsic dependency structure of the data. This mismatch leads to violations of both *sufficiency* and *necessity* in the conditioning set. Throughout this section, $p^*(\mathbf{x})$ denote the oracle data-generating distribution of visual tokens, and $I_{p^*}(\cdot;\cdot \mid \cdot)$ denote conditional mutual information under p^*. Information-theoretic quantities, including conditional entropy, conditional KL divergence, and conditional mutual information, are summarized in Appendix C.3.

Limitation I: Violation of Sufficiency. Under AR factorization, the parent set excludes all future variables $x_{>t}$, regardless of their relevance to x_t. The resulting information deficit is quantified by

$$\mathcal{S}_t := I_{p^*}(x_t;\, x_{>t} \mid x_{<t}), \tag{4}$$

which measures the predictive information contained in the excluded variables. Whenever $\mathcal{S}_t > 0$, the AR parent set is strictly insufficient.

Limitation II: Violation of Necessity. AR conditions on the entire prefix $x_{<t}$, which may include redundant or weakly relevant subsets $R_t \subset x_{<t}$. Their marginal contribution is quantified by

$$\mathcal{N}_t(R_t) := I_{p^*}(x_t;\, R_t \mid x_{<t} \setminus R_t), \tag{5}$$

which measures the additional information provided by R_t after the remaining prefix is already observed. When $\mathcal{N}_t(R_t) \le \epsilon$, these variables are effectively unnecessary, violating the necessity requirement of a minimal parent set.

The above limitations produce measurable prediction shifts. Starting from a fixed set of $K{=}20$ causal parents, we progressively inject additional tokens and measure the change in next-token NLL. As shown in Figure 2, adding causal parents consistently decreases NLL, revealing missing informative context beyond the AR prefix. In contrast, adding low-causal tokens often increases NLL, especially for spatially nearby but weakly relevant tokens, indicating harmful redundant conditioning. These results directly validate the structural suboptimality of AR factorization. This suggests AR prefix is **structurally suboptimal** in two directions: it may miss useful dependencies while retaining unnecessary ones.

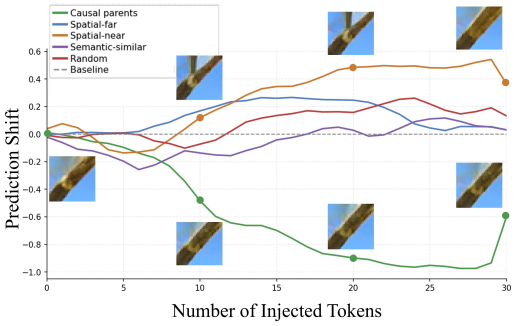


Figure 2: Prediction shift under different context additions.

Structural Consequence. These two limitations arise jointly from the imposed ordering. The following theorem formalizes the resulting suboptimality of AR conditioning.

Theorem 1 (Structural Suboptimality of AR)**.** *Let $P(x_t) = x_{<t}$ denote the AR parent set, and let $S_t^\star$ be a minimal parent set as in Definition 3. If either*

$$\mathcal{S}_t > 0 \quad \textit{or} \quad \exists\, R_t \subset x_{<t} \textit{ such that } \mathcal{N}_t(R_t) \leq \epsilon, \tag{6}$$

then the AR parent set is structurally inconsistent with the minimal parent set, i.e.,

$$x_{<t} \neq S_t^\star.$$

Consequently, under model consistency with the oracle conditional structure,

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{<t}, c) \,\big\|\, p_\theta(x_t \mid x_{S_t^\star}, c)\big) > 0. \tag{7}$$

Theorem 1 shows that AR factorization becomes structurally suboptimal whenever its parent set violates either sufficiency or necessity. The proof is provided in Appendix D.2 under the well-specification condition in Assumption 2. This observation motivates the search for a dependency structure that explicitly satisfies both properties.

3.2 The Minimal Causal Structure

To resolve these limitations, we define a target dependency structure that is both sufficient and necessary, using interventional effects on the target computation to assess causal relevance. Additional predictive information from future tokens identifies candidates for this assessment, rather than establishing causal parenthood.

Definition 3 (Minimal Parent Set)**.** *Let $\mathbf{x}_{-t} := \mathbf{x} \setminus \{x_t\}$ denote all tokens other than x_t. Given a sufficiency tolerance $\epsilon > 0$ and a necessity margin η, a parent set $S_t^\star \subseteq \mathbf{x}_{-t}$ is called minimal if it satisfies:*

(i) Sufficiency. *For any subset $U \subseteq \mathbf{x}_{-t} \setminus S_t^\star$ and any intervention value z_U, the induced distributional shift is negligible:*

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{S_t^\star}) \,\big\|\, p_\theta(x_t \mid x_{S_t^\star},\, \mathrm{do}(U = z_U))\big) \leq \epsilon. \tag{8}$$

(ii) Necessity. *For each variable $v \in S_t^\star$, there exists an intervention value z_v that induces a non-trivial shift:*

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{S_t^\star}) \,\big\|\, p_\theta(x_t \mid x_{S_t^\star \setminus \{v\}},\, \mathrm{do}(v = z_v))\big) \geq \eta. \tag{9}$$

(iii) Order-Agnosticity. *Membership in $S_t^\star$ depends only on causal relevance to x_t, independent of the generation order σ.*

This definition directly enforces the structural requirements from §3.1: sufficiency ensures contextual completeness, while necessity removes redundant dependencies. Together they characterize $S_t^\star$ as the minimal conditioning set fully capturing the predictive structure of x_t.

Remark (Near-Uniqueness). Under thresholds (ϵ, η), any two minimal parent sets must agree up to variables whose contributions lie within the tolerance band. Thus, $S_t^\star$ is identifiable up to ϵ-insensitive components.

From Definition to Computation. The interventional conditions in Definition 3 are not directly tractable during inference. In the next section, we show how these criteria can be approximated within the Transformer computational graph via vector-space ablations and Fisher-weighted metrics, enabling efficient identification of $S_t^\star$.

4 Methodology

Motivated by the structural mismatch identified in §3, we propose **Dynamic Causal Structure Discovery**, a framework that approximates the minimal parent set $S_t^\star$ during inference. In Figure 3, our method consists of **Causal Surrogate** (§4.1), which maps do-interventions onto the Transformer computational graph to derive a tractable Fisher-based criterion, and **Inferential Refinement** (§4.2), which dynamically detects and refines dependencies during decoding. Notation follows Appendix C.1.

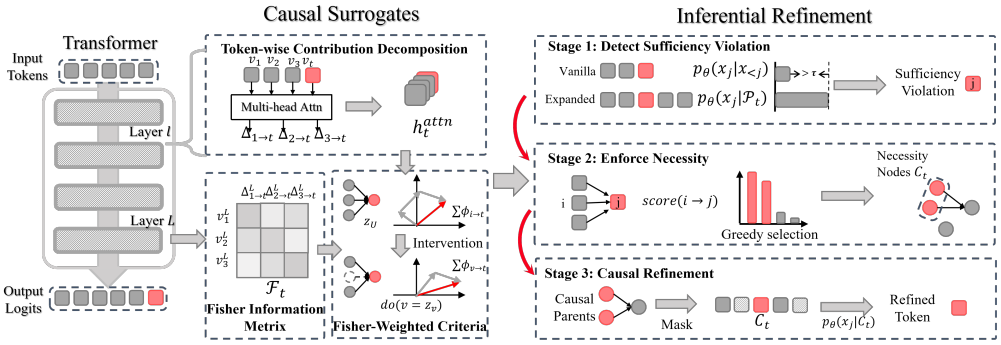


Figure 3: Overview of the proposed dynamic causal structure discovery framework.

4.1 From Interventions to Fisher-Weighted Criteria

The minimal parent set $S_t^\star$ is defined via interventional criteria (Definition 3), which are not directly computable at inference time. We therefore derive a tractable surrogate by exploiting the structural properties of decoder-only Transformers.

Token-wise Contribution Decomposition. At each layer l, the attention output admits an additive decomposition over visible tokens:

$$h_t^{\text{attn},(l)} = \sum_{i \in \mathcal{A}_t} \Delta_{i\to t}^{(l)}, \qquad \Delta_{i\to t}^{(l)} := \sum_{h=1}^{H} \alpha_{t,i}^{(l,h)}\, v_i^{(l,h)}\, W_O^{(l)}[h,:]. \tag{10}$$

Aggregating these paths across all layers yields the total contribution $\phi_{i\to t}$ to the final representation. We assume that these contributions propagate approximately linearly to the output logits, so that removing $\phi_{i\to t}$ from the residual stream provides a pathway-level surrogate for, rather than a full realization of, the do-intervention $\mathrm{do}(x_i = \cdot)$ in Definition 3. This local additivity assumption treats nonlinear downstream interactions as higher-order effects and is stated formally in Assumption 1; the corresponding pathway ablation operation is defined in Appendix C.2. DCD uses this surrogate to evaluate sufficiency and necessity for dependency selection, beyond contribution attribution alone.

Consider an intervention that ablates a subset $\mathcal{I}$ of contributions from the residual stream—i.e., perturbing ℓ_t by $-\sum_{i\in\mathcal{I}} \phi_{i\to t}$. Because the KL divergence between $p_\theta(\cdot \mid \ell_t)$ and a small perturbation of itself vanishes to first order, the leading-order effect of this perturbation is captured by its second-order term, which takes the quadratic form

$$D_{\mathrm{KL}} \;\approx\; \tfrac{1}{2} \Big\| \sum_{i\in\mathcal{I}} \phi_{i\to t} \Big\|_{\mathcal{F}_t}^2 . \tag{11}$$

where $\mathcal{F}_t$ denotes the Fisher information matrix of the output distribution with respect to the residual stream ℓ_t, characterizing the local curvature of the KL divergence around the unperturbed representation. Intuitively, $\mathcal{F}_t$ weights perturbation directions by their influence on the predictive distribution.

