KNOWLEDGE DISTILLATION FOR VISUAL AUTOREGRESSIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive (AR) image generation models are highly expressive but computationally intensive, motivating effective model compression. Knowledge distillation (KD) is a natural approach for model compression and has been widely studied in language modeling, yet its behavior in visual AR generation remains underexplored. In this work, we present the first systematic study of distillation strategies for AR image models. Our analysis shows that while standard distillation can yield meaningful gains, recent methods developed for language do not directly transfer to images: long decoding horizons and visual token ambiguity make teacher supervision unreliable especially under student-conditioned contexts. To address this, we propose VARKD, a distillation framework for visual autoregressive models that distills on student samples while selectively applying teacher supervision and reducing token-level ambiguity. Experiments on text-to-image generation benchmarks and class-to-image generation across multiple autoregressive backbones show that VARKD consistently outperforms prior distillation baselines, substantially narrowing the gap to large-scale models.

1 INTRODUCTION

Autoregressive (AR) models have become a central framework for sequence generation, driving recent advances in large language models and, increasingly, in image generation. Their appeal comes from a simple factorization of the joint distribution into a sequence of conditional predictions, enabling stable maximum likelihood training, strong global coherence, and a natural interface with multimodal and language-conditioned workflows Vaswani et al. (2017); Radford et al. (2019); Van den Oord et al. (2016). In vision, combining this paradigm with discrete tokenization and Transformer decoders has produced scalable AR image generators with competitive fidelity and strong prompt faithfulness Van Den Oord et al. (2017); Razavi et al. (2019); Ramesh et al. (2021); Sun et al. (2024); Tian et al. (2024); Tang et al. (2024); Han et al. (2025).

Despite their strong modeling capabilities, AR image models remain inefficient because image synthesis relies on long-horizon next-token decoding, often requiring hundreds or thousands of sequential steps per image Chen et al. (2020); Lee et al. (2022); Yu et al. (2022). To address this bottleneck, most prior works have focused on accelerating decoding, for example, through speculative, relaxed, or parallel decoding, while leaving the underlying autoregressive model unchanged Jang et al. (2024); Peruzzo et al. (2026); He et al. (2024). Although effective in reducing latency, these approaches implicitly assume access to a large, high-capacity autoregressive model whose robustness is preserved under faster decoding.

Another way to improve efficiency, beyond faster decoding, is through *model-capacity efficiency*: training compact autoregressive image models that retain the robustness and generative quality of their larger counterparts. This is difficult in the visual domain because AR image models operate over much longer horizons than language models Chen et al. (2020); Lee et al. (2022); Yu et al. (2022), rely on highly ambiguous discrete token vocabularies Jang et al. (2024); Wang et al. (2025c), and exhibit strong spatial coupling, where local prediction errors can propagate into global semantic artifacts Ren et al. (2025); Wang et al. (2025a). These properties make visual AR models especially sensitive to the training-inference mismatch induced by teacher forcing: during training, models are optimized on data prefixes, but at inference they must condition on their own predictions Bengio et al. (2015); Xu et al. (2019); Lamb et al. (2016). When model capacity is limited, even small early

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

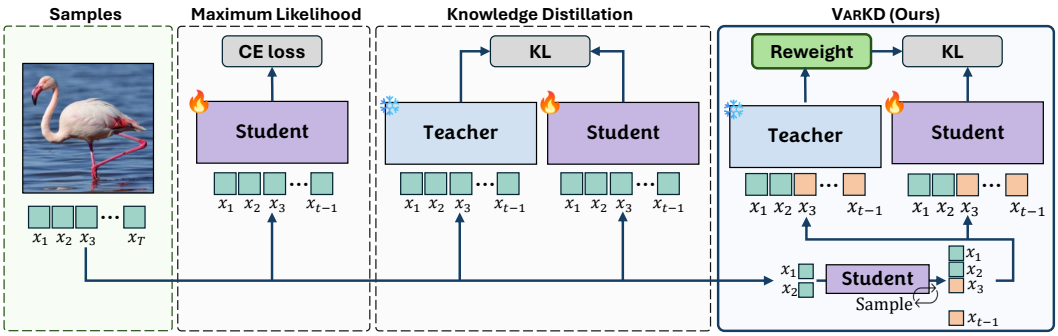


Figure 1: Training paradigms for autoregressive (AR) image models. We assume access to a dataset of real images (left), represented as sequences of discrete tokens. In standard maximum-likelihood training, the model is trained via teacher forcing, optimizing a cross-entropy loss against ground-truth tokens. In knowledge distillation (KD), a pre-trained teacher guides training by matching the student's predictive distribution under ground-truth prefixes. In VARKD, the student generates rollouts from partial ground-truth context, which are scored by the teacher. We further apply confidence-based reweighting to improve distillation reliability.

errors become harder to recover from and can compound over hundreds or thousands of decoding steps, causing compact autoregressive image models to degrade rapidly at inference time Arora et al. (2022).

