RETHINKING REPRESENTATION ALIGNMENT IN AUTOREGRESSIVE IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation alignment improves image generation quality for both diffusion and autoregressive (AR) models by supervising intermediate latents with features from pretrained visual encoders. Nevertheless, in AR models, the utility of representation alignment depends on the causal position of the supervised state. Later states observe more image content but can influence fewer subsequent predictions, so higher representation fidelity does not necessarily improve generation. To this end, we introduce **Causally Accessible Representation Alignment (CARA)**, grounded in a theoretical and empirical investigation of how visual representation learning can facilitate future predictions in autoregressive generation. Our theoretical and empirical analyses suggest that alignment before image generation encourages learning visual information predictable from the condition and provides useful context for subsequent predictions. Mechanistic analyses support using a global teacher target to align the final-layer representation of a learnable prefix token, whose representations remain accessible throughout autoregressive decoding. Experiments demonstrate notable improvements in class-to-image and text-to-image generation tasks across diverse model sizes. Extensive ablations support aligning a dedicated prefix state at the final layer, while cache interventions show that this state helps predict image tokens beyond the first. These findings position causal role as a design principle for representation alignment and motivate learning representations according to both the information they can encode and the future predictions they can support.

1 INTRODUCTION

Representation alignment Yu et al. (2024b) accelerates diffusion model training by aligning intermediate hidden states with features extracted from frozen self-supervised visual encoders, such as DINO Oquab et al. (2024); Siméoni et al. (2025), through dense patch-wise supervision. Building on this idea, subsequent works Yao et al. (2025); Leng et al. (2025); Zheng et al. (2025) extend representation supervision beyond the intermediate hidden states to generative latents and visual tokenizers. This dense alignment paradigm suits diffusion models, where each spatial latent can integrate contextual information from the noisy image while preserving an explicit correspondence with its teacher patch.

For autoregressive (AR) models, researchers have explored representation alignment through both spatial and global supervision. For example, iFSQ and GEAR Lin et al. (2026a;b) align image-token states with dense visual features, whereas ARRA Xie et al. (2025) supervises next-token predictor states with a global teacher representation and also demonstrates the effectiveness of a fixed representation token. Both formulations encounter the issue of *causal information-set mismatch*, where teacher targets are encoded from full-image context while AR states have access only to their own causal prefix. The supervised position therefore determines both the information available for reconstructing the target and the subsequent predictions that can access the resulting representation. Moving this position later provides more image evidence for target reconstruction but reduces its downstream accessibility. Consequently, improved teacher fidelity need not translate into improved generation, motivating the joint consideration of representation quality and causal role. The dependence of generative utility on causal position motivates designing aligned representations as context for subsequent predictions. Representation Forcing (RF) Wang et al. (2026) implements such prospective conditioning by predicting discrete representation tokens before a separate pixel-

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

generation stage. We investigate how representation supervision can serve this role within a standard AR backbone through the joint choice of causal placement and alignment depth. This perspective extends alignment design beyond teacher-target granularity to the supervised state's information set and downstream accessibility. Further discussion of related work appears in Appendix A.

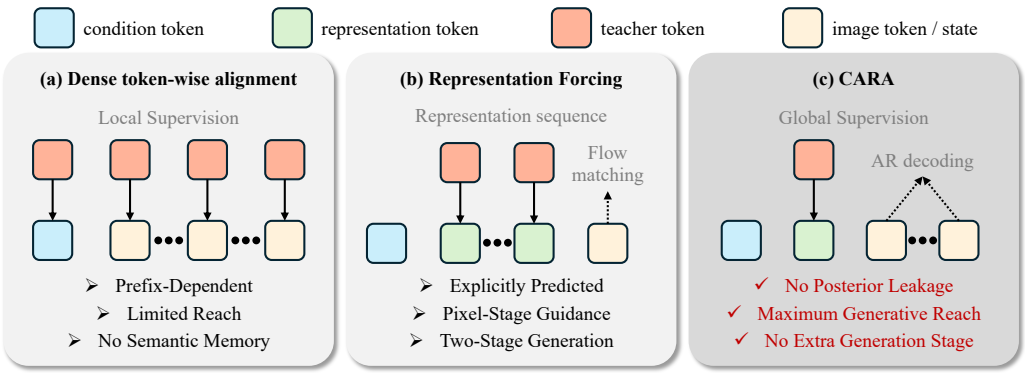

Figure 1: **Architectural comparison**. (a) Dense patch-wise alignment supervises image states with position-dependent context and downstream accessibility. (b) Representation Forcing predicts a representation sequence that conditions subsequent pixel synthesis. (c) **CARA** applies deep global supervision to a condition-only prefix state accessible to every image prediction within the original decoding pipeline.

To this end, we propose **Causally Accessible Representation Alignment (CARA)**. As illustrated in Figure 1, CARA inserts a learnable token between the condition and image sequence and aligns its final-block representation with a global teacher target. The causal mask gives this token access to the complete condition while excluding all image tokens. Under the isolated cosine-alignment objective, the optimal predictor follows the conditional mean direction of normalized teacher features, providing a theoretical interpretation of the aligned state as a condition-dependent semantic prior. Unlike supervision distributed over image-predictor states Lin et al. (2026a;b); Xie et al. (2025), this construction fixes the supervised state's information set independently of the progressively revealed image prefix. Alignment therefore encourages the state to capture semantics predictable from the condition, rather than reconstruct sample-specific content already available through teacher forcing. Prospective placement also makes the token's layerwise states accessible to all subsequent image predictions, preserving their potential influence throughout decoding. Final-block supervision shapes this prefix computation through gradients across the backbone, combining deep representation learning with persistent semantic context. These properties give alignment a prospective role by using full-image teacher supervision to learn a condition-dependent state accessible to subsequent predictions. CARA preserves the original generation pipeline without an additional representation-prediction stage Wang et al. (2026).