Substituting this approximation into Definition 3, with $U = \mathbf{x}_{-t} \setminus S$ in the worst case for the sufficiency condition, yields the following computable criteria for any candidate parent set S:

$$\underbrace{\phi_{v\to t}^\top \mathcal{F}_t\, \phi_{v\to t} \geq 2\eta,}_{\textbf{Necessity}} \qquad \underbrace{\Big\| \sum_{i\,:\,x_i\notin S} \phi_{i\to t} \Big\|_{\mathcal{F}_t}^2 \leq 2\epsilon\,.}_{\textbf{Sufficiency}} \tag{12}$$

These conditions reduce causal structure discovery to identifying a subset that balances Fisher-weighted completeness and minimality, with $S_t^\star$ being one such solution. The full second-order derivation of these Fisher-weighted criteria is provided in Appendix D.1.

Efficient Candidate Screening. Direct evaluation of Eq. (12) is computationally prohibitive: each $\phi_{i\to t}$ requires aggregating attention paths across all layers and heads, and $\mathcal{F}_t$ requires an expectation

over the output vocabulary. We therefore employ a gradient-based surrogate

$$\mathrm{score}(i \to t) := \left\| \nabla_{h_i^{(0)}} \log p_\theta(x_t) \right\|, \tag{13}$$

where $h_i^{(0)}$ is the input embedding of token i. The score is obtained from a single backward pass and approximates the Fisher-weighted relevance of token i to the prediction at t, enabling efficient identification of candidate parents prior to verification of Eq. (12).

4.2 Approximate Structure Discovery at Inference Time

Eq. (12) characterizes the optimal dependency structure but is still not directly enforceable during generation: (i) sufficiency depends on context not yet produced, and (ii) necessity requires combinatorial search over subsets. We therefore operationalize the criteria via a three-stage procedure that triggers periodically during decoding. Each stage corresponds to one component of Eq. (12): **Stage 1** flags positions where Sufficiency is likely violated, **Stage 2** constructs a parent set satisfying Necessity, and **Stage 3** executes the corresponding structural intervention (Algorithm shown in Appendix E.3).

Stage 1: Detecting Sufficiency Violations. Sufficiency is violated when previously generated tokens were committed under a parent set that missed informative future context. At every K generation steps, we re-evaluate each prior token x_j against the expanded context $\mathcal{P}_t := x_{\le t} \setminus \{x_j\}$, and flag position j if

$$\frac{p_\theta(x_j \mid x_{<j})}{p_\theta(x_j \mid \mathcal{P}_t) + \varepsilon_0} > \tau \quad \text{and} \quad \arg\max_v p_\theta(v \mid \mathcal{P}_t) \neq x_j. \tag{14}$$

where ε_0 is a small numerical stabilizer. Under the quadratic approximation in Eq. (12), this likelihood ratio is a monotone function of the sufficiency residual, serving as a cheap proxy for the Fisher-weighted bound. The $\arg\max$ check skips high-entropy positions where refinement would not alter the realized token.

Stage 2: Enforcing Necessity via Greedy Selection. For each flagged position j, we construct a minimal parent set $\hat{S}_j$ through greedy selection. A candidate pool is first formed using the ranking surrogate over the expanded context:

$$\mathcal{C}_j = \{i \in \mathcal{P}_t : \mathrm{score}(i \to j) > \theta_{\mathrm{cau}}\}. \tag{15}$$

Initializing $S \leftarrow \emptyset$, we then greedily admit tokens from $\mathcal{C}_j$ (in descending order of score) only if they strictly reduce the conditional entropy:

$$H_\theta(x_j \mid S \cup \{i\}) < H_\theta(x_j \mid S) - \delta. \tag{16}$$

Upon termination, we set $\hat{S}_j \leftarrow S$. This gating mechanism acts as a direct empirical enforcement of the minimality constraint, being consistent with the necessity condition in Eq. (12).

Stage 3: Causal Refinement via Masked Resampling. Stages 1 and 2 identify *which* tokens to refine and *against which* parent set. Stage 3 performes the refinement by overriding the default causal attention mask at position j, replacing the prefix mask $\{i : i < j\}$ with the discovered parent set $\hat{S}_j$:

$$M_j(i) = \begin{cases} 1, & i \in \hat{S}_j, \\ 0, & \text{otherwise}. \end{cases} \tag{17}$$

This rewires the residual stream so that information at position j flows *only* through the variables in $\hat{S}_j$. We then resample under this modified mask and replace x_j with the new draw:

$$x'_j \sim p_\theta(\cdot \mid x_{\hat{S}_j}), \qquad x_j \leftarrow x'_j. \tag{18}$$

Theoretical Justification. The following result guarantees that this refinement process corresponds to a meaningful structural correction aligned with our causal framework.

Proposition 1 (Refinement Guarantee)**.** *If the refined set $\hat{S}_t$ satisfies the criteria in Eq.* (12) *with parameters (η, ϵ),*

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{S_t}) \,\big\|\, p_\theta(x_t \mid x_{<t})\big) \;\ge\; \eta - \epsilon - o(1). \tag{19}$$

Consequently, each accepted update corresponds to a principled structural correction rather than a stochastic perturbation. Proposition 1 guarantees that the refinement induces a non-trivial prediction shift, while Appendix D.3 further shows that this shift remains controlled by the contribution of newly admitted parents and the residual influence of removed prefix tokens.

Table 1: Main results on ImageNet, where we report generation quality, including FID, IS, Precision, and Recall, as well as inference efficiency, including latency, and throughput.

Type	Model	#Para	FID↓	IS↑	P↑	R↑	L(s)↓	T↑
VAR	VAR-d16 [5]	310M	3.30	274.4	0.84	0.51	0.12	70.58
	VAR-d20 [5]	600M	2.57	302.6	0.83	0.56	0.15	52.53
	VAR-d24 [5]	1.0B	2.09	312.9	0.82	0.59	0.17	39.30
	VAR-d30 [5]	2.0B	1.92	323.1	0.82	0.59	0.26	25.89
Raster	LlamaGen-L [3]	343M	3.07	256.1	0.83	0.52	12.22	2.08
	LlamaGen-XL [3]	775M	2.62	244.1	0.80	0.57	18.51	1.14
	LlamaGen-XXL [3]	1.4B	2.34	253.9	0.80	0.59	24.40	0.72
	LlamaGen-3B [3]	3.1B	2.18	263.3	0.81	0.58	12.37	0.58
Random	RandAR-L [2]	343M	2.55	288.8	0.81	0.58	1.97	28.59
	RandAR-XL [2]	775M	2.25	317.8	0.80	0.60	2.78	17.06
	RandAR-XXL [2]	1.4B	2.15	322.0	0.79	0.62	3.58	11.49
	RAR-B [23]	261M	1.95	290.5	0.82	0.58	4.18	13.76
	RAR-L [23]	461M	1.70	299.5	0.81	0.60	4.04	12.63
	RAR-XL [23]	955M	1.50	306.9	0.80	0.62	5.47	8.76
	ARPG-L [24]	320M	2.44	291.7	0.82	0.55	0.58	104.92
	ARPG-XL [24]	719M	2.10	331.0	0.79	0.61	1.71	36.53
Spatial-aware	PAR-L-4× [25]	343M	3.76	218.9	0.84	0.50	3.16	6.83
	PAR-XL-4× [25]	775M	2.61	259.2	0.82	0.56	4.79	3.69
	PAR-XXL-4× [25]	1.4B	2.35	263.2	0.82	0.57	6.26	2.33
	PAR-3B-4× [25]	3.1B	2.29	255.5	0.82	0.58	3.29	2.32
	NAR-L [26]	372M	3.06	263.9	0.81	0.53	1.01	41.03
	NAR-XL [26]	816M	2.70	277.5	0.81	0.58	1.42	23.36
	NAR-XXL [26]	1.5B	2.58	293.5	0.82	0.57	1.88	15.20
	LPD-L [27]	337M	2.40	284.5	0.81	0.57	0.28	139.11
	LPD-XL [27]	752M	2.10	326.7	0.80	0.59	0.41	75.20
	LPD-XXL [27]	1.4B	2.00	337.6	0.80	0.60	0.55	45.07
Ours	RandAR-L + Ours	343M	2.33	272.8	0.82	0.60	7.63	4.55
	RandAR-XL + Ours	775M	2.11	296	0.82	0.61	11.85	2.37
	ARPG-XL + Ours	719M	2.01	319.6	0.82	0.62	8.36	4.21

5 Experiments

Tasks. We build on RandAR [2] for class-conditional image generation on ImageNet [28].

Comparisons. We compare with representative baselines under three categories: (1) raster-scan AR models, including LlamaGen [3]; (2) randomized-order methods, including RandAR [2], RAR [23], and ARPG [24]; and (3) spatial-aware approaches, including PAR [25], NAR [26], and LPD [27]. All methods are evaluated on ImageNet under the same tokenizer and class-conditional settings.

Metrics. We use FID as the primary metric, together with IS[29], Precision, and Recall[30] for generation quality evaluation. For causal discovery, we measure alignment between gradient-based attribution and intervention-based ground truth using WSHD, WNSHD, WF1, and AUPRC, where WNSHD and AUPRC quantify distributional discrepancy and ranking fidelity, respectively.Experiment-specific sample counts and evaluation protocols are provided in Appendix E.2.