Knowledge distillation (KD) naturally addresses this capacity gap by transferring the behavior of a large teacher to a compact student Buciluǎ et al. (2006); Hinton et al. (2015). In classification and language modeling, supervised KD Hinton et al. (2015) and sequence-level KD Kim & Rush (2016) typically provide supervision under data-distributed inputs or teacher-generated sequences. Recently, the field has shifted toward *on-policy* distillation Agarwal et al. (2024); Lu & Lab (2025), which trains students on their own generated sequences to align supervision with inference-time rollouts, thereby reducing the training–inference mismatch and improving the performance. We summarize different training paradigms for AR image models in Figure 1, highlighting the different data-source and supervision signals.

However, our empirical investigation suggests that directly transferring these strategies to visual autoregressive generation is non-trivial. In particular, on-policy distillation in the visual domain is hampered by image-specific challenges: long decoding horizons and intrinsic token ambiguity cause student samples to rapidly drift off the data manifold Ren et al. (2025); Jang et al. (2024), producing contexts where teacher predictions become diffuse and weakly informative.

Motivated by these findings, we propose VARKD, a training framework that provides more reliable supervision for compact visual autoregressive models. VARKD distills under student-generated contexts, while improving generation quality at training time by conditioning on a ground-truth prefix. To enhance supervision quality, we apply a reweighting of the loss to down-weight and filter ambiguous, low-confidence teacher predictions, and compute the distillation loss in a compressed visual token space to reduce token-level ambiguity. Finally, we improve training efficiency by employing parallel decoding He et al. (2024). All components are used only during training, preserving standard next-token autoregressive decoding at inference.

Our contributions are summarized as follows:

- We present **the first systematic study of knowledge distillation strategies for visual autoregressive (AR) models**. We evaluate supervised KD, sequence-level KD, and on-policy distillation methods adapted from language modeling, and identify image-specific failure modes: long decoding horizons and intrinsic token ambiguity render teacher supervision unreliable under student-conditioned contexts.
- Motivated by these findings, we introduce **VARKD, a distillation framework tailored to visual AR models**. VARKD improves supervision under mixed student-conditioned rollouts by selectively incorporating teacher feedback based on predictive confidence, and consistently outperforms prior distillation baselines across architectures and model scales.

2 Related Work

2.1 Autoregressive Models for Image Generation

Autoregressive (AR) models are among the earliest approaches to image generation, factorizing the joint image distribution into a sequence of conditional predictions optimized via maximum likelihood. Early pixel-space models such as PixelRNN Van Den Oord et al. (2016) and conditional PixelCNN Van den Oord et al. (2016) demonstrated that high-fidelity images could be generated through next-token prediction with stable training and exact likelihood evaluation, establishing autoregressive modeling as a principled alternative to adversarial methods despite the cost.

The modern resurgence of AR image generation was enabled by discrete latent representations, most notably VQ-VAE Van Den Oord et al. (2017) and VQ-VAE-2 Razavi et al. (2019), which compress images into sequences of visual tokens and substantially reduce generation horizon. This paradigm forms the basis of large-scale text-to-image systems such as DALL·E Ramesh et al. (2021) and subsequent autoregressive models including Parti Yu et al. (2022), VQGAN Esser et al. (2021), HART Tang et al. (2024), LlamaGen Sun et al. (2024) and SIMPLEAR Wang et al. (2025b), achieving competitive image quality, semantic coherence, and prompt alignment at scale.

Within this family, autoregressive image models vary in both tokenization and generation structure. Tokenizers range from vector-quantized codebooks Van Den Oord et al. (2017); Razavi et al. (2019) and adversarially trained VQGAN vocabularies Esser et al. (2021) to AR-oriented holistic tokenizers Zheng et al. (2025). Such discrete visual token spaces can have multiple perceptually similar vocabularies, introducing token ambiguity in visual AR modeling Jang et al. (2024). Generation structures span raster-scan pixel prediction in iGPT Chen et al. (2020), patch-level token generation in LlamaGen Sun et al. (2024), and hierarchical or scale-wise formulations such as HQ-Transformer You et al. (2022), VAR Tian et al. (2024), M-VAR Ren et al. (2024), Switti Voronov et al. (2024), and HART Tang et al. (2024). This paradigm has also been adopted in unified decoder-only multimodal systems such as Emu3 Wang et al. (2024), Chameleon Team (2024), and Lumina-mGPT Liu et al. (2024). Despite this diversity, many token-based AR generators still require long visual rollouts increasing latency and susceptibility to error accumulation Chen et al. (2020); Lee et al. (2022); Yu et al. (2022); Ren et al. (2025).

