

---

# Multi-Token Prediction Needs Registers

---

Anastasios Gerontopoulos<sup>1,3</sup>   Spyros Gidaris<sup>2</sup>   Nikos Komodakis<sup>1,3,4</sup>

<sup>1</sup>Archimedes, Athena Research Center

<sup>2</sup>valeo.ai

<sup>3</sup>University of Crete

<sup>4</sup>IACM-Forth

## Abstract

Multi-token prediction has emerged as a promising objective for improving language model pretraining, but its benefits have not consistently generalized to other settings such as fine-tuning. In this paper, we propose MuToR, a simple and effective approach to multi-token prediction that interleaves learnable register tokens into the input sequence, each tasked with predicting future targets. Compared to existing methods, MuToR offers several key advantages: it introduces only a negligible number of additional parameters, requires no architectural changes—ensuring compatibility with off-the-shelf pretrained language models—and remains aligned with the next-token pretraining objective, making it especially well-suited for supervised fine-tuning. Moreover, it naturally supports scalable prediction horizons. We demonstrate the effectiveness and versatility of MuToR across a range of use cases, including supervised fine-tuning, parameter-efficient fine-tuning (PEFT), and pretraining, on challenging generative tasks in both language and vision domains. Our code will be available at: <https://github.com/nasosger/MuToR>.

## 1 Introduction

Autoregressive Transformer architectures have become the cornerstone of modern Large Language Models (LLMs), enabling unprecedented capabilities across a wide range of natural language processing tasks [Achiam et al., 2023, Liu et al., 2024]. Their success has also extended to domains such as image generation [Esser et al., 2021, Sun et al., 2024] and multimodal models [Alayrac et al., 2022, Liu et al., 2023]. These models are primarily trained using a simple yet effective approach: next-token prediction with teacher forcing. By supplying ground truth tokens as context, teacher forcing enables fully parallelized computation via masked self-attention, thus accelerating and stabilizing training.

However, next-token prediction with teacher forcing has notable limitations. Models trained this way often focus on short-term patterns while struggling with harder, long-range decisions. Prior work has shown that this setup can lead to shortcut learning, where valuable supervision diminishes [Bachmann and Nagarajan, 2024], and that it underperforms on tasks requiring planning [Bubeck et al., 2023]. These findings strongly suggest the need for training objectives that go beyond standard next-token prediction.

To address these limitations, multi-token prediction training [Qi et al., 2020, Gloeckle et al., 2024, Liu et al., 2024] has emerged as a promising alternative. Rather than predicting just one token at a time, the model learns to predict multiple future tokens at each position. Recent implementations achieve this through additional transformer output heads: Gloeckle et al. [2024] employs parallel heads, one for each future token position, while Liu et al. [2024] uses sequential heads. Crucially, this approach is used only during training, as its primary goal is to provide a more informative learning signal rather than to speed up inference. Compared to standard teacher forcing, multi-token prediction encourages the model to develop internal "planning" representations, and mitigates overfitting to local patterns.

Experiments by Gloeckle et al. [2024] demonstrate that it leads to improved generative performance and increased data efficiency.

*Motivated by these findings, we explore whether a more effective approach for multi-token prediction can further enhance autoregressive transformers.*

We propose a simple but powerful modification: instead of adding extra transformer layers for future token prediction, we introduce *register tokens*—special tokens interleaved between the regular tokens. Each register token is assigned a randomly sampled offset  $d$ , and the model is trained to predict the token  $d$  steps ahead (rather than just the next token) for these tokens. The original next-token prediction objective remains unchanged for all regular tokens.

These register tokens are used exclusively during training to propagate a richer supervisory signal through the model. At inference time, they are discarded to preserve generation speed. This is made possible by a carefully designed attention mask: register tokens are allowed to attend only to preceding regular tokens—enabling them to learn predictive representations—while regular tokens are entirely blind to register tokens in both directions. This ensures full compatibility with standard autoregressive inference while encouraging the model to internalize forward-looking, multi-step planning during training.

Compared to approaches that add output heads or transformer blocks, our register-based method offers several key advantages:

**No architectural changes:** Only a small number of additional trainable parameters are introduced through register embeddings, while the core transformer layers remain untouched and no extra output heads are required.

**Finetuning compatibility:** Our method is particularly well-suited for fine-tuning pretrained LLMs (e.g., Llama [Grattafiori et al., 2024], Gemma [Gemma Team et al., 2024]). It introduces minimal parameter overhead, preserves original attention patterns for regular tokens, and uses carefully selected position ids and attention masks for register tokens — bringing multi-token prediction closer to the next-token pretraining setup. In contrast, previous methods [Gloeckle et al., 2024, Liu et al., 2024] rely on separate transformer heads, adding many new parameters that must be trained from scratch, making them less effective in fine-tuning scenarios.

**Scalable prediction horizons:** Because the number of register tokens remains fixed regardless of the offset  $d$ , the training cost is independent of the prediction horizon, which can be scaled arbitrarily. Register tokens thus provide greater flexibility in future token prediction. For example, in autoregressive image generation, our method naturally extends to predicting tokens in a two-dimensional neighborhood—a capability not easily achieved by adding output heads.

Overall, our work delivers the following contributions:

- We introduce MuToR (Multi-Token prediction with Registers), a novel multi-token prediction method that employs trainable, interleaved register tokens tasked with predicting multiple future targets. MuToR enables scalable prediction horizons with minimal additional parameters and seamlessly integrates with any pretrained autoregressive language model without architectural modifications.
- Through experiments on language modeling benchmarks, we validate the effectiveness of MuToR in both supervised finetuning and parameter-efficient finetuning (PEFT) settings, consistently surpassing standard finetuning baselines under equivalent training compute.
- We further demonstrate the versatility of MuToR by applying it to autoregressive image generation in a pretraining setting, where it improves performance over standard teacher-forcing—highlighting its broader potential across diverse domains and training settings.