Extensive experiments on class-to-image and text-to-image generation verify the effectiveness of CARA across various model sizes. For example, on LlamaGen-L, CARA reduces FID from 3.2919 to 3.0426 and improves over a matched extra-token control by 0.3815 FID, isolating the contribution of representation supervision within the same prefix architecture. Additionally, comparisons with ARRA-style variants and position-depth ablations further support combining deep alignment with prospective placement. Under the same identity-RoPE configuration, teacher fidelity increases from prospective to terminal placement while generation quality declines, showing that target reconstruction alone is insufficient to characterize generative utility. Furthermore, improvements on GenEval Ghosh et al. (2023) and DPG-Bench Hu et al. (2024) with both SimpleAR-0.5B and SimpleAR-1.5B Wang et al. (2025) extend these benefits to text-conditioned generation. Collectively, these findings identify the causal role of the supervised state as a central consideration in representation alignment. The broader implication is that alignment should be designed around the information a representation can encode and the future predictions it can support, with supervision depth and causal accessibility considered jointly.

2 METHOD

2.1 PROBLEM SETUP AND CAUSAL ROLE

Let $\mathbf{c} = (c_1, \ldots, c_M)$ denote the conditioning sequence and $\mathbf{x} = (x_1, \ldots, x_N)$ the discrete image tokens produced by a frozen visual tokenizer Van Den Oord et al. (2017); Esser et al. (2021). A causal Transformer factorizes the conditional image distribution and minimizes the corresponding negative log-likelihood:

$$p_\theta(\mathbf{x} \mid \mathbf{c}) = \prod_{i=1}^{N} p_\theta(x_i \mid \mathbf{c}, \mathbf{x}_{<i}), \qquad \mathcal{L}_{\mathrm{AR}} = -\sum_{i=1}^{N} \log p_\theta(x_i \mid \mathbf{c}, \mathbf{x}_{<i}). \tag{1}$$

The conditioning sequence $\mathbf{c}$ is a class embedding in the class-to-image setting and a sequence of projected text features in the text-to-image setting.

A frozen visual encoder extracts a target $\mathbf{z}$ from the complete training image. This target encodes full-image context, whereas each causal hidden state depends on its prefix. For a supervised state following the first k image tokens, we define its information set, image visibility, and downstream accessibility as

$$\mathcal{F}_k = (\mathbf{c}, x_1, \ldots, x_k), \qquad v(k) = k, \qquad a(k) = N - k, \tag{2}$$

where $v(k)$ counts visible image tokens and $a(k)$ counts image predictions with causal access to the position, including its own next-token prediction when present. Under teacher forcing, the image prefix consists of ground-truth tokens, with attention restricted to $\mathcal{F}_k$ by the causal mask. Equation 2 captures the placement trade-off: later states have access to more sample-specific evidence but can influence fewer subsequent predictions.

2.2 THE CAUSAL INFORMATION-SET MISMATCH

The prediction problem induced by a full-image target depends on the causal position of the supervised state. Let $\bar{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$ and consider a unit-norm predictor $\mathbf{q}(\mathcal{F}_k)$ that minimizes expected cosine distance. For a nonzero conditional mean, the pointwise optimum is

$$\mathbf{q}_k^*(\mathcal{F}_k) = \frac{\mathbb{E}[\bar{\mathbf{z}} \mid \mathcal{F}_k]}{\|\mathbb{E}[\bar{\mathbf{z}} \mid \mathcal{F}_k]\|_2}, \tag{3}$$

obtained by maximizing $\mathbf{q}(\mathcal{F}_k)^\top \mathbb{E}[\bar{\mathbf{z}} \mid \mathcal{F}_k]$. At $k = 0$, conditioning exclusively on $\mathbf{c}$ yields a condition-dependent semantic prior. For $k > 0$, the estimate additionally incorporates sample-specific information from the observed image prefix. High teacher similarity near the end of the sequence can therefore reflect retrospective reconstruction at a position with little remaining influence on generation.

Equation 3 characterizes the isolated alignment objective, whereas the finite-capacity model is optimized jointly with $\mathcal{L}_{\mathrm{AR}}$. The analysis distinguishes two determinants of representation quality: alignment fidelity depends on the available information, while generative utility additionally depends on downstream accessibility.

2.3 CARA AS PROSPECTIVE ALIGNMENT

CARA implements prospective alignment at $k = 0$ by inserting a learnable representation token r between the condition and image sequence:

$$\mathbf{s}_{\mathrm{CARA}} = [c_1, \ldots, c_M, r, x_1, \ldots, x_N]. \tag{4}$$

The representation token is a learned input embedding and is not itself a token to be generated. The causal mask permits its hidden state to attend to the complete condition while excluding all image tokens. Conversely, every image state has access to the representation-token position. Across the LlamaGen baseline, extra-token control, and CARA, non-spatial prefix positions, including r when present, use identity rotary encoding with $\cos\theta = 1$ and $\sin\theta = 0$. Zero rotary phase preserves the prefix query and key vectors, while the N image tokens retain their original two-dimensional rotary coordinates.

The added position preserves the N next-token predictions in Equation 1. During teacher-forced training, the state at r predicts x_1, and the state at x_i predicts x_{i+1} for $i < N$. The cross-entropy target at the representation-token position is x_1, rather than r. The resulting factorization is:

$$p_\theta(\mathbf{x} \mid \mathbf{c}) = \prod_{i=1}^{N} p_\theta(x_i \mid \mathbf{c}, r, \mathbf{x}_{<i}), \tag{5}$$

where r denotes the learned prefix embedding and its condition-dependent hidden states.

2.3.1 Global representation target

Let E_ϕ be a frozen visual teacher. Each image is assigned a global target from the DINOv3 class-token output Siméoni et al. (2025), or from pooled features for teachers without a class token. Targets are extracted from the complete training image. The ImageNet pipeline stores discrete visual codes, which a frozen tokenizer decodes before teacher feature extraction, yielding $\mathbf{z} = E_\phi(\mathrm{Dec}(\mathbf{x}))$. For the reference DINOv3 configuration, the reconstruction is resized to 224×224 and normalized with ImageNet channel statistics. Tokenizer decoding and teacher feature extraction are performed without gradient computation.