2.2 Improving Efficiency of Autoregressive Image Models

A fundamental limitation of autoregressive image generation is the inefficiency of next-token decoding at high resolution. When images are represented as sequences of discrete visual tokens, sequence length grows quadratically with spatial resolution, resulting in hundreds or thousands of sequential decoding steps per sample Chen et al. (2020); Yu et al. (2022). This behavior is characteristic of large-scale autoregressive image Transformers such as iGPT Chen et al. (2020), LlamaGen Sun et al. (2024), and more recent visual autoregressive models He et al. (2024), leading to high inference latency and low throughput. Long decoding horizons further amplify error accumulation, as early prediction errors propagate across subsequent tokens Bengio et al. (2015); Ren et al. (2025); Arora et al. (2022).

To reduce inference latency, prior works accelerate autoregressive decoding through speculative, draft-based, or parallel generation schemes. Speculative Decoding Leviathan et al. (2023) introduced draft-and-verify generation, later extended by multi-head or tree-based methods such as Medusa Cai et al. (2024), feature-level speculative frameworks like EAGLE Li et al. (2024a;b), and distillation-based draft alignment as in DistillSpec Zhou et al. (2023). Extensions to the visual domain explicitly account for properties of visual token sequences: LANTERN Jang et al. (2024) accounts for token ambiguity during verification, while MuLo-SD Peruzzo et al. (2026) exploits multi-scale structure and spatial locality. Other methods relax strict raster-order decoding via locality-aware or parallel generation, including ZipAR He et al. (2024), LPD Zhang et al. (2025), and ARPG Li et al. (2026). Although effective, these methods target decoding-time efficiency rather than rollout-aware training of compact visual AR models.

A complementary line of work improves efficiency through formulation-level changes, including next-scale or hierarchical AR designs such as VAR Tian et al. (2024), M-VAR Ren et al. (2024), Switti Voronov et al. (2024), and HART Tang et al. (2024), as well as training-optimization pipelines such as SimpleAR Wang et al. (2025b). These approaches improve the efficiency or scalability of

visual generation, but they do not directly address how to train compact visual AR students to remain stable under their own rollouts.

2.3 Knowledge Distillation for Autoregressive Sequence Models

Knowledge distillation (KD) is a longstanding approach for compressing large neural networks into smaller models, originally based on logit matching and soft-label supervision Buciluǎ et al. (2006); Hinton et al. (2015). In autoregressive sequence modeling, token-level KD typically supervises the student under teacher-forced contexts by matching teacher and student next-token distributions, often treated as a forward Kullback-Leibler objective Zhang et al. (2023); Wei et al. (2024); Gu et al. (2024). Sequence-Level Knowledge Distillation (SeqKD) Kim & Rush (2016) extended this paradigm by training students on full sequences generated by the teacher, thereby replacing the original target distribution with a teacher-induced supervision distribution. While widely adopted in language modeling, these approaches remain off-policy with respect to the student and therefore do not address the training–inference mismatch associated with exposure bias Arora et al. (2022); Ranzato et al. (2015).

Classic imitation-learning work shows that policies trained only on expert states can fail when evaluated on their own induced states, leading to compounding errors Ross & Bagnell (2010); Ross et al. (2011). Recent autoregressive distillation methods build on this view by supervising students under student-induced contexts: Lin et al. (2020) applied imitation-learning-style KD to language generation, while f-DISTILL Wen et al. (2023) generalized sequence-level KD with alternative f-divergences. Generalized Knowledge Distillation (GKD) Agarwal et al. (2024) (also called on-policy distillation Lu & Lab (2025)), further trains students on self-generated rollouts using teacher probabilities as dense supervision, with related LLM distillation work exploring on-policy or imitation-learning-based objectives to reduce exposure bias Gu et al. (2024); Pozzi et al. (2025). However, these methods have been developed and validated primarily on text tasks, leaving their behavior in visual AR generation unexplored.

Unlike text, visual AR models combine long, spatially coupled rollouts with ambiguous token vocabularies, making student-conditioned contexts prone to error accumulation and teacher feedback less reliable under naive on-policy supervision Chen et al. (2020); Lee et al. (2022); Ren et al. (2025). This motivates distillation objectives that adapt on-policy learning to the visual domain while providing reliable teacher supervision.

3 Method

3.1 Distillation for Visual Autoregressive Image Generation

We study knowledge distillation in the setting of autoregressive (AR) image generation. Let $x_{1:T} = (x_1, \ldots, x_T)$ denote a tokenized image with tokens $x_t \in \mathcal{V}$, with $\mathcal{V}$ the discrete token vocabulary. An autoregressive image model defines:

$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_{<t}), \tag{1}$$

where $x_{<t} := x_{1:t-1}$ is the prefix context.

We assume access to a teacher autoregressive model p_T and a lower-capacity student model p_S, both defining next-token distributions over the same vocabulary. The goal is to train the student to match the teacher's predictive behavior while retaining the student's lower inference cost.