## 2 Related Work

**Limitations of Next-Token Prediction** Bachmann and Nagarajan [2024] introduce a path-finding task on star-graphs to highlight key limitations of standard next-token prediction (i.e., teacher forcing). Their findings reveal that next-token prediction encourages shortcuts, making the underlying task intractable and reducing validation accuracy to random-guessing levels. Interestingly, this "cheating" behavior can be mitigated by multi-token prediction with lookahead embeddings, as in Monea et al. [2023]. While the task is an extreme case, alternative transformer architectures can solve it

[Frydenlund, 2024], reinforcing the intuition that tasks requiring planning might need tailored training objectives [Bubeck et al., 2023].

**Multi-token and Lookahead Prediction** Several works have explored decoding multiple future tokens, primarily to accelerate inference rather than improve generation quality [Stern et al., 2018, Monea et al., 2023, Li et al., 2024, Cai et al., 2024]. In contrast, Qi et al. [2020] propose a multi-token prediction pre-training objective that enhances performance on some sequence-to-sequence tasks. However, their method scales poorly to large decoder models because their multi-stream attention mechanism becomes computationally expensive as prediction depth increases. More recently, Gloeckle et al. [2024] proposed an architecture with multiple parallel decoding heads for multi-token prediction, leading to better generative performance (mostly in coding tasks). Liu et al. [2024] modified this approach to use sequential decoding heads instead. Both studies suggest that multi-token prediction helps by providing richer supervision, better information sharing between tokens, and implicit planning in hidden states. However, these benefits are mainly observed during pretraining, with limited success in fine-tuning scenarios—a gap our work addresses.

A less related line of research trains autoregressive models on permuted sequences [Yang et al., 2019, Pannatier et al., 2024, Kakogeorgiou et al., 2024, Yu et al., 2024, Pang et al., 2024]. By predicting next tokens in shuffled order, these methods force the model to recover distant dependencies in the original sequence. While conceptually interesting, they differ significantly from our approach.

**Trainable Dummy / Register Tokens** Recent work investigates the use of trainable tokens as transformer inputs, revealing emergent properties. Burtsev et al. [2020] prepend dummy tokens to prompts to induce working memory, but observe minimal gains. Goyal et al. [2024] demonstrate that appended trainable tokens can improve performance by increasing operations between tokens, thus encouraging deeper "thinking". Pfau et al. [2024] study the conditions under which a language model can leverage the extra computation provided by the dummy tokens. Related efforts incorporate planning tokens trained to predict latent reasoning steps [Wang et al., 2023, Yin et al., 2024]. In the vision domain, Darcet et al. [2024] employ register tokens during pretraining to prevent the appearance of high-norm artifacts. Similarly to these works, we use learnable tokens—but we also associate them with an auxiliary multi-token prediction objective to create a denser training signal.

More related to our work, Monea et al. [2023] appends lookahead tokens to the input sequence and trains them (while freezing the base model) to enable parallel decoding for speculative sampling. Bachmann and Nagarajan [2024] adapt this idea to path-finding on star-graphs, by simultaneously predicting multiple answer tokens from a prefix. While effective for these specific graph problems, the approach’s complete "teacherless" training paradigm and parallel inference requirements make it fundamentally unsuitable for the broader range of generative tasks we consider. Unlike these methods, our approach only uses register tokens during training to enhance supervision, without modifying the inference procedure or requiring additional compute at test time.

## 3 Method

### 3.1 Preliminaries

**Next-Token Prediction** Building on the foundational work of Shannon [1948, 1951], next-token prediction remains the core objective for autoregressive language models. Given a sequence  $(x_1, \dots, x_t)$ , the model is trained to predict the next token  $x_{t+1}$  by maximizing the joint probability under left-to-right factorization:

$$P(x_1, \dots, x_T) = \prod_t P(x_{t+1} \mid x_{\leq t}), \quad (1)$$

where  $T$  is the sequence length. For a model  $P_\theta$  and dataset  $\mathcal{D}$ , the training objective will be to minimize the the expected negative log-likelihood loss over the dataset:

$$\mathcal{L}_{\text{ntp}} = \mathbb{E}_{\mathcal{D}} \left[ - \sum_t \log P_\theta(x_{t+1} \mid x_{\leq t}) \right]. \quad (2)$$

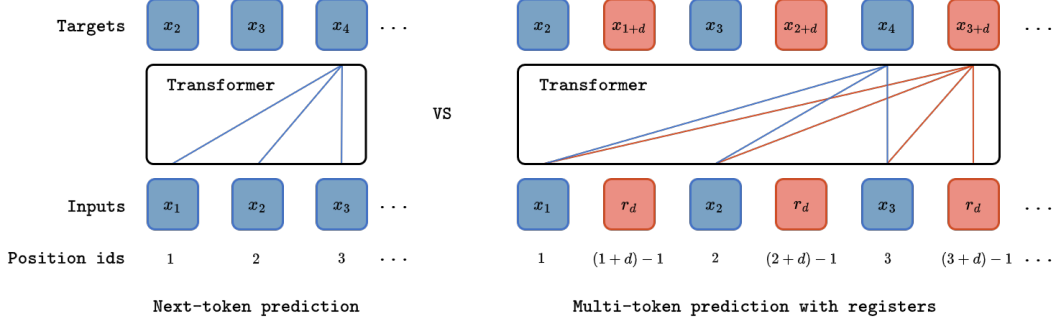


Figure 1: **Next-token prediction vs. Multi-token prediction with registers (MuToR).** The transformer block represents any decoder-only autoregressive model, with colored lines indicating permitted attention connections between tokens. **Left:** Standard next-token prediction, where each  $x_t$  predicts  $x_{t+1}$  conditioned on preceding tokens. **Right:** MuToR interleaves register tokens  $r_d$  to predict tokens  $d$  steps ahead ( $x_{t+d}$ ), conditioned only on previous *regular* tokens. Register tokens are assigned position ids (e.g.,  $t + d - 1$  for  $r_d$  targeting  $x_{t+d}$ ) that mimic next-token prediction. Regular tokens follow the standard next-token prediction formulation, unaffected by the registers.