2.3.2 Alignment objective

Let $\mathbf{h}_r^{(L)}$ be the representation-token state after the final Transformer block. CARA applies the backbone's shared output RMSNorm and a two-layer projection head g_ψ with a SiLU nonlinearity:

$$\widehat{\mathbf{z}}_r = g_\psi\left(\mathrm{RMSNorm}_{\mathrm{out}}(\mathbf{h}_r^{(L)})\right). \tag{6}$$

For a minibatch of size B, we combine the cosine-alignment loss with the AR objective:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{B}\sum_{b=1}^{B}\left(1 - \frac{\widehat{\mathbf{z}}_{r,b}^{\top}\mathbf{z}_b}{\|\widehat{\mathbf{z}}_{r,b}\|_2\|\mathbf{z}_b\|_2}\right), \tag{7}$$

$$\mathcal{L}_{\mathrm{CARA}} = \mathcal{L}_{\mathrm{AR}} + \lambda\mathcal{L}_{\mathrm{align}}. \tag{8}$$

We use $\lambda = 1$ in the reference configuration. The alignment loss updates the representation token, projection head, and AR backbone. The visual teacher and tokenizer remain frozen.

At inference, the model processes the condition and representation token before predicting the first image token, caching their layerwise keys and values for subsequent decoding. The teacher and projection head are discarded, while the representation token remains as an additional context position throughout image-token sampling. Subsequent predictions access its cached layerwise states, which are shaped by alignment during training, rather than receiving the final aligned output directly.

2.4 Controlled Causal Placements and Alignment Variants

For intermediate placement, r is inserted after the first k image tokens:

$$\mathbf{s}_k = [c_1, \ldots, c_M, x_1, \ldots, x_k, r, x_{k+1}, \ldots, x_N]. \tag{9}$$

The resulting state has $v(k) = k$ and $a(k) = N - k$. We assign r zero rotary phase while preserving the original image-token coordinates. Logits on either side of the insertion are concatenated to retain exactly N image-token targets. The state at r predicts x_{k+1} when $k < N$. Retrospective alignment supervises the terminal image-token state without adding a position after x_N. This state incorporates all N image tokens under teacher forcing and lies beyond all image-prediction positions.

These placements use the same global teacher target and cosine objective. The dense-alignment variant supervises the corresponding N causal image states with N teacher patch features and introduces no representation token. For the ImageNet experiments, a 256×256 teacher input produces a 16×16 patch grid with a one-to-one correspondence to the 256 visual tokens.

Intermediate placement is evaluated using the normalized final-block state. For prospective and terminal placement, we examine states at one-third, two-thirds, and full network depth. The shared feature path applies the backbone's output RMSNorm at the selected depth, whereas the bypass path uses an independently learned RMSNorm. These paths distinguish the effects of supervision depth and normalization. Experiments with the earlier rotary configuration, including target-granularity, loss-weight, teacher, and bypass-path ablations, appear in Appendix D.

2.4.1 ARRA-STYLE ALIGNMENT

We compare CARA with layer-1 global alignment on a fixed `[REP]` state and on ordinary image-token predictor states, denoted `[HYBNEXT]` following ARRA Xie et al. (2025). For token-wise global alignment, each selected predictor state receives the same full-image target through a two-layer GELU adaptor. The predictor for x_i conditions on $(\mathbf{c}, \mathbf{x}_{<i})$, giving each position a different image information set. A separate variant follows the public ARRA implementation's last-256-state alignment window. On SimpleAR, the layer-1 variant supervises the states responsible for predicting image tokens after the first Transformer block, inserts no additional representation token, and preserves the original inference architecture.

3 EXPERIMENTS

3.1 CLASS-TO-IMAGE GENERATION

3.1.1 EXPERIMENTAL SETUP

We evaluate CARA based on LlamaGen Sun et al. (2024) on the ImageNet-256 benchmark, with identity-RoPE for non-spatial prefix positions, final-block alignment to a frozen DINOv3-H+ teacher Siméoni et al. (2025), shared RMSNorm, and $\lambda = 1$. We report FID, IS, precision, and recall from 50K samples following Dhariwal & Nichol (2021). Other training and checkpoint details are provided in Appendix B.1.

3.1.2 MAIN RESULTS

Table 1 provides the quantitative comparison between CARA and representative image generation methods. CARA achieves lower FID and higher IS than the corresponding LlamaGen baseline at every model size, while maintaining or improving precision and recall. These gains span different model capacities, although their magnitude and expression across metrics vary with scale. CARA also strengthens LlamaGen's performance relative to the other methods listed in the table. CARA-XXL achieves lower FID than the larger LlamaGen-3B reference and comparable FID to TensorAR-XXL, while retaining the original visual-token decoding pipeline. These comparisons show that representation supervision can improve an existing AR backbone across model scales. The contribution of the alignment objective and the roles of supervision position and depth are examined through controlled experiments in Section 3.3.

3.2 TEXT-TO-IMAGE GENERATION

3.2.1 EXPERIMENTAL SETUP

We apply CARA to SimpleAR-0.5B-SFT and SimpleAR-1.5B-SFT Wang et al. (2025) and conduct evaluation on GenEval Ghosh et al. (2023), DPG-Bench Hu et al. (2024), and TIIF-Bench Wei et al. (2025). The token follows the text prompt and receives final-block supervision from DINOv3-H+. The ARRA-style Xie et al. (2025) comparison uses layer-1 token-wise global alignment. Full sampling settings, SimpleAR's fixed-template style augmentation, and checkpoint details are provided in Appendix B.2.

3.2.2 MAIN RESULTS

CARA improves on both SimpleAR SFT baselines across the benchmarks in Table 2. The gains on GenEval Ghosh et al. (2023) and DPG-Bench Hu et al. (2024) span object-level composition and adherence to detailed prompts, indicating benefits across complementary aspects of text-conditioned generation. On TIIF-Bench Wei et al. (2025), CARA improves instruction following for both short and long prompts at both model scales. The gain over the SFT baseline is larger for short prompts at 0.5B and for long prompts at 1.5B, showing that the benefit extends across instruction lengths while its magnitude depends on model scale and prompt format. These benefits remain present with the stronger 1.5B backbone, suggesting that CARA provides useful supervision beyond the capabilities acquired through the original SimpleAR SFT procedure.

Table 1: Quantitative results of class-conditional generation on ImageNet-256. Blank entries indicate unavailable measurements.