In autoregressive image models, **Knowledge Distillation** can be expressed as matching the teacher and student next-token distributions under a chosen distribution of conditioning prefixes:

$$\mathcal{L} = \mathbb{E}_{x_{<t} \sim p_{\mathrm{src}}} \left[\mathcal{D}(p_T(\cdot \mid x_{<t}) \,\|\, p_S(\cdot \mid x_{<t})) \right], \tag{2}$$

where p_{src} denotes the source distribution over prefix contexts, and the expectation implicitly averages over token positions. Here, $\mathcal{D}$ is a chosen divergence between categorical next-token distributions (e. g. KL or Jensen–Shannon divergence). Different distillation strategies can be expressed as different choices of data source p_{src} and divergence $\mathcal{D}$.

Generalized Knowledge Distillation (GKD). GKD Agarwal et al. (2024) applies supervision under contexts generated by the student model itself, often referred to as *on-policy distillation* Lu & Lab (2025). Formally, we set $p_{\text{src}} = p_S$, so prefixes are induced by student rollouts. This exposes the student to its own rollout distribution during training. The corresponding objective is

$$\mathcal{L}_{\text{GKD}} = \mathbb{E}_{x_{<t} \sim p_S} \left[\mathcal{D}(p_T(\cdot \mid x_{<t}) \,\|\, p_S(\cdot \mid x_{<t})) \right]. \tag{3}$$

By training on student-generated samples, on-policy distillation reduces the training-inference mismatch. However, we find that directly applying it to visual AR models yields smaller gains than those observed in language modeling.

The next section analyzes the underlying causes and introduces practical modifications that make on-policy distillation more stable and effective for visual autoregressive generation.

3.2 Reliable teacher feedback under student-conditioned contexts

Mixed data–student context distributions. A direct application of Eq. (3) requires fully student-generated prefixes $x_{<t} \sim p_S$. For visual autoregressive models, such free-running rollouts can drift away from data-like contexts, especially during early training or for capacity-limited students. To regulate the distribution of encountered contexts while retaining exposure to inference-time behavior, we introduce mixed data–student rollouts.

Given a prefix length k, we define a mixed rollout distribution as:

$$p^k_{\text{mix}}(x_{1:T}) = p_{\text{data}}(x_{1:k}) \prod_{t=k+1}^{T} p_S(x_t \mid x_{<t}), \tag{4}$$

where $p_{\text{data}}(x_{1:k})$ denotes the data distribution over length k prefixes. Under this distribution, the first k tokens are anchored to ground-truth visual tokens, while the remaining suffix is generated autoregressively by the student. As a rollout distribution, p^k_{mix} interpolates between fully data-conditioned sequences ($k = T$) and fully student-generated sequences ($k = 0$).

Unless stated otherwise, we apply the distillation objective under p^k_{mix}, which empirically improves supervision quality without altering the inference-time decoding procedure.

Student-sampled, teacher-scored objective. Let $x_{1:T} \sim p^k_{\text{mix}}$ (see Eq. (4)). For suffix positions $t > k$, corresponding to tokens generated by the student under the mixed context distribution, we apply distillation via a reverse-KL divergence:

$$\mathcal{L}_{\text{GKD}} = \mathbb{E}_{x_{1:T} \sim p^k_{\text{mix}},\, t>k} \left[\mathcal{D}(p_S(\cdot \mid x_{<t}) \,\|\, p_T(\cdot \mid x_{<t})) \right]. \tag{5}$$

This corresponds to on-policy distillation under controlled student rollouts, with supervision applied only to student-generated suffix positions. While Eq. (5) mitigates severe distribution shift, teacher predictions on student-generated visual contexts can remain ambiguous, so uniform supervision may still introduce noisy feedback. We address this through selective teacher supervision.

Selective teacher supervision via entropy reweighting. Visual autoregressive models operate over ambiguous discrete token spaces, where teacher uncertainty varies substantially across contexts. We therefore use teacher predictive entropy to estimate the reliability of each distillation target and down-weight uncertain supervision to prevent them from dominating the training signal. For a given context $x_{<t}$, the teacher entropy is:

$$\mathcal{H}(p_T(\cdot \mid x_{<t})) = -\sum_{v \in \mathcal{V}} p_T(v \mid x_{<t}) \log p_T(v \mid x_{<t}). \tag{6}$$

We define teacher confidence as the normalized inverse entropy:

$$\alpha_{x_{<t}} = 1 - \frac{\mathcal{H}(p_T(\cdot \mid x_{<t}))}{\log(|\mathcal{V}|)}, \tag{7}$$

where $\log(|\mathcal{V}|)$ is the maximum entropy, corresponding to a uniform distribution over the vocabulary. By construction, $\alpha_{x_{<t}} \in [0, 1]$: low-entropy (i.e., more confident) teacher predictions receive higher weights, while high-entropy predictions are down-weighted.