**Multi-Token Prediction** In contrast, Gloeckle et al. [2024] propose predicting multiple future tokens per position. Their objective minimizes:

$$\mathcal{L}_{\text{mtp}} = \mathbb{E}_{\mathcal{D}} \left[ - \sum_t \log P_{\theta}(x_{t+1:t+d_{\max}} \mid x_{\leq t}) \right], \quad (3)$$

where we denote the maximum prediction horizon as  $d_{\max}$ , to align notation with our method. Recent implementations utilize additional transformer heads—either parallel [Gloeckle et al., 2024] or sequential [Liu et al., 2024].

### 3.2 Our approach

In this work, we introduce MuToR, an alternative multi-token prediction method (illustrated in Figure 1). Our approach inserts interleaved learnable tokens—termed *registers*, following Darcet et al. [2024]—into training sequences, where each register is tasked with predicting a future token at a uniformly sampled offset  $d$ . By optimizing this auxiliary prediction objective alongside the primary next-token prediction task, the model benefits from richer supervisory signals, which enhances the quality of learned representations.

**Register Tokens** Let  $x = (x_1, x_2, \dots, x_T)$  be a training sequence. We augment  $x$  by interleaving<sup>1</sup> register tokens  $r_d$ , yielding:

$$x' = (x_1, r_d, x_2, r_d, \dots, x_{T-1}, r_d, x_T), \quad (4)$$

where each  $r_d$  predicts the future token at offset  $d^2$ . By default, all  $r_d$  share a single learnable embedding, adding minimal trainable parameters, while the target offset  $d$  is specified via the register’s position id (detailed later). The augmented sequence  $x'$  is processed by the transformer  $P_{\theta}$ , which shares all components—embeddings, layers, and prediction head—between the regular tokens  $x_t$  and register tokens  $r_d$ .

**Attention Masking** We modify the causal attention mask to satisfy two conditions: (i) regular tokens  $x_t$  cannot attend to any register tokens, and (ii) register tokens cannot attend to other register tokens in the sequence (see Figure 2). This pattern preserves the standard next-token prediction objective (Equation 2) for regular tokens, as their representations remain unaffected by registers. As a result, the registers can be discarded during inference.

<sup>1</sup>The interleaving pattern is flexible—e.g., registers may be sparse or appear consecutively.

<sup>2</sup>For simplicity, we use a fixed  $d$  per sequence, though registers with different offsets could be mixed.

**Multi-token Prediction with Registers** Each register token  $r_d$  inserted after  $x_t$  predicts the future token  $x_{t+d}$ , with offset  $d$  sampled uniformly per sequence from  $\{1, \dots, d_{\max}\}$ , where  $d_{\max}$  determines the maximum prediction horizon. The auxiliary register loss over dataset  $\mathcal{D}$  is:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathcal{D}} \left[ - \sum_t \log P_{\theta}(x_{t+d} \mid x_{\leq t}, r_d) \right]. \quad (5)$$

Our attention masking ensures predictions depend only on preceding regular tokens, excluding other registers. This approach enables flexible prediction horizons while maintaining parameter efficiency, as all register tokens—regardless of  $d$ —share a single learnable embedding.

**Position Embeddings for Registers** Since we do not use specialized heads for future token prediction, we encode prediction offsets through positional bias. We utilize the token indices in the original (register-free) sequence to compute positional embeddings. While regular tokens  $x_t$  keep their natural positions  $t$ , each register  $r_d$  inserted after  $x_t$  (predicting  $x_{t+d}$ ) receives position  $t + d - 1$  (see position ids in Figure 1). This matches the position id of the regular token that would normally predict  $x_{t+d}$  under standard next-token prediction.

Our implementation focuses on RoPE [Su et al., 2024], the dominant positional encoding in modern language [Touvron et al., 2023b] and autoregressive image models [Sun et al., 2024]. However, our position manipulation works with any embedding scheme (sinusoidal, relative, etc.), requiring no architectural changes while effectively encoding prediction offsets through positional bias.

**Overall Training Loss** We jointly optimize the standard next-token prediction loss  $\mathcal{L}_{\text{ntp}}$  and the auxiliary register loss  $\mathcal{L}_{\text{reg}}$ . The overall loss combines these two objectives through a weighted sum:

$$\mathcal{L}_{\text{mtp}} = (1 - a)\mathcal{L}_{\text{ntp}} + a\mathcal{L}_{\text{reg}}, \quad (6)$$

where  $a \in (0, 1)$  controls the relative contribution of each loss term.

**Inference** During inference, we discard register tokens entirely, leaving the model’s computational graph and latency unchanged. This is made possible by our attention masking strategy, which prevents regular tokens from ever attending to registers during training. Unlike prior approaches (Goyal et al. [2024], Pfau et al. [2024]) that use inserted tokens to increase inference computation, our method maintains the standard autoregressive process without modification.

### 3.3 Adaptation in Language Modeling

A key application of our method is supervised fine-tuning of pre-trained language models. For generative tasks (e.g., mathematical reasoning), where datasets contain (prefix, answer) pairs—with answer sequences potentially including chain-of-thought tokens—we interleave register tokens only within the answer sequence. This aligns with standard practice where prefix predictions are excluded from loss computation. Beyond this adaptation, the method follows the implementation described in subsection 3.2 without modification.