5.1 Main Comparison

We compare our method with different AR methods in Table 1, where our method achieves competitive performance with comparable parameter sizes. VAR combines next-scale prediction and model scaling for strong generation quality, while raster-order methods offer simple decoding but limited throughput. Random-order methods improve flexibility but retain order-defined conditioning, whereas spatial-aware methods remain constrained by spatial priors. Our DCD dynamically selects relevant context without retraining, reducing FID from 2.55 to 2.33 on RandAR-L and from 2.25 to 2.11 on RandAR-XL. It further reduces FID on ARPG-XL from 2.10 to 2.01, demonstrating transferability across random-order AR architectures. DCD's causal context refinement improves distributional

Table 2: Ablation study on refinement strategy, where we choose different tokens as parent nodes to conduct refinement.

Category	Strategy	FID↓	IS↑	Precision↑	Recall↑
Baseline	w/o Refinement	5.46	199.7	0.89	0.44
Static	Static Causal	4.82	188.5	0.91	0.44
Dynamic	Spatial Neighborhood	5.11	197.2	0.90	0.45
	Full Context	4.95	196.5	0.89	0.43
	DCD (Ours)	4.79	192.4	0.91	0.47

Table 3: Comparison with simpler refinement baselines under a matched token-refinement budget.

Method	Refined/image ↓	FID ↓	IS ↑	P ↑	R ↑
No Refinement	0	5.46	199.7	0.89	0.44
Random Selection	69	5.25	177.9	0.86	0.43
Low Confidence	69	5.11	191.2	0.87	0.43
High Entropy	69	4.87	184.1	0.87	0.45
DCD	69	4.79	192.4	0.91	0.47

fidelity and coverage, as indicated by lower FID and higher Precision and Recall, while yielding a lower IS. We analyze DCD's inference cost and quality–efficiency trade-off in Section 5.3.

5.2 Ablation studies

Ablation study on refinement strategy. Table 2 reveals the effectiveness of our causal refinement, where we replace our dynamic causal discovery strategy with various alternative solutions. As shown in row 2, refinement using a static causal graph discovered after full image generation achieves a comparable FID to DCD, but a lower IS, since static causal graph fail to meet the dynamic mechanism of AR, leading to low diversity. Meanwhile, rows 3-4 show that using spatial neighborhood or full context also obtain suboptimal performance, since they rely on a structural prior misaligned with the intrinsic structure of images. In contrast, our method achieves the lowest FID of 4.79, validating our dynamic causal discovery strategy is essential for structural fidelity.

Comparison with simpler refinement criteria. Table 3 compares DCD with Random Selection, Low Confidence, and High Entropy, each refining an average of 69 tokens per image. DCD outperforms all three baselines in FID, IS, Precision, and Recall, demonstrating the effectiveness of our causal refinement strategy under the same token-refinement budget.

Table 4: Fidelity analysis of causal graph traced with Fisher-weighted criteria. We compare our approximate causal discovery against the do-intervention graph under different causal set sizes $|S_j|$.

Method	Causal Set	WNSHD↓	WF1↑	W-Prec↑	AUPRC↑	Runtime(s)↓
Do-Intervention	Full	0.0000	1.0000	1.0000	1.0000	5.2
Ours	Full	0.2827	0.6603	1.0000	0.5284	0.12
	$\|S_j\|$=50	0.2827	0.5355	0.922	0.2745	0.12
	$\|S_j\|$=20	0.2827	0.4726	0.8737	0.2037	0.12

Fidelity of Fisher-weighted criteria. To verify the effectiveness of our Fisher-weighted criteria, we quantitatively compare the inference-time causal graphs against do-intervention[31]. In Table 4, our method achieves a WF1=0.6603 using full causal parents as context, faithfully captures most causal correlations. Notably, our approach reduces the runtime from 5.2s per sample to a mere 0.12s, which underscores the computational efficiency of our Fisher-weighted criteria for real-time decoding.

Sensitivity to hyperparameters. To validate the influence of hyperparameters, we conduct ablation studies on sufficiency violation detection threshold τ and the refinement step interval K. As shown in Table 5 (a), when τ varies, our method maintains consistent performance, successfully captures

Table 5: Ablation studies on key refinement parameters.

(a) Sensitivity to detection threshold τ.

	FID↓	IS↑	Prec.↑	Rec.↑
$\tau = 3.0$	71.79	51.0	0.80	0.80
$\tau = 5.0$	71.87	52.6	0.79	0.80
$\tau = 8.0$	71.72	50.1	0.80	0.79

(b) Sensitivity to refinement step interval K.

	FID↓	IS↑	Prec.↑	Rec.↑
K = 32	72.46	48.2	0.81	0.81
K = 64	73.55	53.4	0.83	0.82
K = 128	72.26	47.4	0.80	0.79

Table 6: Stage-wise computational and memory costs of DCD.

Stage	Latency (s/image) ↓	Forward	Backward	Peak VRAM (MB) ↓
Original AR decoding	2.780	256.0	0.0	4,880.9
Violation detection	6.209	16.0	0.0	22,474.5
Gradient scoring	1.838	68.7	68.7	5,789.7
Greedy verification	0.517	68.7	0.0	4,811.9
Masked resampling	0.506	68.7	0.0	4,880.5
Total	11.850	478.1	68.7	22,474.5

the crucial context for AR generation. Table 5 (b) further illustrates that higher refinement step does not influence the generation quality strikingly. We choose $\tau = 5.0$ and $K = 64$ in our experiments.

5.3 Efficiency Analysis

Stage-wise computational breakdown. Table 6 reports the inference cost of individual stages in DCD. Violation detection takes 6.209 seconds out of the total 11.850 seconds per image and incurs the highest peak memory usage of 22,474.5 MB. Gradient scoring takes 1.838 seconds, while greedy verification and masked resampling require 0.517 and 0.506 seconds, respectively. Overall, DCD performs an average of 478.1 forward and 68.7 backward evaluations per image. These measurements identify expanded-context evaluation as the primary efficiency bottleneck.

Quality–efficiency trade-off. Table 7 compares three profiles by jointly varying the detection interval K, the maximum refinements per detection step B, and the maximum parent tokens per target M. Fast, Balanced, and Quality use $(K, B, M) = (128, 16, 15)$, $(64, 32, 20)$, and $(32, \infty, \infty)$, respectively, with all other hyperparameters fixed. Balanced reduces FID from 5.46 to 4.92, retaining approximately 81% of Quality's FID improvement. Compared with Quality, it reduces latency from 11.85 to 7.72 seconds and increases throughput from 2.37 to 5.55. Fast further reduces latency to 4.83 seconds, with FID 5.33. These profiles offer adjustable trade-offs between generation quality and inference efficiency.

Table 7: Quality–efficiency trade-off under different refinement budgets.

Profile	Refined/image	FID ↓	IS ↑	P ↑	R ↑	L (s) ↓	T ↑
RandAR	0	5.46	199.7	0.89	0.440	2.78	17.06
DCD Fast	27	5.33	180.5	0.87	0.447	4.83	10.20
DCD Balanced	50	4.92	190.9	0.89	0.458	7.72	5.55
DCD Quality	69	4.79	192.4	0.91	0.470	11.85	2.37

5.4 Visualization

Qualitative Results. We provide a qualitative analysis of the generation quality achieved by our method in Figure 4. Compared with the baseline (RandAR), our method demonstrates consistent improvements across multiple categories. This indicates that refining generation with minimal causal dependencies enables each image token to effectively incorporate global contextual information while suppressing redundant semantic cues.

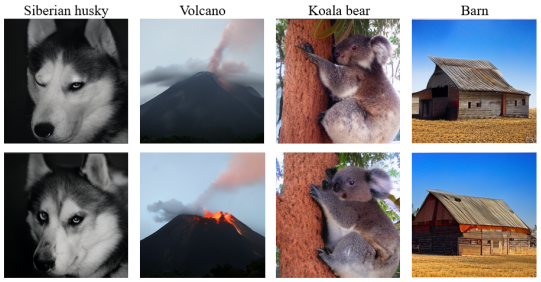


Figure 4: Qualitative comparison before and after causal refinement across diverse ImageNet categories. **Top row**: baseline results generated without refinement. **Bottom row**: refined results using our method.

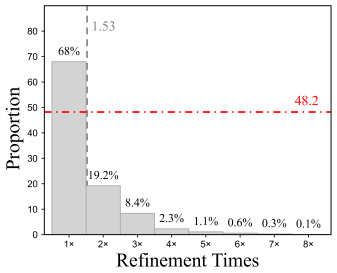


Figure 5: Refinement frequency analysis. Horizontal and vertical lines denote the average proportion and the expected frequency, respectively.

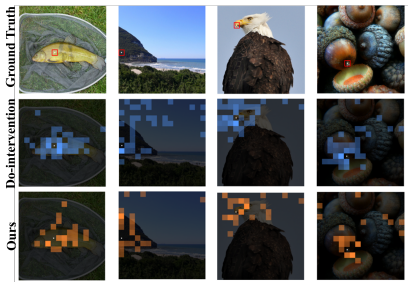


Figure 6: Visualization of causal parent maps. Parent sets are highlighted.