Model	Params	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Previous SOTA methods					
DiT-XL/2 Peebles & Xie (2023)	675M	2.27	278.20	0.83	0.57
LlamaGen-3B Sun et al. (2024)	3.1B	2.18	263.30	0.81	0.58
TensorAR-XXL Cheng et al. (2026)	1.4B	2.03	267.70	0.82	0.61
VAR-d30 Tian et al. (2024)	2.0B	1.92	323.10	0.82	0.59
MAR-H Li et al. (2024)	943M	1.55	303.70	0.81	0.62
RAR-XXL Yu et al. (2024a)	1.5B	1.48	326.00	0.80	0.63
xAR-H Ren et al. (2025)	1.1B	1.24	301.60	0.83	0.64
Controlled comparison on LlamaGen with identity-RoPE					
LlamaGen-B	111M	5.46	193.60	0.83	0.45
+ **CARA-B**	111M	**4.71**	228.29	0.86	0.45
LlamaGen-L	343M	3.80	248.28	0.83	0.52
+ **CARA-L**	343M	**3.04**	282.67	0.85	0.54
LlamaGen-XL	775M	2.62	244.10	0.80	0.57
+ **CARA-XL**	775M	**2.44**	297.93	0.84	0.57
LlamaGen-XXL	1.4B	2.34	253.90	0.81	0.60
+ **CARA-XXL**	1.4B	**2.02**	270.66	0.81	0.60

CARA also outperforms the evaluated layer-1 ARRA-style variants on all three benchmarks at both scales. At 0.5B, shallow ARRA-style alignment improves GenEval but reduces DPG-Bench, whereas at 1.5B it improves DPG-Bench with little change in GenEval. On TIIF-Bench, ARRA-style alignment yields mixed changes across short and long prompts at 0.5B and small improvements for both at 1.5B, while CARA achieves larger gains in all four settings. These results indicate broader improvements in composition, detailed prompt adherence, and instruction following under CARA. Because both designs use global visual supervision, this comparison favors CARA's combination of deep alignment and a dedicated prospective state to translate that supervision into prompt-conditioned generation. The resulting improvements extend to 4,096-token image sequences while retaining SimpleAR's decoding pipeline and a single aligned state irrespective of sequence length.

Table 2: Text-to-image generation on GenEval, DPG-Bench, and TIIF-Bench TestMini. SFT and CARA entries are reference results. Bold identifies the highest reported score within each model scale.

Model	Params	GenEval ↑	DPG-Bench ↑	TIIF-TestMini ↑	
				Short	Long
SimpleAR-0.5B-SFT	0.5B	0.53	79.34	0.510	0.517
+ ARRA-style layer-1	0.5B	0.55	78.49	0.514	0.515
+ **CARA**	0.5B	**0.57**	**79.92**	**0.562**	**0.541**
SimpleAR-1.5B-SFT	1.5B	0.61	80.11	0.533	0.516
+ ARRA-style layer-1	1.5B	0.61	80.87	0.541	0.523
+ **CARA**	1.5B	**0.63**	**81.32**	**0.576**	**0.586**

3.3 Ablation Studies

We verify the contribution of representation alignment through a matched extra-token control, vary causal placement and alignment depth, and compare ARRA-style alignment variants based on LlamaGen-L with identity-RoPE on ImageNet-256, evaluated on 50K generated samples at CFG 1.75, with FID as the primary metric and sFID, IS, precision, and recall as complementary measures. Appendix D provides teacher, loss weight, normalization path, and target granularity ablations un-

der the earlier rotary configuration. Additional ARRA-style comparisons and further analyses of predictive distributions and cache interventions are reported in Appendix E.

Table 3: Causal-role ablation on LlamaGen-L with identity-RoPE. v counts visible image tokens, and a counts predictions with access to the designated position. All aligned configurations use the same global target and final-block supervision. Bold marks the best metric.

Setting	v	a	**FID** ↓	**sFID** ↓	**IS** ↑	**Prec.** ↑	**Rec.** ↑
LlamaGEN-L Baseline	–	–	3.80	6.81	248.28	0.83	0.52
Extra-token control	0	256	3.42	6.96	291.89	0.86	0.52
CARA / front	0	256	**3.04**	**6.80**	282.67	0.85	0.54
Insert $k = 128$	128	128	3.18	7.11	276.33	0.85	0.53
Terminal/back	256	0	3.35	6.95	279.59	0.85	0.52

3.3.1 Causal placement and matched controls

The extra-token control shares CARA's prefix layout, identity-RoPE convention, image-token coordinates, token and backbone initialization, and optimization schedule. It is trained exclusively with $\mathcal{L}_{\mathrm{AR}}$, without a teacher or projection head. As shown in Table 3, token insertion alone yields worse FID than the unmodified baseline, whereas adding representation supervision improves both FID and recall under the same prefix architecture. The benefit therefore depends on how the additional state is trained, rather than its presence alone.

Across aligned placements, FID worsens as supervision moves from the front to the intermediate and terminal positions, despite the increasing amount of visible image content. More evidence for reconstructing the teacher target is thus insufficient to ensure better generation. The prospective state remains available to every image prediction, while intermediate placement restricts this range and terminal placement leaves no subsequent prediction to use the state directly. Terminal supervision can still influence generation through shared parameter updates, but lacks this direct contextual role. CARA's lower FID and higher recall, together with the extra-token control's higher IS and precision, also show that the benefit involves a different balance of distributional agreement and coverage across the evaluated metrics.

Table 4: Alignment depth and causal placement on LlamaGen-L with identity-RoPE. Deep, mid, and shallow columns report FID (↓) at full, two-thirds, and one-third network depth, respectively, using shared normalization. Positive gains indicate improvement with deeper alignment. Bold marks the best depth within each placement.