### 3.4 Adaptation in Autoregressive Image Generation

Autoregressive transformers achieve strong performance on image generation by modeling discrete visual tokens [Esser et al., 2021, Sun et al., 2024]. A VQ-VAE tokenizer [Van Den Oord et al., 2017] first encodes an image into a 2D token grid  $x^{2D} \in \mathbb{Z}^{h \times w}$ , which is then flattened into a 1D sequence  $x$  via raster-scan ordering. The model then learns autoregressive prediction conditioned on  $c$  (either class labels or captions).

Figure 2: MuToR’s attention mask. Each cell indicates whether the row can attend to the column.

Figure 3: The 2D neighborhood of possible prediction targets (depicted in red) for a register token. The register  $r_d$  is inserted after  $x_7$ , and  $d_{\max\_2D}$  is set to 3.

We adapt MuToR to images by modifying the offset sampling to respect the 2D image structure. For each sequence  $x$ , we sample a 2D offset pair  $(d_h, d_w)$ , with both  $d_h$  and  $d_w$  drawn uniformly from  $\{1, \dots, d_{\max\_2D}\}$ , excluding  $(d_h, d_w) = (1, 1)$ , as it denotes the image token after which the register is inserted (see Figure 3). We then compute the rasterized offset as  $d = (d_h - 1) \cdot w + d_w - 1$ . As in subsection 3.2, each register token  $r_d$  predicts the token  $d$  steps ahead in the sequence, with all other components (attention masking, positional embeddings, and loss) implemented identically—except for the additional conditioning on  $c$ . Each register thus predicts one of  $d_{\max\_2D}^2 - 1$  possible future tokens.

This 2D extension enriches the training signal by capturing spatial dependencies inherent in visual data, while requiring minimal architectural changes. Unlike prior multi-token prediction approaches that would require multiple additional prediction heads (one for each possible 2D offset), MuToR achieves this capability with negligible parameter overhead.

## 4 Results

### 4.1 Language Modeling

**Experimental Setup** We focus on mathematical reasoning with chain-of-thought and abstractive summarization, two challenging generative tasks that provide a rigorous testbed for our method. To evaluate performance, we fine-tune two pretrained decoder-only language models: Gemma 2B [Gemma Team et al., 2024] and Llama 3 8B [Grattafiori et al., 2024]. Our experiments target three widely used mathematical reasoning benchmarks: GSM8K [Cobbe et al., 2021], MATH500 [Lightman et al., 2023], and AQUA-RAT [Ling et al., 2017]. For finetuning, we use curated subsets from OpenMathInstruct-2 [Toshniwal et al., 2025], a high-quality dataset derived from GSM8K and MATH training samples. Specifically, we filter the 1M and 2M splits to isolate grade school (GSM-style) or MATH-style problems, referring to them as *1M-GSM*, *2M-GSM*, and *1M-MATH*. We also finetune on the original GSM8K and AQUA-RAT training splits. Additional details about the experimental setup are provided in Appendix A.1. As for summarization, we target the following benchmarks: SAMSum [Gliwa et al., 2019] and DialogSum [Chen et al., 2021]. For mathematical tasks, we measure exact-match accuracy; for summarization, we use ROUGE scores [Lin, 2004].

**Baselines** We consider two baselines: (i) *Next-Token*, the standard finetuning recipe using the next-token prediction objective, and (ii) *Multi-Token* [Gloeckle et al., 2024], adapted for finetuning by adding  $d_{\max} - 1$  auxiliary prediction heads and applying a loss-weighting strategy for these heads, similar to our method. To ensure a fair comparison, for both *Multi-Token* and MuToR, we tune the number of predicted future tokens,  $d_{\max}$ , alongside the auxiliary loss coefficient. Implementation details are provided in Appendix A.1.

**Comparative Results** Tables 1 and 2 present our results, showing only the best configurations for *Multi-Token* and our MuToR method. In mathematical reasoning (Table 1), our approach consistently outperforms both baselines, including *Multi-Token*, which introduces a substantial number of additional trainable parameters (see Table 3). Moreover, MuToR’s gains are preserved across varying training set sizes, demonstrating its effectiveness even in settings with high-quality finetuning data. In contrast, *Multi-Token*’s benefits seem to diminish with larger models or more data. For summarization (Table 2), our MuToR method improves all ROUGE scores, achieving superior results over *Multi-Token*, demonstrating broad applicability in sequence-to-sequence generative tasks.

**Matching the Training-Time Compute** Our method increases sequence length during training (and thus training compute) by inserting register tokens. To ensure gains are not due to higher compute alone, we train the baselines for more epochs (roughly doubling compute), thus matching or even exceeding the compute of MuToR. As shown in Tables 11 and 12 (Appendix B.1), extending training does not yield further improvements for the baselines and, in some instances, can even degrade performance due to mild overfitting. These results suggest that integrating MuToR into the finetuning pipeline offers a more effective approach to leveraging increased training compute.

**Integration in Parameter-Efficient Finetuning** We test our method with LoRA [Hu et al., 2022]: for both *Next-Token* and our MuToR method, we apply rank-32 adapters to all linear layers (approx-

Table 1: Downstream accuracy (%) in mathematical reasoning benchmarks. The subheaders refer to the training split used in each experiment. Results for Gemma 2B are averaged over 3 seeded runs. MuToR offers a consistent improvement over both standard Next-Token and Multi-Token.