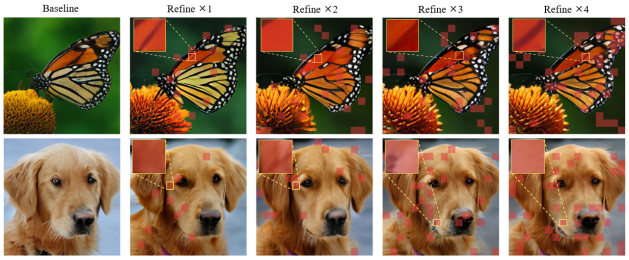


Figure 7: Progressive visualization of the causal refinement process.

Statistical analysis of refinement. In Figure 5, we report the refinement frequency distribution for all generated tokens when $K = 32$. Most tokens require only one refinement step to satisfy the sufficiency condition, with an expected frequency of 1.53. This indicates that a small number of refinements is generally sufficient for tokens to obtain adequate contextual information. In addition, near half of the generated tokens(48.2%) are detected to violate sufficiency, suggesting that sufficiency deficiency is a common phenomenon in AR image generation. This further implies that the strong structural prior in AR generation may cause generation under insufficient contextual conditions, leading to degraded generation quality.

Causal Parent Map Visualization. To demonstrate the effectiveness of our Fisher-weighted criteria, we visualize causal parent nodes of several tokens in Figure 6. Tokens associated with large objects tend to rely on their surrounding tokens as local contextual support, whereas tokens corresponding to small objects are more likely to depend on sparse global semantic cues. The causal relationships discovered by our method are highly consistent with those obtained through the standard do-intervention strategy, indicating that the proposed Fisher-weighted criteria can efficiently approximate the underlying causal dependencies.

Progressive visualization for causal refinement. In Figure 7, we visualize the image evolution during multiple refinement steps under the $K = 64$ setting. As shown in the figure, with the progression of generation steps, more sufficiency violations are detected according to the increasingly complete context, thereby enabling more thorough refinement. Consequently, the refined tokens exhibit clearer and more realistic texture details, leading to progressively improved visual quality.

6 Conclusion

We introduced Dynamic Causal Structure Discovery, a framework that replaces fixed AR conditioning with dynamic causal dependency selection. By connecting causal interventions with Transformer computation, our method enables efficient inference-time refinement and consistently improves generation quality and global coherence without retraining.

Acknowledgements and Disclosures

This work was supported in part by the Natural Science Foundation of Jiangsu Province of China under Grants BK20242050 and BK20251504, in part by the Fundamental Research Funds for the Central Universities under Grants B250201042 and B250201046, in part by the Open Research Project of State Key Laboratory for Novel Software Technology under Grant KFKT2025B04, and in part by the National Natural Science Foundation of China under Grant 6250074646. This work was also supported in part by the High Performance Computing Platform, Hohai University.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *Proceedings of International Conference on Computer Vision*, pages 18431–18441, 2025.

[2] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 45–55, 2025.

[3] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *CoRR*, abs/2406.06525, 2024.

[4] Zhuoyang Zhang, Luke J. Huang, Chengyue Wu, Shang Yang, Kelly Peng, Yao Lu, and Song Han. Locality-aware parallel decoding for efficient autoregressive image generation. In *Proceedings of International Conference on Learning Representations*, 2026.

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

[6] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *Proceedings of International Conference on Computer Vision*, pages 15781–15791, 2025.

[7] Cheng Cheng, Lin Song, Di An, Yicheng Xiao, Xuchong Zhang, Hongbin Sun, and Ying Shan. From prediction to perfection: Introducing refinement to autoregressive image generation. In *Proceedings of International Conference on Learning Representations*, 2026.

[8] Yefei He, Yuanyu He, Shaoxuan He, Feng Chen, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Neighboring autoregressive modeling for efficient visual generation. In *Proceedings of International Conference on Computer Vision*, pages 19000–19010, 2025.

[9] Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. In *Proceedings of International Conference on Learning Representations*, 2025.

[10] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *Proceedings of International Conference on Learning Representations*, 2025.

[11] Yonghao Yu, Lang Huang, Zerun Wang, Runyi Li, and Toshihiko Yamasaki. Mirai: Autoregressive visual generation needs foresight. *CoRR*, abs/2601.14671, 2026.

[12] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[13] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

[14] Jiatao Gu, Yuyang Wang, Yizhe Zhang, Qihang Zhang, Dinghuai Zhang, Navdeep Jaitly, Joshua M. Susskind, and Shuangfei Zhai. Denoising autoregressive transformers for scalable text-to-image generation. In *Proceedings of International Conference on Learning Representations*, 2025.

[15] Zhanzhou Feng, Qingpei Guo, Jingdong Chen, Feng Gao, Ming Yang, and Shiliang Zhang. SCAN: self-calibrated autoregression for high-quality visual generation. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3976–3984, 2026.

[16] Joseph Y. Halpern and Christopher Hitchcock. Actual causation and the art of modeling. *CoRR*, abs/1106.2652, 2011.

[17] Joseph Y. Halpern. A modification of the halpern-pearl definition of causality. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 3022–3033, 2015.

[18] Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

[19] Zhuguanyu Wu, Shihe Wang, Jiayi Zhang, Jiaxin Chen, and Yunhong Wang. FIMA-Q: post-training quantization for vision transformers by fisher information matrix approximation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14891–14900, 2025.

[20] Xiaodong Wu, Wenyi Yu, Chao Zhang, and Philip C. Woodland. An improved empirical fisher approximation for natural gradient descent. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

[21] Haopeng Li, Jinyue Yang, Guoqi Li, and Huan Wang. Autoregressive image generation with randomized parallel decoding. In *Proceedings of International Conference on Learning Representations*, 2026.

[22] Judea Pearl. *Causality*. 2009.

[23] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *CoRR*, abs/2411.00776, 2024.

[24] Haopeng Li, Jinyue Yang, Guoqi Li, and Huan Wang. Autoregressive image generation with randomized parallel decoding. *CoRR*, abs/2503.10568, 2025.

[25] Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12955–12965, 2025.

[26] Yefei He, Yuanyu He, Shaoxuan He, Feng Chen, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Neighboring autoregressive modeling for efficient visual generation. *CoRR*, abs/2503.10696, 2025.

[27] Zhuoyang Zhang, Luke J. Huang, Chengyue Wu, Shang Yang, Kelly Peng, Yao Lu, and Song Han. Locality-aware parallel decoding for efficient autoregressive image generation. *CoRR*, abs/2507.01957, 2025.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[29] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of Advances in Neural Information Processing Systems*, pages 2226–2234, 2016.

[30] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Proceedings of Advances in Neural Information Processing Systems*, pages 3929–3938, 2019.

[31] Shanshan Huang, Haoxuan Li, Chunyuan Zheng, Lei Wang, Guorui Liao, Zhili Gong, Huayi Yang, and Li Liu. Visual representation learning through causal intervention for controllable image editing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23484–23493, 2025.

[32] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of International Conference on Machine Learning*, pages 12606–12633, 2024.

[33] Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *Proceedings of International Conference on Learning Representations*, 2024.

[34] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *Proceedings of International Conference on Learning Representations*, 2025.

[35] Dongyang Liu, Yi Xin, Shitian Zhao, Le Zhuo, Weifeng Lin, Xinyue Li, Qi Qin, Guangtao Zhai, Xiaohong Liu, Hongsheng Li, Yu Qiao, and Peng Gao. Lumina-mgpt: Flexible photorealistic autoregressive text-to-image generation. *Int. J. Comput. Vis.*, 134(4):141, 2026.

[36] Yongtao Liu, Anna N. Morozovska, Eugene A. Eliseev, Kyle P. Kelley, Rama K. Vasudevan, Maxim A. Ziatdinov, and Sergei V. Kalinin. Autonomous scanning probe microscopy with hypothesis learning: Exploring the physics of domain switching in ferroelectric materials. *Patterns*, 4(3):100704, 2023.

[37] Weiwei Lin and Chenhang He. Continuous autoregressive modeling with stochastic monotonic alignment for speech synthesis. In *Proceedings of International Conference on Learning Representations*, 2025.

[38] Jieru Lin, Danqing Huang, Tiejun Zhao, Dechen Zhan, and Chin-Yew Lin. Spot the error: Non-autoregressive graphic layout generation with wireframe locator. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 3413–3421, 2024.

[39] Jakiw Pidstrigach, Elizabeth Louise Baker, Carles Domingo-Enrich, George Deligiannidis, and Nikolas Nüsken. Conditioning diffusions using malliavin calculus. In *Proceedings of International Conference on Machine Learning*, 2025.

[40] Martina Cinquini, Isacco Beretta, Salvatore Ruggieri, and Isabel Valera. A practical approach to causal inference over time. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 14832–14839, 2025.

[41] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *Proceedings of International Conference on Learning Representations*, 2024.

[42] Ning Miao, Yee Whye Teh, and Tom Rainforth. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. In *Proceedings of International Conference on Learning Representations*, 2024.

[43] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of International Conference on Machine Learning*, pages 19274–19286, 2023.

[44] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[45] Aman Madaan, Niket Tandon, and Prakhar Gupta. Self-refine: Iterative refinement with self-feedback. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[46] Divyam Madaan, Taro Makino, Sumit Chopra, and Kyunghyun Cho. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. In *Proceedings of Advances in Neural Information Processing Systems*, 2024.

[47] Xuexin Chen, Ruichu Cai, Zhengting Huang, Yuxuan Zhu, Julien Horwood, Zhifeng Hao, Zijian Li, and José Miguel Hernández-Lobato. Feature attribution with necessity and sufficiency via dual-stage perturbation test for causal explanation. In *Proceedings of International Conference on Machine Learning*, pages 6486–6502, 2024.