Placement	**Deep**	**Mid**	**Shallow**	**Shallow-to-Deep Gain** ↑	**Mid-to-Deep Gain** ↑
Front (CARA)	**3.04**	3.28	3.41	0.37	0.23
Back (terminal)	3.35	**3.34**	3.39	0.04	-0.01

3.3.2 Alignment depth and causal placement

Table 4 shows that prospective alignment improves progressively with supervision depth, whereas terminal alignment changes little and performs best at the middle depth. The relative ordering consequently depends on depth, with prospective placement outperforming terminal placement under middle and final-block supervision but slightly underperforming it under shallow supervision. This contrast supports a joint interpretation of alignment depth and causal role. Increasing depth does not expose the front state to additional image content, since its information set remains restricted to the condition. Instead, final-block supervision trains a longer prefix computation whose earlier layer-wise states can provide context to subsequent predictions. The observed improvement is consistent with learning more useful conditional context, while the weak response at the terminal position shows that applying the same supervision deeper in the network does not produce a comparable benefit for every state role.

Table 5: Unified role and depth comparison on LlamaGen-L with identity-RoPE. Prefix/Final corresponds to CARA in this experimental group. Bold marks the best value in each metric column.

Role	Depth	FID ↓	IS ↑	sFID ↓	Prec. ↑	Rec. ↑
Prefix	Layer-1	3.39	290.26	6.91	**0.86**	0.53
HYBNEXT	Layer-1	3.45	**290.93**	6.84	0.85	0.53
Prefix (CARA)	Final	**2.98**	278.35	**6.53**	0.84	0.54
HYBNEXT	Final	3.06	282.41	6.83	0.84	**0.54**
Normal	–	3.29	284.35	6.81	0.85	0.52

3.3.3 Supervised role and alignment depth

We compare Prefix and ARRA-style HYBNEXT supervision at the first and final Transformer blocks under a unified configuration, separate from the preceding placement experiments. Prefix aligns a dedicated prospective state, whereas HYBNEXT aligns ordinary next-token predictor states. All four models are evaluated from their final checkpoints using 50K generated images (Table 5). Layer-1 denotes the first block, distinct from the one-third-depth setting above.

Both roles improve with final-block supervision, changing from worse to better FID than the unaligned reference. The similar depth gains indicate that deep alignment benefits both state roles. Prefix retains a FID advantage at both depths and achieves lower sFID at the final block, where HYBNEXT yields higher IS, precision, and recall. These results support a shared benefit from deeper supervision alongside an additional Prefix advantage in FID and sFID. The comparison concerns complete role configurations, whose visibility, state identity, and accessibility remain coupled. Additional shallow ARRA-style variants also fail to improve baseline FID in a separate comparison reported in Appendix E.3.

3.4 Mechanistic Analysis

3.4.1 Representation fidelity versus generation

Across all three fidelity measures, agreement with the teacher increases from prospective to intermediate and terminal placement, while generation quality declines (Table 6; diagnostic details in Appendix C). This reversal shows that teacher fidelity alone is an insufficient indicator of generative utility. Later states can use observed image content to estimate the teacher target, whereas the prospective state's isolated alignment objective favors a semantic prior predictable from the condition (Equation 3). Its lower fidelity is compatible with providing useful context for subsequent predictions, while the terminal state has no remaining predictions to support directly. Representation fidelity should therefore be interpreted in relation to the supervised state's information set and role in generation.

Table 6: Representation fidelity and generation quality on the same LlamaGen-L identity-RoPE checkpoints. v and a denote image visibility and downstream accessibility. All three final-state fidelity metrics increase with image visibility, whereas FID worsens.

Placement	v	a	FID ↓	CKA ↑	kNN ↑	Probe cos. ↑
Prospective (CARA)	0	256	**3.04**	0.152	0.335	0.439
Intermediate	128	128	3.18	0.233	0.450	0.559
Terminal	256	0	3.35	**0.438**	**0.498**	**0.563**

3.4.2 Predictive distributions under alignment

We compare the four role and depth configurations on the same 1,024 ImageNet token sequences under teacher forcing, matching predictions by target token. Table 7 reports prediction loss and distributional differences relative to the unaligned model.

Table 7: Teacher-forced predictive diagnostics on the same 1,024 ImageNet token sequences. KL, RMSE, cosine similarity, and top-1 change compare each model with Normal at matched prediction targets.

Model	NLL ↓	ΔNLL ↓	KL	Centered-logit RMSE	Logit cosine	Top-1 change
HYBNEXT/Layer-1	7.13975	+0.00987	0.03100	0.22872	0.99314	30.54%
HYBNEXT/Final	7.14872	+0.01884	0.03988	0.29367	0.98912	34.21%
Prefix/Layer-1	7.13644	+0.00656	0.02677	0.21559	0.99396	28.98%
Prefix/Final (CARA)	7.13793	+0.00805	0.02866	0.22298	0.99268	29.67%

Final-layer alignment improves FID for both roles, although all aligned models have higher NLL than the unaligned reference. Compared with Final HYBNEXT, CARA achieves lower FID and sFID alongside lower NLL and smaller predictive deviations. However, proximity to the reference alone does not explain generation quality, since shallow Prefix changes predictions less but has worse FID. Further diagnostic details and position-wise results appear in Appendix E.1 and Appendix E.2.

3.4.3 Predictive utility of the cached state

We test whether subsequent predictions use the representation state by intervening on its layerwise cache within a fixed CARA identity-RoPE checkpoint. The first image token uses the normal prefix, and interventions begin with the second image-token prediction. Value-shuffle preserves the keys and permutes the values across samples, Value-zero retains the keys but sets the values to zero, and Read-mask prevents subsequent queries from attending to this position. Each intervention is compared with the same checkpoint using its unmodified cache.

Table 8: Cache interventions within a fixed CARA identity-RoPE checkpoint. ΔNLL is intervention NLL minus reference NLL, with positive values indicating impaired target-token prediction. KL and top-1 change measure predictive deviation from the unmodified checkpoint.

Intervention	NLL ↓	ΔNLL	95% CI	KL	Top-1 change
Value-shuffle	7.1645	+0.0092	[+0.0079, +0.0106]	0.0167	18.74%
Value-zero	7.3651	+0.2098	[+0.1850, +0.2371]	0.2429	51.68%
Read-mask	7.7337	+0.5784	[+0.5392, +0.6185]	0.6509	74.89%

All three interventions increase target-token NLL, with confidence intervals above zero (Table 8). The modest loss increase under Value-shuffle supports the predictive value of matching cached content to the current condition. Removing the value signal or blocking access causes greater impairment, although these operations also change value magnitudes or attention normalization. Because the model parameters and first image-token prediction are unchanged, the results show that the cached state contributes to subsequent predictions. Further interpretation of the intervention effects appears in Appendix E.4.