Model	Method	GSM8K			MATH500	AQUA-RAT
		GSM8K	1M-GSM	2M-GSM	1M-MATH	AQUA-RAT
Gemma 2B	Next-Token	38.87	66.09	69.02	26.73	40.16
	Multi-Token	40.66	66.69	69.02	26.87	38.45
	MuToR (ours)	<b>42.10</b>	<b>68.33</b>	<b>70.56</b>	<b>28.13</b>	<b>41.73</b>
Llama3 8B	Next-Token	66.41	85.74	87.33	41.4	-
	Multi-Token	66.56	85.67	86.35	42.6	-
	MuToR (ours)	<b>67.85</b>	<b>87.03</b>	<b>87.64</b>	<b>43.2</b>	-

Table 2: Experimental results for finetuning Gemma 2B on abstractive summarization benchmarks. We select the checkpoint with the higher ROUGE-L in the validation set, and report ROUGE scores on the test set.

Method	SAMSum			DialogSum		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Next-Token	51.47	27.29	43.23	47.23	20.91	38.77
Multi-Token	51.90	27.44	43.50	47.98	21.23	39.25
MuToR (ours)	<b>52.32</b>	<b>28.08</b>	<b>44.09</b>	<b>48.22</b>	<b>21.71</b>	<b>39.48</b>

mately 39M trainable parameters for Gemma 2B; register tokens in MuToR add a negligible number of parameters). As shown in Table 4, our LoRA-MuToR approach improves accuracy over standard LoRA finetuning across both training splits. Interestingly, LoRA-MuToR *matches or even exceeds the full-finetuning* Next-Token performance, demonstrating its utility in PEFT setups. In contrast, the Multi-Token approach is less compatible with PEFT settings, as it requires training several additional transformer layers from scratch.

**Impact of Maximum Lookahead Offset** The key hyperparameter  $d_{\max}$  in our MuToR method controls how many tokens ahead the registers predict. Larger values offer richer supervision but increase task difficulty. Experiments (Table 3) show that  $d_{\max} = 4$  is optimal for this particular setting (in general the optimal value may depend on the training data and the downstream task). Notably, as  $d_{\max}$  increases, Multi-Token’s performance degrades, barely beating Next-Token when using the *1M-GSM* training split. In comparison, MuToR’s gains are maintained across  $d_{\max}$  values and training split sizes.

**Shared vs. Different Register Embeddings Per Offset** We test whether having different register embeddings per offset improves performance. As seen in Table 5, shared embeddings (*Same*), which is the default, perform slightly better than distinct ones (*Different*), suggesting our positional encoding scheme provides sufficient offset information.

## 4.2 Autoregressive Image Generation

**Experimental Setup** We train LlamaGen-B (111M parameters; Sun et al. 2024) on ImageNet [Deng et al., 2009] at 256×256 resolution, using ADM’s preprocessing pipeline [Dhariwal and Nichol, 2021]. The dataset is pre-tokenized using a VQ-VAE tokenizer from Sun et al. [2024].

We compare three approaches: (1) Next-Token, the standard next-token prediction baseline; (2) MuToR-1D with 1D offsets (as in language modeling); and (3) MuToR-2D with 2D offsets (described in subsection 3.4). Implementation details are included in Appendix A.2.

For evaluation, we generate 50,000 images using classifier-free guidance (scale=2.0) [Ho and Salimans, 2022] and compute FID [Heusel et al., 2017], IS [Salimans et al., 2016], Precision, and Recall [Kynkäänniemi et al., 2019] using TensorFlow scripts from Dhariwal and Nichol [2021].

Table 3: Downstream accuracy (%) with respect to the maximum offset  $d_{\max}$ , using Gemma 2B. #Add. Param. denotes the additional trainable parameters for Multi-Token and MuToR (in approximation).

$d_{\max}$	#Add. Param.	Method	GSM8K	
			GSM8K	1M-GSM
1	-	Next-Token	38.87	66.09
2	110M 2K	Multi-Token	40.66	66.69
		MuToR (ours)	41.93	67.15
3	220M 2K	Multi-Token	40.59	66.36
		MuToR (ours)	41.60	68.01
4	330M 2K	Multi-Token	39.78	65.53
		MuToR (ours)	<b>42.10</b>	<b>68.33</b>
6	550M 2K	Multi-Token	40.23	65.55
		MuToR (ours)	39.90	68.16

Table 4: Downstream accuracy (%) in a PEFT scenario, using Gemma 2B and LoRA.

Method	GSM8K	
	GSM8K	1M-GSM
Full-finetuning Next-Token	38.87	66.09
LoRA-Next-Token	36.34	66.11
LoRA-MuToR (ours)	<b>38.59</b>	<b>68.11</b>

Table 5: Ablation regarding the register embeddings, using Gemma 2B and  $d_{\max} = 4$ .

Register embedding	GSM8K	
	GSM8K	1M-GSM
Same	<b>42.10</b>	<b>68.33</b>
Different	41.85	68.18

**Results** Table 6 shows that both MuToR variants consistently outperform Next-Token in FID and IS across different training iterations. Notably, when comparing under similar training-time compute, the MuToR-2D variant at 100K steps surpasses the Next-Token model at 200K steps, demonstrating both improved performance and faster convergence.

**Extending the offset to 2D** The 2D extension in MuToR-2D significantly improves performance by leveraging spatial dependencies in the image data (Table 6), despite requiring prediction of much more possible future tokens. This demonstrates that the 2D formulation effectively captures structural patterns, providing richer training signal.

**Scaling down the number of registers during pretraining** To reduce computational costs while maintaining performance, we investigate using fewer register tokens in MuToR-2D. Specifically, we test inserting only 80 randomly placed registers per image, increasing the sequence length by just 30%. As shown in Table 7, this sparse version achieves very similar performance to the full setup (with 256 registers) while requiring less computation. These results demonstrate that substantial performance gains can be obtained with relatively few register tokens and only a modest increase in training compute. They also highlight untapped potential in MuToR’s design, particularly regarding optimal register placement and sparsity strategies, which merit further investigation.