[48] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *Proceedings of International Conference on Learning Representations*, 2024.

[49] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of International Conference on Machine Learning*, pages 62429–62442, 2024.

[50] Dan Kondratyuk, Lijun Yu, and Xiuye Gu. Videopoet: A large language model for zero-shot video generation. In *Proceedings of International Conference on Machine Learning*, pages 25105–25124, 2024.

[51] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *CoRR*, abs/2310.01889, 2023.

[52] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *Proceedings of International Conference on Learning Representations*, 2024.

[53] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of International Conference on Learning Representations*, 2024.

[54] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending LLM context window beyond 2 million tokens. In *Proceedings of International Conference on Machine Learning*, pages 11091–11104, 2024.

[55] Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5819–5828, 2024.

[56] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2O: heavy-hitter oracle for efficient generative inference of large language models. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

[57] Cheng Cheng, Song Lin, An Di, Xiao Yicheng, Zhang Xuchong, Sun Hongbin, and Shan Ying. From prediction to perfection: Introducing refinement to autoregressive image generation. In *Proceedings of International Conference on Learning Representations*, pages 1–15, 2026.

[58] Sangkyu Lee, Changho Lee, Janghoon Han, Hosung Song, Tackgeun You, Hwasup Lim, Stanley Jungkyu Choi, Honglak Lee, and Youngjae Yu. Spanning tree autoregressive visual generation. *CoRR*, abs/2511.17089, 2025.

A Related Work

A.1 Autoregressive Visual Generation

Recent AR visual generation methods have begun exploring alternatives to fixed raster-scan sequence modeling [32, 2, 3] to diverse, dynamically scalable ordering paradigms [5, 33, 34]. Standard scaling approaches ensure production control by treating images as flattened sequences and optimizing next-token prediction across immense parameters [35, 13]. Alternative ordering methods leverage randomized permutations, space-filling curves, or bidirectional generation paths to capture visual context from multiple spatial orientations [36, 2, 23, 24, 25, 26, 27]. Meanwhile, parallel and continuous decoding strategies have emerged to accelerate the strict step-by-step generative bottleneck [37]. Despite these advances, most generative models prioritize exposing varying prefix contexts or increasing decoding speed over fundamentally altering the prefix-based conditioning logic itself. In the context of visual dependency modeling, forcing a token's conditioning set to rely on its ordinal position rather than semantic content inevitably introduces accumulated trajectory errors [38]. Breaking this ordinal precedence to reconstruct a content-driven, order-agnostic dependency structure is critical for global coherence yet remains a significant technical hurdle.

A.2 Causal Dependency Modeling

Recent studies on autoregressive dependency optimization have evolved from heuristic decoding strategies and token-level truncation methods toward more principled causal modeling and inference-time refinement techniques [39, 5, 40]. Beyond heuristic search, causal-based approaches improve generation robustness by modeling structural dependencies and mitigating spurious correlations through structural causal frameworks [41]. Meanwhile, modern inference-time refinement methods leverage speculative decoding, backtracking search, and self-correction mechanisms to improve alignment and output fidelity without retraining the base model [42, 43, 44, 45]. However, most existing refinement strategies primarily adjust output probability distributions or re-rank candidate sequences, rather than directly revising the underlying autoregressive dependency structure itself. In dynamic sequential generation, reliable prediction requires dependencies that are both causally sufficient and necessary for the current generation state [46, 47]. Consequently, operating directly on the token-level computational dependency graph to enable adaptive refinement beyond fixed sequential ordering remains a significant technical challenge, which our lightweight framework explicitly addresses.

A.3 Context Utilization

Context utilization in visual generation has transitioned from basic local perception architectures [33, 48] to sophisticated, multi-granular dependency frameworks [49, 50]. Token-level approaches ensure feature richness by enhancing visual tokenization schemes and exploring optimized discrete or continuous representations [48]. Pairwise context methods leverage dynamic masking, linear attention mechanisms, and KV-cache optimization to manage direct token-to-token interactions across extended sequences efficiently [51, 52, 53, 54]. Structure-level innovations, are designed to capture long-range semantic dependencies in high-resolution images and videos [49, 50, 55]. Despite these architectural upgrades, most context utilization frameworks prioritize expanding context capacity or accelerating attention computations over distinguishing true dependency necessity. In autoregressive visual prediction, operating under the implicit assumption that all prior tokens are uniformly beneficial inevitably introduces redundant variables, leading to weakened global coherence [56, 13]. While existing methods focus on expanding the context window, the real challenge lies in dynamically discarding irrelevant tokens to extract an ideal causal parent set. This structural problem remains largely unaddressed.

B Motivation: Dependency Structures in AR Visual Generation

Building on the strong empirical performance of autoregressive (AR) models in sequence modeling, recent works[36, 2, 23, 24, 25] have extended AR paradigms to visual generation, leveraging their scalability, flexibility, and compatibility with multimodal inputs. At their core, AR models adopt a simple factorization strategy, modeling complex data distributions through sequential next-token

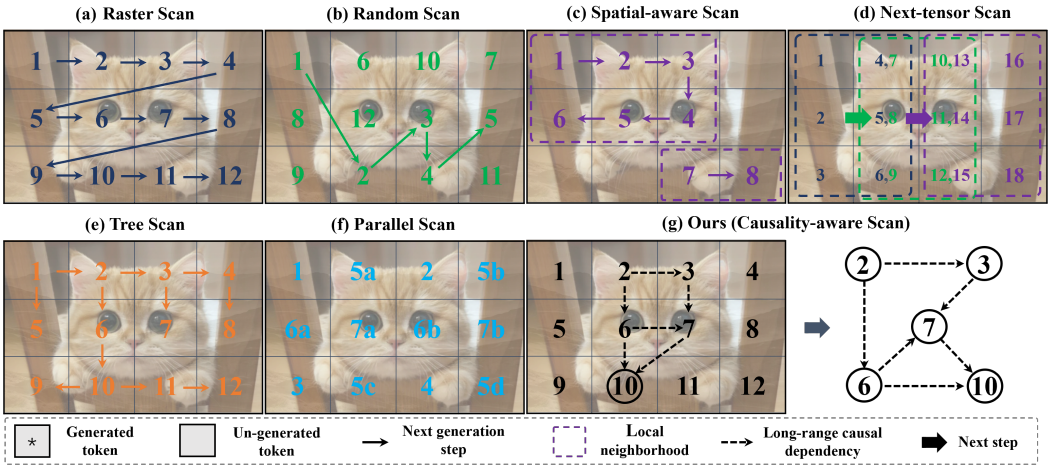


Figure 8: Comparison of representative autoregressive models for visual generation: (a) Raster Scan [3], (b) Random Scan [2], (c) Spatial-aware Scan [27], (d) Next-tensor Scan [57], (e) Tree Scan [58], (f) Parallel Scan [25], and (g) our causality-aware dependency structure after dynamically discovering.

prediction. This formulation enables structured generation and has proven effective across a range of vision tasks[32, 13, 45].

Fig. 8(a)-(f) illustrate common AR paradigms in image generation. For example, standard AR approaches typically represent images as sequences of discrete tokens (e.g., patches) and model dependencies according to a predefined ordering, such as Raster Scan[3]. While this serialization makes the problem tractable, it imposes a rigid one-dimensional structure on inherently two-dimensional data. As a result, the dependency relationships captured by the model are dictated by the chosen ordering rather than the underlying spatial structure of the image, potentially limiting modeling efficiency.

To alleviate these limitations, a growing body of work[2, 57, 58, 27, 25] has explored various extensions of the autoregressive paradigm. One line of research explores randomized or continuous generation processes, such as Random Scan [2], which enables iterative correction during generation. Another direction redesigns the prediction space by modeling tokens as continuous-valued tensors rather than discrete symbols [57]. In parallel, several works investigate alternative traversal strategies, including Tree Scan [58], Spatial-aware Scan [27], and Parallel Scan [25], to better align autoregressive generation with the spatial structure of images.

While these approaches have led to substantial improvements in generation quality, they typically introduce additional architectural components, modify the underlying training objectives, or incur increased computational and memory overhead. More importantly, despite their differences in formulation, most of these methods retain the core autoregressive factorization and its reliance on a predefined ordering. As a result, the dependency structure governing generation remains largely unchanged, and the fundamental mismatch between sequential ordering and spatial dependencies is not explicitly addressed.

This observation suggests that the limitations of AR visual generation may not solely arise from the choice of parameterization or generation dynamics, but rather from the dependency structure induced by the factorization itself. Consequently, improving AR models may require rethinking not only how tokens are generated, but also how their conditioning context is defined.

In this work, we take a different perspective and ask: can the limitations of AR visual generation be addressed by rethinking its underlying dependency structure? Specifically, instead of committing to a fixed generation order, can the model adapt its conditioning context based on the relevance of different tokens? This viewpoint suggests that the core challenge is not only how to generate tokens sequentially, but also how to identify appropriate dependencies for each prediction step, motivating a more flexible and data-aligned formulation of autoregressive visual generation. As illustrated in Figure 8(g), our method dynamically constructs content-dependent causal dependencies during

decoding, allowing each prediction to selectively attend to both local and long-range informative context rather than relying solely on predefined sequential ordering.

C Supplementary Definitions

This appendix collects formal definitions and assumptions used in the main text. Cross-references to the main paper are given in parentheses.

C.1 Transformer Quantities

We work with a decoder-only Transformer with L layers and H attention heads per layer. The notation in this subsection underlies the development of §4.1 and §4.2.