4 Conclusion

We introduced CARA, which applies deep global supervision to a dedicated, image-blind prefix state for autoregressive image generation. Under the isolated cosine-alignment objective, conditioning this state exclusively on the generation condition makes its optimal prediction a condition-dependent semantic prior. Prospective placement makes the state's layerwise representations available throughout decoding, connecting global supervision to semantic context that subsequent predictions can use. Experiments on class-to-image and text-to-image generation demonstrate improved performance, while positional and depth ablations support the combination of deep alignment and prospective placement. CARA thus motivates a view of representation alignment in which teacher fidelity is interpreted together with the supervised state's information set, downstream accessibility, and alignment depth to understand its role in generation.

AI USE STATEMENT

In this work, we used generative AI tools to obtain feedback on research methodology and experimental comparisons, assist with translation, and interpret experimental results. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, implement methods, generate synthetic datasets, assist in writing proofs, or propose or refine hypotheses. Dataset cleaning and reformatting, as well as qualitative and thematic data analysis, are not applicable to this work. We have reviewed all AI-assisted work, checking the descriptions of methods, numerical comparisons, and interpretations against the experimental evidence and revising the text where necessary. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

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 *International Conference on Learning Representations*, volume 2026, pp. 8737–8749, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Bin Lin, Zongjian Li, Yuwei Niu, Kaixiong Gong, Yunyang Ge, Yunlong Lin, Mingzhe Zheng, JianWei Zhang, Miles Yang, Zhao Zhong, et al. iFSQ: Improving FSQ for image generation with 1 line of code. *arXiv preprint arXiv:2601.17124*, 2026a.

Bin Lin, Zheyuan Liu, Chenguo Lin, Sixiang Chen, Yunyang Ge, Yunlong Lin, Jianwei Zhang, Miles Yang, Zhao Zhong, Liefeng Bo, and Li Yuan. GEAR: Guided end-to-end autoregression for image synthesis. *arXiv preprint arXiv:2606.32039*, 2026b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-X prediction for autoregressive visual generation. *arXiv preprint arXiv:2502.20388*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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

Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. SimpleAR: Pushing the frontier of autoregressive visual generation through pretraining, SFT, and RL. *arXiv preprint arXiv:2504.11455*, 2025.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, Youliang Zhang, Haoqi Fan, and Xihui Liu. Representation forcing for bottleneck-free unified multimodal models. *arXiv preprint arXiv:2605.31604*, 2026.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Xing Xie, Jiawei Liu, Ziyue Lin, Huijie Fan, Zhi Han, Yandong Tang, and Liangqiong Qu. Unleashing the potential of large language models for text-to-image generation through autoregressive representation alignment. *arXiv preprint arXiv:2503.07334*, 2025. URL `https://arxiv.org/abs/2503.07334`.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15703–15712, 2025.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024a.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Related Work

A.1 Representation Alignment in Diffusion Models

REPA Yu et al. (2024b) introduces representation alignment as an auxiliary objective for diffusion transformers. It projects intermediate denoising states through a lightweight head and matches them to patch features from a frozen self-supervised visual encoder such as DINOv2 Oquab et al. (2024), improving both training efficiency and generation quality. Subsequent work extends representation supervision beyond the denoising backbone. VA-VAE Yao et al. (2025) aligns the VAE latent

with a vision foundation model, REPA-E Leng et al. (2025) jointly adapts the VAE and diffusion transformer, and RAE Zheng et al. (2025) constructs a representation autoencoder from a frozen pretrained encoder. Analyses of teacher representations Singh et al. (2025) further identify spatial structure as a factor in generation quality. These studies establish teacher choice, alignment depth, and target granularity as relevant design dimensions. Dense alignment in diffusion models benefits from the shared spatial correspondence and full-image context of denoising states and teacher features. AR models introduce an additional constraint: sequence position determines both the information available to an aligned state and its accessibility to subsequent predictions.

A.2 Representation Supervision in Autoregressive Models

Dense representation alignment has also been applied to AR image generators. iFSQ Lin et al. (2026a) aligns LlamaGen image-token states with visual-teacher features and studies the regularization strength required under teacher forcing. GEAR Lin et al. (2026b) retains intermediate token-wise supervision while extending its gradients into a jointly optimized visual tokenizer. Both methods retain the diffusion-inspired correspondence between image states and spatial teacher targets. ARRA Xie et al. (2025) instead uses a global teacher target, a two-layer adaptor, and cosine alignment on next-token predictor states denoted `[HYBNEXT]`. Its ablations include a fixed `[REP]` token and favor early-layer supervision. Our LlamaGen-L comparisons examine these alignment strategies through the information sets, downstream accessibility, and supervision depth of the aligned states. CARA combines final-block alignment with a dedicated, image-blind prefix state. Representation Forcing Wang et al. (2026) makes visual representations explicit generative intermediates. It autoregressively predicts discrete representation tokens that serve as context for a subsequent image-generation stage. This explicit intermediate representation changes both the prediction target and the generation pipeline. CARA retains visual-token prediction, adding one persistent context position and using the teacher and projection head only during training.

B Additional Experimental Details

B.1 Class-to-Image Generation

A frozen VQ tokenizer with a downsampling factor of 16 and a codebook of 16,384 entries represents each ImageNet-256 image as 256 tokens. LlamaGen-L has 24 Transformer layers, a hidden dimension of 1,024, and 16 attention heads. The reference teacher target is the global `[CLS]` representation of DINOv3-H+. Models at each scale start from the corresponding public LlamaGen checkpoint, with identical backbone initialization across variants within each ablation. Training uses 40 epochs of AdamW with a global batch size of 256, an initial learning rate of 10^{-5}, and cosine decay after 5,000 warm-up steps. The identity-RoPE core controls use the final checkpoints with the GPT-L backbone, evaluated at 256×256 resolution with CFG 1.75 and 50K samples. Experiments preceding the rotary correction are identified separately.