While teacher entropy is a useful proxy for model confidence, it can become unreliable once the context drifts off the data manifold: the model may remain confident even when the prefix is incorrect. Inspired by speculative decoding Leviathan et al. (2023), we therefore introduce a truncation strategy that stops applying the distillation loss once confidence drops below a threshold. In practice, given a time step $\bar{t}$ such that $\alpha_{x_{<\bar{t}}} < \tau$, we set $\alpha_{x_{<t}} = 0$ for all $t > \bar{t}$, effectively ignoring the remainder of the sequence. This truncation prevents increasingly unreliable student-conditioned contexts from contributing to the distillation loss.

Using this confidence score, we reweight the distillation objective as:

$$\mathcal{L}_{\text{VarKD}} = \mathbb{E}_{x_{1:T} \sim p_{\text{mix}}^{(k)},\, t>k} \left[\alpha_{x_{<t}} \mathcal{D}(p_S(\cdot \mid x_{<t}) \,\|\, p_T(\cdot \mid x_{<t})) \right]. \tag{8}$$

Codebook relaxation via compressed-space distillation. Discrete visual vocabularies often contain visually similar tokens, so teacher probability mass may be split across multiple tokens that correspond to similar image content. To reduce this token-level ambiguity, we introduce *codebook relaxation*: a compressed-space distillation loss that groups visually similar tokens during supervision.

Let $f : \mathcal{V} \to \tilde{\mathcal{V}}$ map the original vocabulary to a smaller set of token groups obtained by clustering visual token embeddings. For each compressed token $\tilde{x} \in \tilde{\mathcal{V}}$, we aggregate the teacher and student distributions as:

$$\tilde{p}_T(\tilde{x} \mid x_{<t}) = \sum_{x \in f^{-1}(\tilde{x})} p_T(x \mid x_{<t}), \qquad \tilde{p}_S(\tilde{x} \mid x_{<t}) = \sum_{x \in f^{-1}(\tilde{x})} p_S(x \mid x_{<t}). \tag{9}$$

The compressed-space objective is:

$$\mathcal{L}_{\text{VarKD}} = \mathbb{E}_{x_{1:T} \sim p_{\text{mix}}^{(k)},\, t>k} \left[\alpha_{x_{<t}} \mathcal{D}(\tilde{p}_S(\cdot \mid x_{<t}) \,\|\, \tilde{p}_T(\cdot \mid x_{<t})) \right]. \tag{10}$$

This changes only the supervision signal, the student still predicts over the original vocabulary $\mathcal{V}$ and inference uses the same decoding process. By grouping visually similar tokens during supervision, compressed-space distillation reduces sensitivity to token-level ambiguity under long student-conditioned rollouts.

Parallel decoding for efficient training. Constructing student-conditioned rollouts under p_{mix}^{k} requires autoregressive sampling over long suffixes, which can dominate training cost. To mitigate this overhead, we employ parallel decoding He et al. (2024) during rollout generation.

Given a partially decoded visual token grid, parallel decoding identifies a *set* of positions $\mathcal{S}$ that can be predicted simultaneously, and generates all corresponding tokens in a single forward pass:

$$\{x_i : i \in \mathcal{S}\} \leftarrow \text{ParallelDecode}(p_S, x_{\text{obs}}), \tag{11}$$

where x_{obs} denotes the current observed context available to the student.

This procedure is used solely to accelerate rollout construction during training. It does not modify the distillation objective, the model architecture, or the standard autoregressive decoding used at inference time.

4 Experiments

We evaluate distillation strategies for visual autoregressive (AR) image generation on both text-to-image and class-conditional generation. We first study Janus-Pro, a state-of-the-art text-to-image model, and then validate our findings across multiple teacher–student configurations on ImageNet using LlamaGen Sun et al. (2024) and ARPG Li et al. (2026).

Method	Global	Entity	Attribute	Relation	Other	Overall
Janus-Pro-1B	82.97	88.33	86.55	92.64	76.4	82.02
+KD	79.93	88.46	87.06	92.99	**78.8**	81.89
+SeqKD	79.93	88.33	**87.28**	92.22	76.4	81.51
+GKD	81.45	**89.06**	87.13	92.03	76.8	82.00
+VARKD	**83.89**	88.85	86.98	**93.23**	78.4	**82.34**

(a) DPG-Bench Hu et al. (2024)

Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall (↑)
Janus-Pro-1B	98.2	82.1	51.9	88.7	63.1	54.8	73.1
+KD	**99.3**	89.0	52.1	89.8	67.4	**64.1**	77.0
+SeqKD	97.8	87.2	51.9	89.0	68.5	62.0	76.1
+GKD	99.0	88.1	**52.3**	90.8	68.2	64.0	77.1
+VARKD	99.0	**89.1**	52.2	**92.0**	**69.2**	64.0	**77.6**

(b) GenEval Ghosh et al. (2023).