Definition 4 (Visible Set)**.** *For a position $t \in \{1, \dots, T\}$, the* visible set $\mathcal{A}_t \subseteq \{1, \dots, T\} \setminus \{t\}$ *is the set of positions whose hidden states are admitted by the attention mask at position t. Under the standard causal mask, $\mathcal{A}_t = \{i : i < t\}$. Under the modified mask used in §4.2, $\mathcal{A}_t$ is determined by the discovered parent set $\hat{S}_t$ as $\mathcal{A}_t = \hat{S}_t$.*

Definition 5 (Residual Stream)**.** *The* residual stream *at position t and layer l is the hidden state $h_t^{(l)} \in \mathbb{R}^d$, obtained by accumulating attention and MLP outputs from layer 0 up to layer l. We write $\ell_t := h_t^{(L)}$ for the final-layer residual; the output distribution $p_\theta(x_t \mid \cdot)$ is a function of ℓ_t alone.*

Definition 6 (Token-wise Contribution)**.** *At layer l, the attention output decomposes additively over the visible set as*

$$h_t^{\mathrm{attn},(l)} = \sum_{i \in \mathcal{A}_t} \Delta_{i \to t}^{(l)}, \qquad \Delta_{i \to t}^{(l)} := \sum_{h=1}^{H} \alpha_{t,i}^{(l,h)} \, v_i^{(l,h)} \, W_O^{(l)}[h, :], \tag{20}$$

where $\alpha_{t,i}^{(l,h)}$ is the attention weight, $v_i^{(l,h)}$ is the value vector, and $W_O^{(l)}[h, :]$ is the per-head output projection. The token-wise contribution *of i to position t is the aggregate $\phi_{i \to t} := \sum_{l=1}^{L} \widetilde{\Delta}_{i \to t}^{(l)}$, where $\widetilde{\Delta}_{i \to t}^{(l)}$ is the projection of $\Delta_{i \to t}^{(l)}$ onto ℓ_t via the residual connections of subsequent layers. By construction, $\ell_t = \sum_{i \in \mathcal{A}_t} \phi_{i \to t} + \mathrm{const}_t$, where const_t collects MLP and bias terms not attributable to any single visible token.*

Definition 7 (Residual-Stream Fisher Information)**.** *The* Fisher information matrix at position t *is the Hessian of the KL divergence with respect to the residual stream ℓ_t:*

$$\mathcal{F}_t := \mathbb{E}_{x_t \sim p_\theta(\cdot \mid \ell_t)}\big[\nabla_{\ell_t} \log p_\theta(x_t \mid \ell_t) \, \nabla_{\ell_t} \log p_\theta(x_t \mid \ell_t)^\top\big]. \tag{21}$$

Equivalently, $\mathcal{F}_t = -\mathbb{E}_{x_t}[\nabla_{\ell_t}^2 \log p_\theta(x_t \mid \ell_t)]$. We write $\|v\|_{\mathcal{F}_t}^2 := v^\top \mathcal{F}_t \, v$ for the induced quadratic form.

C.2 Causal Operations on the Transformer

Definition 8 (Pathway Ablation)**.** *Given a subset $\mathcal{I} \subseteq \mathcal{A}_t$, the* pathway ablation *of $\mathcal{I}$ at position t is the modification*

$$\ell_t \;\mapsto\; \ell_t - \sum_{i \in \mathcal{I}} \phi_{i \to t}, \tag{22}$$

which removes the contributions in $\mathcal{I}$ from the residual stream while keeping all other quantities fixed. Under Assumption 1, this provides a pathway-level surrogate for the structural intervention $\mathrm{do}(x_i = \cdot)$ in Definition 3 for $i \in \mathcal{I}$.

Definition 9 (Mask Restriction)**.** *Given a target subset $S \subseteq \mathcal{A}_t$, the* mask restriction *to S replaces the visible set with S at position t, equivalently setting the attention mask*

$$M_t(i) = \begin{cases} 1, & i \in S, \\ 0, & \textit{otherwise}. \end{cases} \tag{23}$$

Under mask restriction, the residual stream becomes $\ell_t = \sum_{i \in S} \phi_{i \to t} + \mathrm{const}_t$, suppressing all pathways outside S. Mask restriction is the operational form of pathway ablation in §4.2.

Definition 10 (Expanded Context)**.** *At a refinement step t during decoding, the* expanded context *for a prior position $j < t$ is*

$$\mathcal{P}_t(j) := x_{\leq t} \setminus \{x_j\}, \tag{24}$$

i.e., all currently generated tokens except x_j. Evaluating $p_\theta(x_j \mid \mathcal{P}_t(j))$ requires lifting the causal mask at position j so that j can attend to tokens $\{i : i > j\} \cap x_{\leq t}$; this is implemented by a single additional forward pass with the modified mask. We write $\mathcal{P}_t$ for $\mathcal{P}_t(j)$ when j is clear from context.

C.3 Information-Theoretic Quantities

Definition 11 (Conditional Entropy and Conditional KL)**.** *For random variables X, Y, Z over discrete alphabets and a joint distribution p, we use the standard quantities:*

$$H_p(X \mid Y) := \mathbb{E}_{y \sim p(Y)}\big[-\sum_x p(x \mid y) \log p(x \mid y)\big], \tag{25}$$

$$D_{\mathrm{KL}}\big(p(X \mid Y) \,\big\|\, q(X \mid Y)\big) := \mathbb{E}_{y \sim p(Y)}\big[\textstyle\sum_x p(x \mid y) \log \frac{p(x|y)}{q(x|y)}\big], \tag{26}$$

$$I_p(X; Y \mid Z) := H_p(X \mid Z) - H_p(X \mid Y, Z). \tag{27}$$

In the main text, H_θ denotes conditional entropy under p_θ, and I_{p^*} denotes conditional mutual information under the data distribution p^*.

C.4 Assumptions

Assumption 1 (Local Additivity of Contributions)**.** *The token-wise contributions $\{\phi_{i \to t}\}_{i \in \mathcal{A}_t}$ propagate approximately linearly to the output logits in a neighborhood of the unperturbed residual stream ℓ_t. Equivalently, ablating a small subset of contributions induces a first-order change in the logits, with nonlinear corrections bounded in the operating regime of inference.*

This assumption underpins the identification of pathway ablation (Definition 8) with the structural do-intervention of Definition 3, and is implicit in the second-order expansion of §D.1.

Assumption 2 (Well-Specification)**.** *The model distribution p_θ matches the data-generating distribution p^* on the conditional sets $\{x_S : S \cup \mathbf{x}_{-t}\}$ that appear in Definition 3 and Theorem 1, in the sense that $D_{\mathrm{KL}}(p^*(x_t \mid x_S) \,\|\, p_\theta(x_t \mid x_S))$ is negligible for the analysis.*

This assumption is invoked in the proof of Theorem 1 to translate information-theoretic claims about p^* (the source of Limitations I and II in §3.1) into operational claims about p_θ (the target of the suboptimality bound). It is a standard regularity condition in the analysis of likelihood-based generative models and is not used elsewhere in the paper.

D Proofs

This appendix provides proofs of Theorem 1 (§D.2) and Proposition 1 (§D.3), together with the derivation of the Fisher-weighted criteria of Eq. (12) from the second-order expansion in §D.1. We work under the standard assumption that p_θ matches the data-generating distribution p^* on the conditional sets that appear in our analysis (well-specification); this is the regime in which information-theoretic claims about p^* translate to operational claims about p_θ.

D.1 Derivation of the Fisher-Weighted Criteria

We begin with the second-order expansion underlying Eq. (12), since both Theorem 1 and Proposition 1 rely on it.

Let $\ell_t = h_t^{(L)} \in \mathbb{R}^d$ denote the final-layer residual at position t, so that $p_\theta(x_t \mid \cdot)$ is determined by ℓ_t. Consider an intervention that ablates a subset $\mathcal{I}$ of token-wise contributions, which corresponds to perturbing ℓ_t by $\Delta := -\sum_{i \in \mathcal{I}} \phi_{i \to t}$. Define

$$F(\Delta) := D_{\mathrm{KL}}\big(p_\theta(\cdot \mid \ell_t) \,\big\|\, p_\theta(\cdot \mid \ell_t + \Delta)\big). \tag{28}$$

Standard properties of KL give $F(0) = 0$ and $\nabla F(0) = 0$, so a second-order Taylor expansion around $\Delta = 0$ yields

$$F(\Delta) = \tfrac{1}{2}\, \Delta^\top \mathcal{F}_t\, \Delta + o(\|\Delta\|^2), \tag{29}$$

where $\mathcal{F}_t = \nabla^2 F(0) = \mathbb{E}_{x_t \sim p_\theta(\cdot|\ell_t)}\big[\nabla_{\ell_t} \log p_\theta(x_t \mid \ell_t)\, \nabla_{\ell_t} \log p_\theta(x_t \mid \ell_t)^\top\big]$ is the Fisher information matrix of the output distribution with respect to the residual stream. Substituting $\Delta = -\sum_{i \in \mathcal{I}} \phi_{i \to t}$ gives

$$D_{\mathrm{KL}}^{\mathrm{ablate}(\mathcal{I})} = \tfrac{1}{2} \Big\| \sum_{i \in \mathcal{I}} \phi_{i \to t} \Big\|_{\mathcal{F}_t}^2 + o\Big(\Big\| \sum_{i \in \mathcal{I}} \phi_{i \to t} \Big\|^2 \Big), \tag{30}$$

which is the approximation appearing in §4.1.