**Maximum Offset Analysis** We examine how expanding the prediction neighborhood in MuToR-2D affects performance by varying  $d_{\max\_2D}$ , which determines how many future tokens each register must predict. Using 80 randomly placed registers per sequence, we test different  $d_{\max\_2D}$  values. Table 8 shows that  $d_{\max\_2D} = 4$  achieves optimal performance, while increasing it leads to degrading results. In this optimal setup, each register predicts up to 15 future tokens—a prediction horizon

Table 6: Conditional generation performance on Imagenet  $256 \times 256$  (cfg scale = 2.0). Both  $d_{\max}$  and  $d_{\max\_2D}$  are set to 4, for MuToR-1D and MuToR-2D respectively.

# Iter.	Method	FID ↓	IS ↑	Pre. ↑	Rec. ↑
100K	Next-Token	7.71	146.5	0.830	0.439
	MuToR-1D	7.01	155.6	0.828	0.438
	MuToR-2D	<b>6.57</b>	<b>163.2</b>	<b>0.831</b>	<b>0.445</b>
200K	Next-Token	6.83	158.4	0.833	0.443
	MuToR-1D	6.43	163.0	0.836	0.444
	MuToR-2D	<b>5.65</b>	<b>183.5</b>	<b>0.842</b>	<b>0.448</b>
360K	Next-Token	6.18	171.5	<b>0.841</b>	0.443
	MuToR-1D	5.79	178.3	<b>0.841</b>	0.441
	MuToR-2D	<b>5.09</b>	<b>195.3</b>	0.839	<b>0.457</b>



Table 7: Ablation using MuToR-2D,  $d_{\max\_2D} = 4$ , and varying number of registers (# Reg.) inserted in random positions.

# Iter.	# Reg.	FID ↓	IS ↑
100K	80	6.87	162.1
	256	<b>6.57</b>	<b>163.2</b>
200K	80	5.89	175.6
	256	<b>5.65</b>	<b>183.5</b>
360K	80	<b>5.03</b>	<b>195.6</b>
	256	5.09	195.3

Table 8: Ablation using MuToR-2D, with varying  $d_{\max\_2D}$  and 80 registers inserted in random positions. # Targets denote the number of future tokens ( $d_{\max\_2D}^2 - 1$ ) that the registers predict during training (see Figure 3).

# Iter.	$d_{\max\_2D}$	# Targets	FID ↓	IS ↑
100K	3	8	7.20	150.8
	4	15	<b>6.87</b>	<b>162.1</b>
	5	24	7.08	154.8
200K	3	8	6.19	169.0
	4	15	<b>5.89</b>	<b>175.6</b>
	5	24	6.21	166.2
360K	3	8	5.56	178.3
	4	15	<b>5.03</b>	<b>195.6</b>
	5	24	5.53	179.2

that would require 14 additional transformer heads in prior multi-token approaches, making them computationally impractical. This demonstrates MuToR-2D’s unique ability to effectively leverage long-range predictions while maintaining training efficiency.

### 4.3 Synthetic Data

We further evaluate MuToR on the star-graph path-finding problem [Bachmann and Nagarajan, 2024], which highlights limitations of the next-token prediction objective. Consider a directed graph  $G = (n, l)$ , where  $n$  denotes the number of paths emanating from the start node,  $u_{\text{start}}$ , and  $l$  denotes their length. Given an end node,  $u_{\text{end}}$ , the model must identify the unique path from  $u_{\text{start}}$  to  $u_{\text{end}}$ .

Despite the task’s simplicity, transformer models trained by teacher forcing fail to solve it, due to shortcut learning and the loss of meaningful training signal. We thus investigate whether the look-ahead induced by MuToR’s objective can provide the necessary supervision to learn the underlying task. To this end, we experiment with finetuning a pretrained GPT2 model [Radford et al., 2019] following the setup from Bachmann and Nagarajan [2024]. The relevant implementation details are provided in the Appendix A.3.

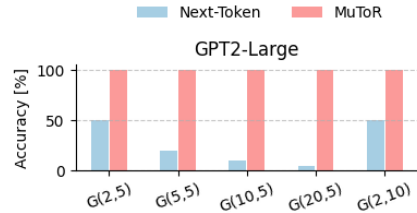


Figure 4: Solve rate (%) of finetuned GPT2-L model on different star graph configurations.

**Results** The results are presented in Figure 4. MuToR solves the task across various graph configurations, effectively overcoming the "cheating phenomenon" that causes standard teacher forcing to fail. These findings indicate that MuToR can be particularly effective in scenarios where some form of shortcut learning applies, recovering valuable training signal.

## 5 Conclusion

We introduced MuToR, a simple yet effective approach to multi-token prediction that leverages interleaved, trainable register tokens to predict future targets. MuToR enables scalable prediction horizons with minimal parameter overhead and is fully compatible with existing pretrained model architectures, allowing seamless integration into standard finetuning pipelines. Empirical results demonstrate that MuToR consistently improves performance in both language modeling tasks—such as mathematical reasoning and summarization—and autoregressive image generation, highlighting its versatility across modalities. This positions MuToR as a promising foundation for using token-based lookahead mechanisms to propagate richer supervisory signals during training.

It is worth noting that MuToR currently uses uniformly interleaved or randomly positioned register tokens—strategies that may not align optimally with the structure or semantics of specific tasks. While this simple placement scheme has proven effective across modalities, it leaves room for

substantial improvement. By learning or adapting the placement of register tokens—potentially guided by model uncertainty or task-specific priors—MuToR could deliver more targeted supervision with fewer auxiliary tokens, further enhancing efficiency and performance.