Table 1: Text-to-image generation results across two benchmarks.

4.1 SETTING

Datasets and tasks. For text-to-image generation, we distill Janus-Pro-1B from Janus-Pro-7B Chen et al. (2025b) using LoRA fine-tuning on the dataset from Chen et al. (2025a), and evaluate on GenEval Ghosh et al. (2023) and DPG-Bench Hu et al. (2024). For class-conditional generation, we conduct experiments on ImageNet Russakovsky et al. (2015). Following prior work, we report FID and Inception Score computed from 50k generated samples.

Baselines. We compare against Knowledge Distillation (KD) Hinton et al. (2015), Sequence-level KD (SeqKD) Kim & Rush (2016), and Generalized KD (GKD) Agarwal et al. (2024). KD matches teacher logits under teacher-forced contexts, SeqKD trains on teacher-generated samples, and GKD mixes ground-truth and student-generated contexts. We refer to Sec. A.1 for more details.

4.2 MAIN RESULTS

Quantitative results. We first evaluate on Janus-Pro-1B distilled from a Janus-Pro-7B teacher and report results on GenEval and DPG-Bench (see Table 1b, Table 1a). Across both benchmarks, standard KD already recovers a substantial fraction of the teacher–student gap, highlighting the effectiveness of simple logit matching for visual AR models. Generated-context methods provide additional gains, with VARKD achieving the strongest overall performance. These results demonstrate that our approach transfers effectively to a modern text-to-image foundation model.

We next evaluate class-conditional image generation on ImageNet. Table 2 compares KD, SeqKD, GKD, and VARKD across three student architectures (LlamaGen-B, LlamaGen-L, and ARPG-L). Several consistent trends emerge. First, supervised KD is a surprisingly strong baseline, recovering a large fraction of the teacher–student gap while avoiding the additional sampling cost required by on-policy methods. Second, training on teacher-generated sequences (SeqKD) is consistently less effective than using data-sampled contexts or student-generated rollouts. Third, switching to student-generated contexts (GKD) yields performance comparable to KD, with larger gains for stronger students such as LlamaGen-L and ARPG-L. This suggests that on-policy distillation becomes increasingly beneficial as the student approaches the teacher distribution. Consistent with this hypothesis, VARKD achieves the best (or tied-best) performance across nearly all settings and delivers the most consistent improvements in FID. The gains are observed across student sizes and across both LlamaGen and ARPG backbones, indicating that the proposed refinements are complementary to existing distillation objectives rather than architecture-specific. As expected, the largest improvements are obtained for smaller students, where the teacher–student gap is greatest.

Table 2: Comparison to knowledge distillation methods (KD, SeqKD, GKD) adapted for visual autoregressive models. Results are reported on ImageNet across multiple student sizes and architectures (LlamaGen and ARPG). VARKD consistently matches or surpasses prior baselines.

Method	**FID** (↓)	**IS** (↑)
Llamagen-B	6.53 ± 0.06	156.82 ± 1.48
+ KD Hinton et al. (2015)	4.93 ± 0.05	196.26 ± 2.50
+ SeqKD Kim & Rush (2016)	5.36 ± 0.16	200.64 ± 5.23
+ GKD Agarwal et al. (2024)	4.94 ± 0.05	196.31 ± 2.13
+ VARKD	**4.62** ± 0.04	**206.34** ± 2.17
Llamagen-L	3.16 ± 0.06	257.18 ± 1.32
+ KD Hinton et al. (2015)	2.99 ± 0.02	244.58 ± 2.04
+ SeqKD Kim & Rush (2016)	3.30 ± 0.11	**258.1** ± 2.39
+ GKD Agarwal et al. (2024)	2.92 ± 0.04	248.15 ± 0.94
+ VARKD	**2.84** ± 0.04	245.32 ± 2.49
ARPG-L	2.36 ± 0.04	296.82 ± 5.23
+ KD Hinton et al. (2015)	2.23 ± 0.03	293.51 ± 2.33
+ SeqKD Kim & Rush (2016)	2.24 ± 0.03	**310.78** ± 2.43
+ GKD Agarwal et al. (2024)	2.20 ± 0.02	308.78 ± 2.57
+ VARKD	**2.16** ± 0.03	304.18 ± 3.36