Necessity bound. Definition 3-(ii) requires that for every $v \in S_t^\star$,

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid S_t^\star) \,\big\|\, p_\theta(x_t \mid S_t^\star \setminus \{v\}, \mathrm{do}(v = z_v))\big) \geq \eta. \tag{31}$$

Removing the variable v from $S_t^\star$ corresponds (under the local additivity of §4.1) to ablating $\phi_{v \to t}$, so $\mathcal{I} = \{v\}$ in Eq. (30). The Necessity condition then reads

$$\tfrac{1}{2}\, \phi_{v \to t}^\top \mathcal{F}_t\, \phi_{v \to t} \geq \eta + o(1) \quad \Longleftrightarrow \quad \phi_{v \to t}^\top \mathcal{F}_t\, \phi_{v \to t} \geq 2\eta, \tag{32}$$

matching the Necessity branch of Eq. (12).

Sufficiency bound. Definition 3-(i) requires that for every $U \subseteq \mathbf{x}_{-t} \setminus S_t^\star$,

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid S_t^\star) \,\big\|\, p_\theta(x_t \mid S_t^\star, \mathrm{do}(U = z_U))\big) \leq \epsilon. \tag{33}$$

The worst case over U is $U = \mathbf{x}_{-t} \setminus S_t^\star$. Setting $\mathcal{I} = \mathbf{x}_{-t} \setminus S_t^\star$ in Eq. (30) gives

$$\tfrac{1}{2} \Big\| \sum_{i\,:\,x_i \notin S_t^\star} \phi_{i \to t} \Big\|_{\mathcal{F}_t}^2 \leq \epsilon + o(1) \quad \Longleftrightarrow \quad \Big\| \sum_{i\,:\,x_i \notin S_t^\star} \phi_{i \to t} \Big\|_{\mathcal{F}_t}^2 \leq 2\epsilon, \tag{34}$$

matching the Sufficiency branch of Eq. (12).

D.2 Proof of Theorem 1

Proof. Under well-specification, $p_\theta(\cdot \mid x_S) = p^*(\cdot \mid x_S)$ for the conditional sets that appear below. We argue by cases.

Case 1: $\mathcal{S}_t > 0$. By definition, $\mathcal{S}_t = I_{p^*}(x_t;\, x_{>t} \mid x_{<t}) > 0$ implies

$$p^*(x_t \mid x_{<t}) \neq p^*(x_t \mid x_{<t}, x_{>t}) \quad \text{(in the } x_{>t}\text{-mean-square sense)}, \tag{35}$$

i.e., the $x_{>t}$ component carries non-trivial predictive information about x_t given $x_{<t}$. By the chain rule of KL,

$$D_{\mathrm{KL}}\big(p^*(x_t \mid x_{<t}, x_{>t}) \,\big\|\, p^*(x_t \mid x_{<t})\big) > 0 \tag{36}$$

on a set of positive measure. Since $S_t^\star$ contains a sufficient set of predictors (Definition 3-(i)), the predictive distribution $p_\theta(x_t \mid x_{S_t^\star})$ matches $p^*(x_t \mid x_{<t}, x_{>t})$ up to the tolerance ϵ, while $p_\theta(x_t \mid x_{<t})$ ignores $x_{>t}$. Therefore

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{<t}) \,\big\|\, p_\theta(x_t \mid x_{S_t^\star})\big) \geq \mathcal{S}_t - \epsilon > 0, \tag{37}$$

provided $\epsilon < \mathcal{S}_t$.

Case 2: $\exists R_t \subset x_{<t}, \mathcal{N}_t \leq \epsilon$. By definition, R_t contributes negligible information about x_t given $x_{<t} \setminus R_t$. Yet $R_t \subseteq P(x_t) = x_{<t}$ is part of the AR parent set, so $p_\theta(x_t \mid x_{<t})$ unnecessarily conditions on R_t. The minimal set $S_t^\star$, by Definition 3-(ii), excludes R_t since each $v \in R_t$ would violate Necessity. Hence $x_{S_t^\star}$ and $x_{<t}$ index distinct conditioning sets, and $p_\theta(x_t \mid x_{<t})$ inherits the redundancy of R_t. The KL divergence

$$D_{\mathrm{KL}}\big(p_\theta(x_t \mid x_{<t}) \,\big\|\, p_\theta(x_t \mid x_{S_t^\star})\big) > 0 \tag{38}$$

follows because conditioning on redundant variables yields a strictly different predictive distribution unless the model is invariant to R_t, which would contradict $R_t \subset x_{<t}$ entering the parent set non-trivially. □ □

Remark. The bound in Case 1 simplifies to $D_{\mathrm{KL}} \geq \mathcal{S}_t - \epsilon$, giving a quantitative refinement of the qualitative claim $D_{\mathrm{KL}} > 0$ in the statement.

D.3 Proof of Proposition 1

Let $C_j^{\mathrm{AR}} := \{i : i < j\}$ denote the default AR prefix. Replacing the AR prefix by the refined parent set $\hat{S}_j$ changes the residual stream at position j by

$$\Delta_j = \sum_{i \in \hat{S}_j} \phi_{i \to j} - \sum_{i \in C_j^{\mathrm{AR}}} \phi_{i \to j}. \tag{39}$$

We decompose this shift into two parts:

$$\Delta_j = \underbrace{\sum_{i \in \hat{S}_j \setminus C_j^{\mathrm{AR}}} \phi_{i \to j}}_{\Delta_j^+} - \underbrace{\sum_{i \in C_j^{\mathrm{AR}} \setminus \hat{S}_j} \phi_{i \to j}}_{\Delta_j^-}. \tag{40}$$

Here, Δ_j^+ denotes the contribution of newly admitted non-prefix parents, and Δ_j^- denotes the contribution of prefix tokens removed by refinement.

Under the local second-order KL approximation,

$$D_{\mathrm{KL}}\big(p_\theta(x_j \mid x_{\hat{S}_j}) \parallel p_\theta(x_j \mid x_{<j})\big) = \frac{1}{2}\|\Delta_j\|_{\mathcal{F}_j}^2 + o(1). \tag{41}$$

Therefore, it remains to lower-bound $(\|\Delta_j\|_{\mathcal{F}_j}^2$.

By assumption, the newly admitted parents have non-negligible aggregate effect: $\|\Delta_j^+\|_{\mathcal{F}_j}^2 \geq 2\eta$.

Meanwhile, the removed prefix contribution is controlled by the sufficiency residual: $\|\Delta_j^-\|_{\mathcal{F}_j}^2 \leq 2\epsilon$. We further assume bounded cross-cancellation:

$$\langle \Delta_j^+, \Delta_j^- \rangle_{\mathcal{F}_j} \leq \|\Delta_j^-\|_{\mathcal{F}_j}^2 + o(1). \tag{42}$$

This condition states that the removed prefix contribution cannot fully cancel the newly admitted causal contribution.

Using $\Delta_j = \Delta_j^+ - \Delta_j^-$, we have

$$\|\Delta_j\|_{\mathcal{F}_j}^2 = \|\Delta_j^+\|_{\mathcal{F}_j}^2 + \|\Delta_j^-\|_{\mathcal{F}_j}^2 - 2\langle \Delta_j^+, \Delta_j^- \rangle_{\mathcal{F}_j}.$$

By the bounded-cancellation condition,

$$\|\Delta_j\|_{\mathcal{F}_j}^2 \geq \|\Delta_j^+\|_{\mathcal{F}_j}^2 - \|\Delta_j^-\|_{\mathcal{F}_j}^2 - o(1).$$

Substituting the two bounds above yields

$$\|\Delta_j\|_{\mathcal{F}_j}^2 \geq 2\eta - 2\epsilon - o(1).$$

Finally,

$$D_{\mathrm{KL}}\big(p_\theta(x_j \mid x_{\hat{S}_j}) \parallel p_\theta(x_j \mid x_{<j})\big) = \frac{1}{2}\|\Delta_j\|_{\mathcal{F}_j}^2 + o(1) \geq \eta - \epsilon - o(1).$$

This completes the proof. □

E Implementation Details and Additional Results

E.1 Implementation Details

All experiments use RandAR-XL (0.7B), a 36-layer autoregressive transformer (dim=1280, 20 heads) with a LlamaGen VQ tokenizer (16×16 grid, codebook size 16,384). We evaluate class-conditional generation on ImageNet 256×256 (1,000 classes) and report FID, sFID, IS, Precision, and Recall against the standard ImageNet-10K reference. All experiments run on a single NVIDIA A100 GPU using BF16 mixed precision.