## Acknowledgments and Disclosure of Funding

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Hardware resources were granted with the support of GRNET.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.
- Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2296–2318, 2024.
- Sébastien Bubeck, Varun Chadrsekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.
- 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. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5209–5235, 2024.
- Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, 2021.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019.
- 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*, pages 12873–12883, 2021.
- Arvid Frydenlund. The mystery of the pathological path-star task for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12493–12516, 2024.

- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsun corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, 2019.
- Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15706–15734, 2024.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*, 2024.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Ioannis Kakogeorgiou, Spyros Gidaris, Konstantinos Karantzas, and Nikos Komodakis. Spot: Self-training with patch-order permutation for object-centric learning with autoregressive transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22776–22786, 2024.
- Dawid J Kopiczko, Tijmen Blankevoort, and Yuki M Asano. Bitune: Bidirectional instruction-tuning. *arXiv preprint arXiv:2405.14862*, 2024.
- Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28935–28948, 2024.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, 2017.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Giovanni Monea, Armand Joulin, and Edouard Grave. Pass: Parallel speculative sampling. *arXiv preprint arXiv:2311.13581*, 2023.
- Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024.
- Arnaud Pannatier, Evann Courdier, and François Fleuret.  $\sigma$ -gpts: A new approach to autoregressive models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 143–159. Springer, 2024.
- Jacob Pfau, William Merrill, and Samuel R Bowman. Let’s think dot by dot: Hidden computation in transformer language models. In *First Conference on Language Modeling*, 2024.
- Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, 2020.
- 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1): 50–64, 1951.
- Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- 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.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.
- Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordani. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*, 2023.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- Yongjing Yin, Junran Ding, Kai Song, and Yue Zhang. Semformer: Transformer language models with semantic planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18669–18680, 2024.
- Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. *arXiv preprint arXiv:2411.00776*, 2024.

## A Implementation Details

In this section, we provide all the training details, as well as the hyperparameters used in our experiments. In order to facilitate reproducibility, we plan to release the code for MuToR’s implementation as well.

### A.1 Language Modeling

#### A.1.1 Mathematical Reasoning

**Datasets** GSM8K comprises approximately 8.7K grade-school level math problems, with around 1.3K of them forming the test set. On the other hand, MATH500 test set consists of 500 problems, uniformly sampled from the original MATH test set [Hendrycks et al., 2021], covering more advanced and diverse mathematical domains. Finally, AQUA-RAT includes multiple-choice math problems with chain-of-thought solutions, split among the training set ( $\sim 97$ K samples), the validation ( $\sim 250$  samples) and the test set ( $\sim 250$  samples).

As mentioned in subsection 4.1, we finetune our base models on standard downstream training datasets, such as GSM8K and AQUA-RAT, as well as on curated subsets derived from OpenMathInstruct-2. The utilized training splits are listed below, including information about the employed filtering and the amount of samples:

- GSM8K training set ( $\sim 7.4$ K samples),
- 1M-GSM split ( $\sim 152$ K samples), obtained by filtering OpenMathInstruct-2’s 1M split for grade school math-like problems,
- 2M-GSM split ( $\sim 277$ K samples), similarly derived from OpenMathInstruct-2’s 2M split using the same grade-school filtering criteria,
- 1M-MATH split ( $\sim 200$ K samples), constructed by filtering OpenMathInstruct-2’s 1M split for MATH-style problems and randomly sampling 200K of them,
- AQUA-RAT training set ( $\sim 97$ K samples).

Our filtering depends on the source dataset of each sample, which is included in OpenMathInstruct-2’s metadata.

**Training details** Throughout all experiments and for all methods, we use bidirectional attention among the prefix tokens, as proposed in previous works [Dong et al., 2019, Raffel et al., 2020, Kopiczko et al., 2024], since it benefits prefix-answer tasks. We finetune all models for 5 epochs, using AdamW optimizer [Loshchilov and Hutter, 2017] without weight decay and a batch size of 10. We also employ a learning rate scheduler with linear decay and warmup, setting the peak learning rate to be  $5e-5$  for Gemma 2B and  $2e-5$  for Llama 3 8B. Both language models are loaded and finetuned in bfloat16 precision. To match our available resources, we filter out training sequences that are longer than 512 tokens<sup>3</sup>. Moreover, all experiments with the 2B model are conducted using three random seeds to ensure statistical robustness. This setting is not replicated for the 8B model though, due to the substantial computational resources that are required.

**Evaluation** For evaluation, we assess performance by using greedy decoding and applying exact match techniques to verify the correctness of the generated answer. For experiments on GSM8K and MATH500, the best-performing checkpoint from each run is then used for comparisons. For experiments on AQUA-RAT, we choose the checkpoint with the highest accuracy on the provided validation set. Since register tokens are not used during inference, the inference process is identical across both our method and the baseline approaches.

#### A.1.2 Abstractive summarization

**Datasets** In our experiments, we target SAMSum and DialogSum, two widely used dialogue summarization benchmarks. SAMSum consists of approximately 16K messenger-like conversations with summaries. They are split between the training set ( $\sim 14$ K samples), the validation set ( $\sim 818$

---

<sup>3</sup>In 1M-MATH, we keep sequences up to 768 tokens.

samples) and the test set ( $\sim 819$  samples). On the other hand, DialogSum contains approximately 14K conversation-summary pairs, focusing more on daily-life formal conversations, such as business negotiations. These pairs are split between the training set ( $\sim 12.5$ K samples), the validation set ( $\sim 500$  samples) and the test set ( $\sim 1.5$ K samples).

**Training details** Across all experiments, we finetune the model for 3 training epochs, following the same setup (optimization details, weights’ precision), with the mathematical reasoning experiments. We filter out training sequences that are longer than 768 tokens.