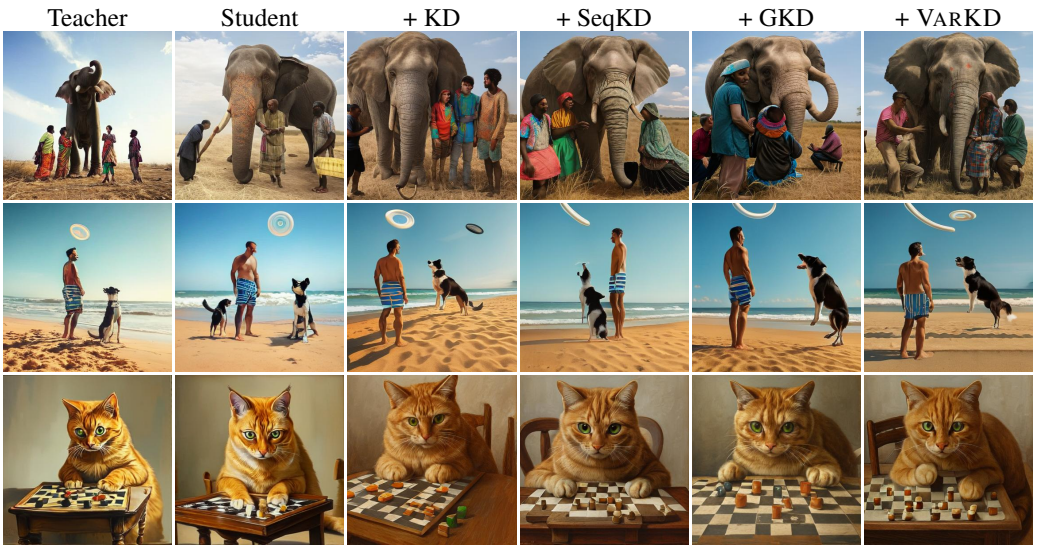


Figure 2: Qualitative comparison showing that VARKD reduces spatial artifacts and improves global coherence over prior distillation baselines. Prompts from the DPG-Bench benchmark.

For each method, we generate samples using five different random seeds and report the mean FID and Inception Score together with its variance. The results show that the improvements achieved by VarKD are consistent across seeds and remain statistically significant. In particular, pairwise t-tests comparing VarKD against the corresponding baselines yield $p < 0.05$ across all evaluated settings, indicating that the observed gains are unlikely to be explained by sampling variability alone.

Qualitative results. Figure 2 shows class-conditional samples across methods. Compared to KD and GKD, VARKD produces samples with fewer spatial artifacts and improved global coherence. Additional qualitative comparisons are provided in the supplementary material Sec B.

4.3 ABLATIONS

Table 3: Method ablation. We analyze the contribution of each component of VARKD.

Entropy Rwt.	Codebook Rlx.	Parallel Decoding	FID (↓)	IS (↑)
–	–	–	6.51	156.3
✗	✗	✗	4.94	194.7
✓	✗	✗	4.66	190.1
✓	✓	✗	4.60	195.7
✓	✓	✓	**4.58**	**203.3**

Table 4: Ablation on teacher size. Effect of teacher capacity on VARKD.

Teacher → Student	FID (↓)	IS (↑)
–	6.51	156.3
L → B	**4.58**	**203.3**
XL → B	4.98	198.0
XXL → B	4.82	198.2

Method We ablate the main design choices of VARKD in the LlamaGen-L → B in Tab. 3. The second row corresponds to GKD, where supervision is applied on mixed-prefix, student-conditioned samples. Starting from this setting, we isolate the three components of VARKD. Adding teacher-confidence reweighting yields the largest improvement, supporting the need to down-weight unreliable teacher feedback under student-conditioned contexts. Adding codebook relaxation further improves performance, consistent with reduced ambiguity among visually similar tokens and a cleaner supervision signal. Finally, parallel decoding does not degrade quality and even yields slight improvements in generative metrics, possibly due to a regularization effect. Moreover, it accelerates training by $1.3\times$, making it more practical in our setting (see supplementary).

Effect of Teacher size We investigate the impact of teacher capacity on student performance in Tab. 4. Overall, we observe a trend consistent with prior findings in language-model distillation Zhong et al. (2024): a larger teacher does not always produce a better student. In fact, we find that distillation can improve when using an intermediate or even weaker teacher, and the strongest teacher is not uniformly optimal across tasks or metrics. We hypothesize that this behavior arises because very large teachers may produce output distributions that are overly sharp or complex for the student to match, especially under limited student capacity or constrained training budgets.

Training efficiency and two-stage distillation. A practical consideration is the additional cost of online distillation. While VARKD incurs lower training throughput than KD due to student rollout generation, this overhead does not directly translate into a proportional increase in end-to-end training cost. Rather than replacing KD, we use VARKD as a refinement stage: the student is first trained with KD for 250k iterations and then further optimized with VARKD for only 20k iterations from the best KD checkpoint.

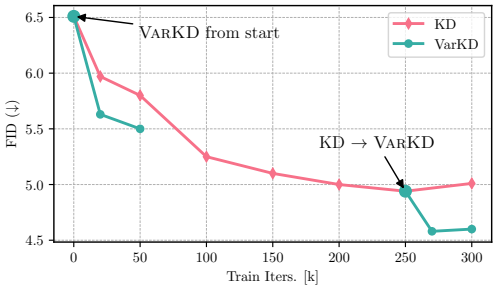


Figure 3: FID vs. training iterations (lower is better) for KD and VARKD. We initialize VARKD from two checkpoints, leveraging KD as an efficient alignment stage.