Algorithm 1 Dynamic Causal Structure Discovery for AR Generation

Require: AR model p_θ, condition c, sequence length T, check interval K
Require: Thresholds τ (Sufficiency), θ_{cau} (Score), δ (Necessity); stabilizer ε_0
Ensure: Refined sequence $x_{1:T}$

$\mathcal{P}_0 \leftarrow \emptyset$
for $t = 1$ to T **do**
  $x_t \sim p_\theta(\cdot \mid x_{<t})$ ▷ *Standard AR forward pass*
  $P_{\text{old}}(t) \leftarrow p_\theta(x_t \mid x_{<t})$
  $\mathcal{P}_t \leftarrow \mathcal{P}_{t-1} \cup \{x_t\}$
  if $t \bmod K = 0$ **then** ▷ *Trigger structure discovery*
    for each $j < t$ **do**
      $\mathcal{A}_j \leftarrow \mathcal{P}_t \setminus \{x_j\}$ ▷ *Expanded context*
      $P_{\text{pool}}(j) \leftarrow p_\theta(x_j \mid \mathcal{A}_j)$
      $x_j^{\text{full}} \leftarrow \arg\max_v p_\theta(v \mid \mathcal{A}_j)$
      if $P_{\text{old}}(j)/(P_{\text{pool}}(j) + \varepsilon_0) \leq \tau$ **or** $x_j^{\text{full}} = x_j$ **then**
        continue ▷ *Stage 1: Detect Sufficiency Violation*
      $\mathcal{C}_j \leftarrow \{i \in \mathcal{A}_j : \text{score}(i \to j) > \theta_{\text{cau}}\}$ ▷ *Stage 2: Ranking surrogate*
      $S \leftarrow \emptyset$
      for $i \in \mathcal{C}_j$ in descending order of $\text{score}(i \to j)$ **do**
        if $H_\theta(x_j \mid S \cup \{i\}) < H_\theta(x_j \mid S) - \delta$ **then**
          $S \leftarrow S \cup \{i\}$ ▷ *Stage 2: Verify Necessity*
      $\hat{S}_j \leftarrow S$
      $x'_j \sim p_\theta(\cdot \mid x_{\hat{S}_j})$ ▷ *Stage 3: Masked resampling*
      $x_j \leftarrow x'_j$ ▷ *Accept update (Prop. 1)*
return $x_{1:T}$

E.2 Evaluation protocol

We use 50K generated images for the main comparison and 5K for the hyperparameter ablation. All other generation-quality evaluations use 10K generated images. Smaller sample sets are used in ablations and controlled comparisons to reduce evaluation cost.

E.3 Detailed Inference Procedure

For completeness, we provide the full inference-time procedure of Dynamic Causal Structure Discovery in Algorithm 1. The algorithm operationalizes the three stages introduced in §4.2: detecting sufficiency violations, constructing minimal necessary parent sets, and refining predictions under the discovered causal structure.

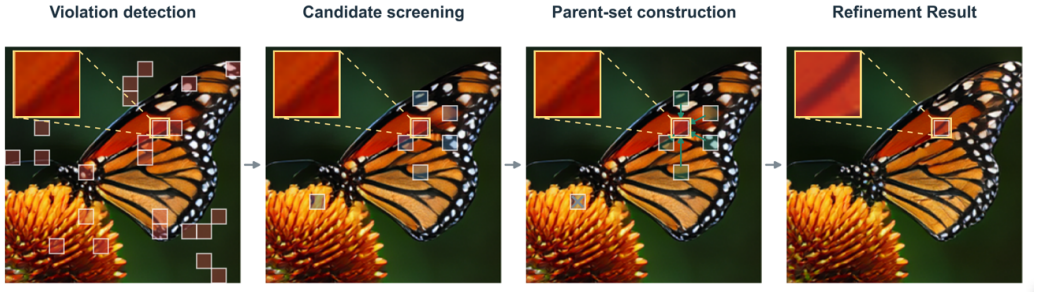


Figure 9: Schematic illustration of the token-refinement pipeline: violation detection, candidate screening, parent-set construction, and refinement.



Figure 10: Additional qualitative comparisons across diverse ImageNet categories. **Rows 1 and 3**: baseline results generated without refinement. **Rows 2 and 4**: refined results using our method. All pairs share the same class condition and random seed.

F Additional Qualitative Examples

Visualization of the Refinement Pipeline. Figure 9 schematically illustrates the refinement of a selected token. Red overlays indicate flagged positions, while the yellow outline identifies the target token. Blue boxes denote candidate parents. Green boxes and arrows indicate retained parents, whereas the gray cross marks a rejected candidate. The final panel illustrates the local refinement, with magnified patches highlighting the target region.

Class-Conditional Comparison. To further demonstrate the effectiveness of our method, we provide more side-by-side comparisons across diverse ImageNet categories in Figure 10. The samples cover a broad range of object types, including birds, mammals, insects, vehicles, and food. Compared with the baseline (RandAR), the refined results consistently exhibit sharper local details and more coherent global structure, while preserving the original semantic content. This further confirms that suppressing redundant dependencies and enforcing sufficient contextual support benefits AR image generation across a wide spectrum of visual scenes.

Progressive Refinement Visualization. Figure 11 provides additional examples of the progressive refinement process. From left to right, we visualize the generated images after an increasing number of refinement steps, with red patches marking the tokens that have been refined up to the current step. As more sufficiency violations are detected with the accumulating context, additional tokens are progressively refined, leading to clearer local textures and more realistic global structure—e.g., finer terrain in the valley, more coherent rocket exhaust, and sharper patterns on the hot-air balloon. This further confirms that our framework incrementally corrects suboptimal AR factorizations as generation proceeds.

Qualitative Comparisons with RandAR and RAR. Figure 12 presents visual comparisons on three ImageNet classes, with RandAR, DCD, and RAR shown from top to bottom. In the displayed husky examples, DCD produces more coherent eye and muzzle geometry than RandAR and RAR. The volcano example further illustrates clearly defined lava flows and eruption details. RAR samples are reproduced from the original paper [23] for class-matched comparison.



Figure 11: Progressive visualization of the causal refinement process on additional samples. From left to right, the number of refined tokens increases as generation proceeds. **Red patches** indicate tokens refined up to the current step. The accumulation of refinements progressively enhances visual fidelity.

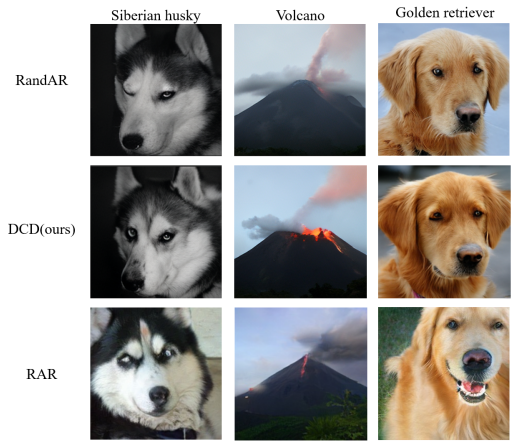


Figure 12: Qualitative comparisons on three ImageNet classes. From top to bottom: RandAR, DCD (ours), and RAR.

F.1 Evaluation Metrics for Causal Graph Discovery

Evaluation Metrics. To quantitatively assess how faithfully our Fisher-weighted criteria recover the underlying causal structure, we adopt four edge-level metrics that compare the predicted causal graph against the ground-truth graph obtained by brute-force do-intervention [22]. Let $\mathcal{G}^\star = (\mathcal{V}, \mathcal{E}^\star)$ denote the ground-truth causal graph and $\hat{\mathcal{G}} = (\mathcal{V}, \hat{\mathcal{E}})$ the predicted graph, with edge weights derived from the Fisher-weighted contribution scores. The four metrics together capture both edge-level correctness and ranking fidelity.

Weighted Structural Hamming Distance (WSHD). WSHD measures the weighted edge-level discrepancy between $\hat{\mathcal{G}}$ and $\mathcal{G}^\star$, counting mis-added, missing, and reversed edges with their associated causal weights. A lower WSHD indicates a more accurate recovery of both the topology and the relative strength of causal dependencies.

Weighted Normalized SHD (WNSHD). WNSHD further normalizes WSHD by the total edge weight of $\mathcal{G}^\star$, eliminating the influence of graph size and yielding a scale-invariant distributional discrepancy between two causal graphs. This makes WNSHD suitable for comparing graphs of varying token lengths.

Weighted F1 (WF1). WF1 treats causal edge prediction as a binary classification problem and computes the weighted harmonic mean of precision and recall. It reflects the overall accuracy of causal structure identification, balancing false-positive and false-negative edges under causal-strength weighting.

Area Under the Precision-Recall Curve (AUPRC). AUPRC computes the area under the PR curve obtained by ranking predicted edges according to their Fisher-weighted confidence scores. It measures the ranking fidelity of the predicted causal edges and is particularly suitable for the sparse-positive nature of causal graphs, where the number of true parent edges is much smaller than the total number of candidate edges.

G Limitations

Although our framework improves dependency modeling in autoregressive visual generation, several limitations remain. First, Dynamic Causal Structure Discovery introduces additional inference-time overhead due to repeated dependency evaluation and refinement, which may limit efficiency for large-scale generation. Second, our method relies on local approximations of causal interventions through Fisher-based criteria, and therefore may not fully capture higher-order or long-range dependencies in highly complex scenes. Third, while our experiments demonstrate consistent improvements on autoregressive image generation benchmarks, the scalability of the framework to larger multimodal or video generation models remains to be further explored.

H Broader Impacts

Ethical Impacts. Our work focuses on improving dependency modeling and generation quality in autoregressive visual generation. We do not introduce new datasets containing personal or sensitive information, and our method operates directly on pretrained generative models without modifying training data.

Expected Societal Implications. This work contributes to a better understanding of the structural limitations of autoregressive generation and provides a causal perspective for improving generative modeling. We hope our framework encourages future research on more interpretable, adaptive, and data-aligned generation mechanisms. At the same time, as with other generative modeling technologies, improvements in visual generation quality may also increase the potential misuse of synthetic media, highlighting the importance of responsible deployment and usage.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed about limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete proof in our paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully disclose all the information needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our source code and reproduction instructions are publicly available at `https://github.com/Songliang112/DCD`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provide sufficient experimantal details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report experimental results with average of multiple running.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide compute resource in our paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conduct our research in the paper conform.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed about potential societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credited assets used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our released code is documented in the GitHub repository, including installation instructions and commands for inference and evaluation: https://github.com/Songliang112/DCD.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in our research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.