B.2 Text-to-Image Generation

SimpleAR represents each 1024×1024 image using 4,096 Cosmos visual tokens. Each SFT baseline and its CARA counterpart use identical prompts, seeds, sampling settings, resolution, and evaluator checkpoints. GenEval assesses object-level composition, DPG-Bench measures adherence to detailed prompts, and TIIF-Bench TestMini evaluates instruction following for semantically matched short and long prompts. For GenEval and DPG-Bench, sampling uses CFG 6.0, temperature 1.0, top-k 64,000, top-p 1.0, and seed 0, with four images per prompt. Following the original SimpleAR inference protocol, each benchmark prompt is augmented with a style prefix sampled from a fixed template set under a fixed random seed, without external LLM-based prompt rewriting. The ARRA-style baseline applies token-wise global alignment after the first Transformer block through a two-layer GELU adaptor and adds no `[REP]` token. These measurements are reported alongside the existing SFT and CARA reference scores.

C Additional Mechanistic Analyses

For the final-state diagnostics in Table 6, linear CKA measures global feature-space similarity, and mutual 10-nearest-neighbor overlap quantifies local neighborhood agreement between the supervised state and the teacher. A ridge probe fitted on half of the samples measures teacher-feature predictability through cosine similarity on the held-out half. Earlier layerwise diagnostics are reported separately in Appendix D.2.

C.1 Attention Allocation under Identity-RoPE

For teacher-forced image sequences, we compute the attention mass assigned by image-token queries to the condition, representation token, and preceding image tokens. We aggregate attention to r over queries with causal access to its position and also measure total attention to all prefix positions. Table 9 reports prefix attention under identity-RoPE, which preserves the prefix queries and keys. CARA assigns 0.1391 of overall attention mass and 0.2777 of deep-layer attention mass to prefix positions, compared with 0.1435 and 0.2911 for the baseline. The representation token itself receives 2.04% overall and 4.52% in deep layers. Thus, subsequent image states attend to the aligned token without an increase in aggregate prefix attention. Attention entropy and top-1 attention mass also differ little across the evaluated models. The observed improvement is therefore not accompanied by greater total attention to the prefix or a substantial change in these attention-concentration statistics. This pattern is consistent with a contribution from the accessible state's content, although attention weights alone do not identify its causal effect on generation.

Table 9: Prefix attention mass on LlamaGen-L with identity-RoPE. Values are attention fractions aggregated over all layers or the deepest third of the network. CARA does not increase aggregate prefix attention relative to the baseline. Its representation token receives 0.0204 overall and 0.0452 in deep layers.

Model	**Overall prefix mass**	**Deep prefix mass**
Baseline	0.1435	0.2911
CARA	0.1391	0.2777
Insert $k = 128$	0.1383	0.2830
Terminal	0.1385	0.2824

C.2 Causal Accessibility and Future Utility

The identity-RoPE placement and fidelity results support studying causal accessibility jointly with representation learning, while visible information v and downstream accessibility a still vary together. Prospective and intermediate alignment use an added token, whereas terminal alignment uses an existing image-token state, so state identity and prediction role also differ. Independently varying information access and downstream use with a common state construction would distinguish these factors. Interventions on the cached representation-token values demonstrate their contribution to subsequent target-token prediction within the trained CARA model (Section 3.4.3). The near-baseline LlamaGen-L performance of shallow token-wise global alignment also motivates distinguishing teacher-feature predictability from usefulness for future generation. Identifying teacher subspaces relevant to future predictions and learning gated access to the semantic state are potential extensions, neither of which is evaluated here.

D Additional Ablations under the Earlier Rotary Configuration

The following sensitivity, layerwise, and target-granularity results precede the prefix-RoPE correction. They characterize the earlier configuration, with the identity-RoPE causal-placement and depth results reported in Section 3.3.

D.1 Sensitivity to Supervision Choices

The loss-weight, teacher, and normalization-path ablations in Table 10 exhibit a narrow FID range under the earlier rotary configuration. Sensitivity to these choices under identity-RoPE remains unevaluated.

Table 10: Sensitivity to supervision choices before the prefix-RoPE correction. Each variant changes one setting relative to the DINOv3-H+, $\lambda = 1$, shared-path reference.

Variant	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Earlier reference	**2.88**	278.80	0.84	0.54
Bypass path	2.89	275.50	0.84	**0.55**
$\lambda = 0.5$	2.90	**281.09**	0.84	0.54
$\lambda = 2.0$	**2.88**	277.50	0.84	0.54
DINOv3-L	**2.88**	279.91	0.84	**0.55**
SigLIP2	2.94	279.07	0.84	0.54

Shared and bypass normalization produce different parameter updates through their respective gradient paths, yet yield similar generation quality (Table 10). This observation concerns the earlier rotary configuration, whereas the identity-RoPE depth ablation evaluates the shared path.

D.2 Layerwise Representation Diagnostics

The following layerwise trajectories were measured before the prefix-RoPE correction and are distinct from the updated final-state measurements in Table 6. The prefix state's similarity to the teacher decreases from 0.335 at one-third depth to 0.107 in the deepest layers. Intermediate and terminal states follow the opposite trend, increasing from 0.192 to 0.381 and from 0.149 to 0.412, respectively. Mean-pooled image-token representations remain similar across those earlier checkpoints. These measurements describe the earlier configuration and do not establish the layerwise dynamics of identity-RoPE CARA.

D.3 Complete Dense-Alignment Sweep

Table 11 reports dense patch-level and global-terminal alignment before the prefix-RoPE correction. The identity-RoPE experiments in Table 3 retain global targets, so the effect of target granularity under identity-RoPE remains unmeasured.

Table 11: Dense patch-level and global-terminal alignment on LlamaGen-L before the prefix-RoPE correction. All configurations use cosine alignment and the shared feature path.