As shown in Figure 3, KD rapidly converges and provides a strong initialization, but exhibits diminishing returns with additional training. In contrast, continuing optimization with VARKD consistently improves FID beyond the KD plateau. This two-stage recipe achieves the best overall performance while keeping the additional training budget modest relative to the full distillation pipeline.

5 CONCLUSION

In this paper, we study knowledge distillation for visual autoregressive image generation and show that language-model distillation methods do not directly address the long horizons, spatial coupling, and token ambiguity of visual AR models. We present VARKD, a rollout-aware distillation framework that improves teacher supervision under mixed data–student contexts through confidence reweighting and compressed-space distillation. Experiments on ImageNet with LlamaGen and ARPG show that VARKD outperforms standard KD, SeqKD, and GKD without changing inference-time decoding. These results highlight the importance of reliable supervision under student-conditioned rollouts for compressing visual AR models.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, 2022.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020.

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

Patrick Esser, Robin Rombach, and Bjorn 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.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Miniplm: Knowledge distillation for pre-training language models. *arXiv preprint arXiv:2410.17215*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15733–15744, 2025.

Yefei He, Feng Chen, Yuanyu He, Shaoxuan He, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipar: Parallel auto-regressive image generation through spatial locality. *arXiv preprint arXiv:2412.04062*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

Doohyuk Jang, Sihwan Park, June Yong Yang, Yeonsung Jung, Jihun Yun, Souvik Kundu, Sung-Yub Kim, and Eunho Yang. Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. *arXiv preprint arXiv:2410.03355*, 2024.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11523–11532, 2022.

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

Haopeng Li, Jinyue Yang, Guoqi Li, and Huan Wang. Autoregressive image generation with randomized parallel decoding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024a.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-2: Faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7421–7432, 2024b.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020.

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

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Elia Peruzzo, Guillaume Sautière, and Amirhossein Habibian. Multi-scale local speculative decoding for image generation. *arXiv preprint arXiv:2601.05149*, 2026.

Andrea Pozzi, Alessandro Incremona, Daniele Tessera, and Daniele Toti. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Computing and Applications*, 37(18):12013–12029, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Sucheng Ren, Yaodong Yu, Nataniel Ruiz, Feng Wang, Alan Yuille, and Cihang Xie. M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. *arXiv preprint arXiv:2411.10433*, 2024.

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 the IEEE/CVF International Conference on Computer Vision*, pp. 15781–15791, 2025.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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.

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 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, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anton Voronov, Denis Kuznedelev, Mikhail Khoroshikh, Valentin Khrulkov, and Dmitry Baranchuk. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Jiamian Wang, Ziqi Zhou, Chaithanya Kumar Mummadi, Sohail Dianat, Majid Rabbani, Raghuveer Rao, Chen Qiu, and Zhiqiang Tao. Visual self-refinement for autoregressive models. *arXiv preprint arXiv:2510.00993*, 2025a.

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*, 2025b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Tokenbridge: Bridging continuous and discrete tokens for autoregressive visual generation. In *International Conference on Computer Vision (ICCV)(19/10/2025-23/10/2025, Honolulu, Hawai'i)*, 2025c.

Jingxuan Wei, Linzhuang Sun, Yichong Leng, Xu Tan, Bihui Yu, and Ruifeng Guo. Sentence-level or token-level? a comprehensive study on knowledge distillation. *arXiv preprint arXiv:2404.14827*, 2024.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. F-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023.

Yifan Xu, Kening Zhang, Haoyu Dong, Yuezhou Sun, Wenlong Zhao, and Zhuowen Tu. Rethinking exposure bias in language modeling. *arXiv preprint arXiv:1910.11235*, 2019.

Tackgeun You, Saehoon Kim, Chiheon Kim, Doyup Lee, and Bohyung Han. Locally hierarchical auto-regressive modeling for image generation. *Advances in Neural Information Processing Systems*, 35:16360–16372, 2022.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Yufeng Chen, Wenjuan Han, Jian Liu, and Jinan Xu. Towards understanding and improving knowledge distillation for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8062–8079, 2023.

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. *arXiv preprint arXiv:2507.01957*, 2025.

Anlin Zheng, Haochen Wang, Yucheng Zhao, Weipeng Deng, Tiancai Wang, Xiangyu Zhang, and Xiaojuan Qi. Holistic tokenizer for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16916–16926, 2025.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10900–10913, 2024.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. Distillspec: Improving speculative decoding via knowledge distillation. *arXiv preprint arXiv:2310.08461*, 2023.