**Evaluation** We calculate ROUGE scores against the ground truth reference summaries, select the best checkpoint with respect to ROUGE-L on the validation set, and report ROUGE-1, ROUGE-2 and ROUGE-L scores on the test set.

### A.1.3 Best performing configurations

In Tables 9 and 10, we provide the best performing configurations ( $d_{\max}$  and  $a$ ) for MuToR, across all the training splits that were utilized in our experiments. For Multi-Token, after carefully tuning the same hyperparameters, we found that the optimal configuration across all experiments is  $d_{\max} = 2, a = 0.1$ .

Table 9: MuToR’s best performing hyperparameters for Gemma 2B model.

Training data	$d_{\max}$	$a$
GSM8K	4	0.3
1M-GSM	4	0.3
2M-GSM	6	0.1
1M-MATH	3	0.2
AQUA-RAT	4	0.3
SAMSum	3	0.5
DialogSum	4	0.5

Table 10: MuToR’s best performing hyperparameters for Llama 3 8B model.

Training data	$d_{\max}$	$a$
GSM8K	3	0.3
1M-GSM	6	0.1
2M-GSM	6	0.1
1M-MATH	4	0.1

## A.2 Autoregressive Image Generation

**Training details** In the autoregressive image generation experiments, we pretrain LlamaGen-B model (approximately 111M parameters). Specifically, the model itself is an autoregressive decoder-only transformer, similar to Llama [Touvron et al., 2023a]. It utilizes RoPE as positional bias, extended to two-dimensions. The model also involves learnable class embeddings, which are used as conditionals for image generation. During training, the class embeddings are dropped with a probability of 0.1, to enable the use of classifier-free guidance at inference time.

To tokenize the image patches, we use a VQ-VAE tokenizer provided by Sun et al. [2024], with codebook size equal to 16384 and embedding dimension equal to 8. Unlike LlamaGen, we pre-tokenize the dataset with the ADM’s preprocessing scheme [Dhariwal and Nichol, 2021], resulting in two crops per image.

Both Next-Token baseline and MuToR are trained for 360K update steps, using AdamW optimizer with  $\beta_1 = 0.9, \beta_2 = 0.95$  and weight decay = 0.05. We employ a constant learning rate, equal to 0.0004, and a batch size of 1024.

For MuToR’s implementation, we empirically tune the loss coefficient  $a$  to be 0.5, so that  $\mathcal{L}_{\text{reg}}$  has equal contribution with  $\mathcal{L}_{\text{ntp}}$  (Equation 6). This indicates that the auxiliary loss provides valuable supervision that the model leverages during pretraining, to improve its learned representations. In the experiments that fewer registers are inserted, we sample random positions for each training sample.

**Evaluation** To benchmark generative performance, we sample 50,000 images at  $256 \times 256$  resolution, setting temperature = 1.0 and a fixed random seed for fair comparison. We use classifier-free guidance [Ho and Salimans, 2022] with scale  $s = 2.0$ , which was reported as an optimal value in the original LlamaGen. Thus, the logit  $l_g$  is formed as such:  $l_g = l_u + 2(l_c - l_u)$ , where  $l_u$  denotes the unconditional logit (with the class embedding dropped) and  $l_c$  denotes the conditional logit. Then,

the generated samples are used to calculate the performance metrics, using the Tensorflow scripts from Dhariwal and Nichol [2021].

### A.3 Synthetic Data

We setup our experiments using the official implementation from Bachmann and Nagarajan [2024] (regarding the model’s architecture and optimization). All model configurations (Next-Token and MuToR) are trained for a sufficient number of epochs.

MuToR’s implementation for the star graphs problem follows subsection 3.3, with a task-specific modification; we sample the offset  $d$  from  $\{2, \dots, d_{\max}\}$ , thus excluding the next-token from the register’s prediction. In this way, we prevent the registers from learning the "Clever Hans Cheat", and enable them to focus on *planning* look-ahead predictions. In these experiments, we set  $d_{\max} = 4$  and  $a = 0.5$  for the graphs with path length = 5 and  $d_{\max} = 6$ ,  $a = 0.3$  for graphs with path length = 10, which are more challenging.

## B Additional experiments

### B.1 Matching the training compute

Tables 11 and 12 report the results of finetuning baseline methods for twice the number of training epochs than MuToR, in order to investigate whether improved downstream performance can be derived from increasing the training compute. Due to our limited computational resources, we conduct these experiments using Gemma 2B model.

Table 11: Downstream accuracy (%) in mathematical reasoning benchmarks across different number of training epochs. The subheaders refer to the training split used in each experiment. Results are averaged across three seeded runs.

Method	# Epochs	GSM8K		MATH500
		GSM8K	1M-GSM	1M-MATH
Next-Token	5	38.87	66.09	26.73
	10	37.91	65.07	27.07
Multi-Token	5	40.66	66.69	26.87
	10	39.98	64.92	26.73
MuToR	5	<b>42.10</b>	<b>68.33</b>	<b>28.13</b>

Table 12: Rouge metrics comparison across different number of training epochs. We select the checkpoint with the higher ROUGE-L in the validation set, and report ROUGE scores on the test set.

Dataset	Method	#Epochs	ROUGE-1	ROUGE-2	ROUGE-L
SAMSum	Next-Token	3	51.47	27.29	43.23
		5	51.67	27.64	43.32
	Multi-Token	3	51.90	27.44	43.50
		5	51.04	26.35	42.48
	MuToR	3	<b>52.32</b>	<b>28.08</b>	<b>44.09</b>
DialogSum	Next-Token	3	47.23	20.91	38.77
		5	46.92	20.39	38.25
	Multi-Token	3	47.98	21.23	39.25
		5	47.46	20.76	38.97
	MuToR	3	<b>48.22</b>	<b>21.71</b>	<b>39.48</b>