Granularity	Depth	λ	Teacher	FID ↓	IS ↑	Prec.	Rec.
Patch	Final	1.0	DINOv3-H+	3.33	292.70	0.85	0.52
Patch	Final	1.0	DINOv3-L	3.30	293.20	0.85	0.53
Patch	Final	1.0	SigLIP2	**3.22**	288.22	0.85	0.53
Patch	Final	2.0	DINOv3-H+	3.25	291.11	0.85	0.52
Patch	Two-thirds	1.0	DINOv3-H+	3.29	290.70	0.85	0.52
Patch	One-third	1.0	DINOv3-H+	3.27	286.40	0.85	0.53
Patch	One-third	2.0	DINOv3-H+	3.32	284.04	0.85	0.52
Global	Final	1.0	DINOv3-H+	**3.15**	278.80	0.85	0.53
Global	Two-thirds	1.0	DINOv3-H+	3.22	274.50	0.85	0.53
Global	One-third	1.0	DINOv3-H+	3.28	285.00	0.85	0.52

E UNIFIED ROLE AND DEPTH COMPARISON AND PREDICTIVE DIAGNOSTICS

E.1 ADDITIONAL PREDICTIVE-DISTRIBUTION DIAGNOSTICS

The diagnostics in Section 3.4.2 evaluate the four aligned models and Normal on the same 1,024 ImageNet token sequences under teacher forcing and a common seed. Comparisons between Prefix and non-Prefix models are indexed by the predicted image token x_t, accounting for the additional prefix position. Table 7 reports NLL and its change relative to Normal, together with KL divergence, centered-logit RMSE, logit cosine similarity, and the fraction of predictions whose top-1 token differs from Normal. The latter four quantities describe deviation from a reference model, rather than improvement relative to a ground-truth distribution.

All four aligned models have higher NLL than Normal on these sequences. Increasing alignment depth raises NLL by 0.00149 for Prefix and 0.00897 for `HYBNEXT`, despite improving FID for both roles. The evaluated teacher-forced prediction loss and sampled-image distribution metrics therefore respond differently to alignment. NLL evaluates target-token prediction under observed image prefixes, whereas FID and sFID evaluate the feature distributions of generated images.

Among the Final configurations, CARA combines lower FID and sFID with lower NLL and smaller KL, centered-logit RMSE, and top-1 change relative to Normal. This shows that its generation advantage over `HYBNEXT`/Final is accompanied by a smaller change to the reference model's predictive distribution. Proximity to Normal alone does not explain generation quality, however, since Prefix/Layer-1 is closer to Normal than Prefix/Final on every reported deviation measure but has worse FID. These model-level comparisons characterize the outcome of training across the backbone; they do not isolate the effect of reading the representation token's cache.

E.2 VARIATION ACROSS IMAGE-TOKEN POSITIONS

Table 12 reports KL and top-1 change in the initial, middle, and final position intervals. Both diagnostics are highest in the middle interval among the three reported intervals for all four models. For CARA, KL increases from 0.02140 at positions 1–32 to 0.03252 at positions 129–160, then decreases to 0.02779 at positions 225–256. Its top-1 change follows the same pattern, moving from 28.87% to 31.30% and then 27.05%.

Table 12: Predictive deviation from Normal across image-token position intervals on the same diagnostic sequences. Both KL and top-1 change are largest in the middle interval among the three reported intervals. Top-1 change measures disagreement with Normal, not prediction error.

	KL			Top-1 change		
Model	**1–32**	**129–160**	**225–256**	**1–32**	**129–160**	**225–256**
`HYBNEXT`/Layer-1	0.02000	0.03498	0.03087	28.61%	31.88%	28.21%
`HYBNEXT`/Final	0.02929	0.04518	0.03936	33.83%	36.05%	31.13%
Prefix/Layer-1	0.01657	0.03100	0.02664	27.11%	30.68%	26.61%
Prefix/Final (CARA)	0.02140	0.03252	0.02779	28.87%	31.30%	27.05%

The positional pattern indicates that alignment changes predictive distributions throughout the sequence, with nonuniform effects across positions. Its occurrence in both Prefix and `HYBNEXT` models makes it a shared diagnostic pattern rather than evidence specific to the persistent Prefix state. Moreover, these measurements compare trained models on common teacher-forced prefixes, so they characterize position-dependent predictive differences rather than divergence accumulated along free-generation trajectories.

E.3 ADDITIONAL ARRA-STYLE ALIGNMENT VARIANTS

Table 13 compares three layer-1 ARRA-style variants with deep CARA under the configuration used for the placement experiments. The variants align a fixed `[REP]` token, `[HYBNEXT]` predictor states, or the public implementation's last-256 window. Among the shallow variants, `[HYBNEXT]`

achieves the lowest FID, but none improves on the unaligned baseline. Deep CARA achieves lower FID and sFID than all three.

Table 13: ARRA-style layer-1 alignment and CARA on LlamaGen-L with identity-RoPE. Bold marks the best metric across all rows.

Setting	**FID ↓**	**sFID ↓**	**IS ↑**	**Prec. ↑**	**Rec. ↑**
Unaligned baseline	3.29	6.81	284.35	0.85	0.52
ARRA-style `[REP]`	3.43	6.98	290.14	0.85	0.52
ARRA-style `[HYBNEXT]`	3.31	6.90	287.17	0.85	0.52
ARRA-style public last-256	3.36	6.95	284.22	0.85	0.52
CARA, deep / front	**3.04**	6.80	282.67	0.85	0.54

E.4 Additional Analysis of Cache Interventions

The interventions in Section 3.4.3 use the same CARA checkpoint with an unmodified-cache reference NLL of 7.1553. This reference differs from the unaligned Normal model used in Section 3.4.2. The NLL increments in Table 8 correspond to perplexity ratios of 1.009 for Value-shuffle, 1.233 for Value-zero, and 1.783 for Read-mask.

Value-shuffle produces a modest NLL increase of 0.0092 while changing 18.74% of top-1 predictions. Permuting values preserves their marginal distribution across samples but disrupts their correspondence with the current condition, keys, and queries. The positive loss change supports the predictive utility of this correspondence, beyond the presence of an accessible cache position. It does not isolate a purely semantic component of the values.

Value-zero and Read-mask produce larger NLL increases of 0.2098 and 0.5784, respectively. Zeroing changes the value magnitudes, whereas masking removes an attention position and redistributes weight over the remaining positions. Their effect sizes therefore measure distinct intervention outcomes rather than an additive decomposition of content and routing contributions.