Proxy-Tuning: Tailoring Multimodal Autoregressive Models for Subject-Driven Image Generation

Supplementary Material

1. Analysis of Proxy-Tuning

To further understand the underlying mechanisms driving the "weak-to-strong" generalization [1] phenomenon observed in our Proxy-Tuning framework, we conduct a comprehensive feature-level analysis. Specifically, we hypothesize that the phenomenon stems from the autoregressive (AR) model's inherent capacity to act as a statistical noise filter during the next-token prediction process [6, 10].

Mechanistic Insights. In the proxy dataset generated by the diffusion supervisor, the feature of the target subject represents a consistent, high-frequency signal. Conversely, the generative artifacts, distortions, or background inconsistencies introduced by the diffusion model are inherently stochastic and exhibit high variance [5, 7]. When training the AR model on this proxy data, these random artifacts manifest as low-probability outliers within the discrete token space [2, 11]. Driven by the cross-entropy objective, the AR model exhibits strong mode-seeking behavior [3, 9]. Instead of perfectly memorizing the noisy distribution, the AR model assigns negligibly low probabilities to these infrequent artifact tokens, effectively truncating the distribution to favor the dominant mode, which corresponds to the consistent subject identity [4].

Empirical Validation. To empirically validate this mechanism, we perform a large-scale feature distribution analysis. We generate 100 samples per subject using both the diffusion supervisor and the fine-tuned AR student, subsequently extracting and analyzing their feature deviations (as shown in Figure 1). Our results reveal two critical findings:

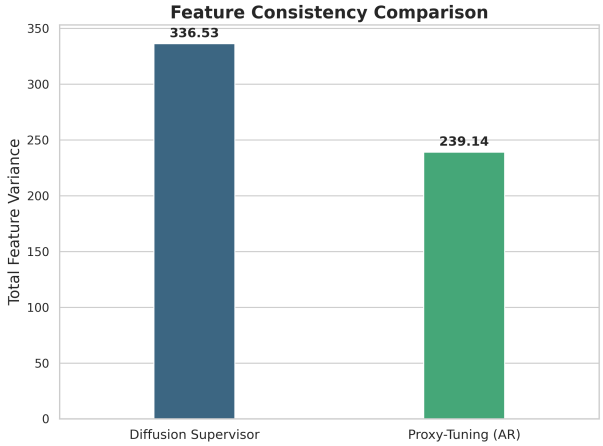
- **Variance Reduction:** The feature representations of the images generated by the proxy-tuned AR model exhibit significantly lower variance compared to those from the diffusion supervisor, demonstrating the AR model's effectiveness in filtering stochastic noise [9, 13].
- **Distribution Truncation:** Histogram analysis of the feature deviations explicitly shows that the AR model successfully truncates the "long-tail" outliers [4] that are prevalently generated by the diffusion supervisor. By cutting off these noisy tails, the AR model converges on a much more consistent and faithful representation of the subject's identity.

These analytical results firmly validate that AR models leverage their structural priors [12] and discrete token spaces [2] to actively filter the supervisor's stochastic noise rather than passively overfitting to it.

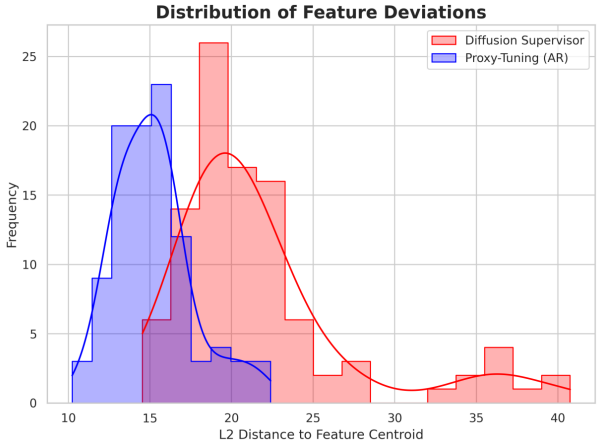
2. Generalization of Proxy-Tuning

Generalization to complex scenarios. While our primary evaluations demonstrate the efficacy of Proxy-Tuning under standard subject-driven generation protocols, real-world applications often demand high fidelity under complex and highly compositional textual conditions. To rigorously evaluate the robustness and generalization capabilities of our method, we design an additional set of qualitative experiments utilizing intricately detailed, open-ended prompts. These prompts are specifically crafted to test the model's ability to maintain subject identity while simultaneously adhering to complex attribute bindings, such as specific attire, nuanced lighting, and detailed background contexts. As illustrated in Figure 2, the AR student (Proxy-Tuning) consistently synthesizes images that exhibit superior visual alignment and fidelity compared to the diffusion supervisor. Notably, the AR student achieves impressive prompt adherence—accurately rendering dense environmental requests while faithfully preserving the unique characteristics of the customized subject. In contrast, the diffusion supervisor frequently struggles with concept bleeding or loses precise subject identity when overwhelmed by dense textual conditions.

Generalization to Superior Customization Methods. A natural question arises regarding whether the Proxy-Tuning framework can generalize to and benefit from superior subject-driven customization methods specifically designed for diffusion models. To investigate whether the weak-to-strong generalization phenomenon persists when the supervisor is already highly optimized, we employ AttnDreamBooth [8] which is a leading attention-optimization method to act the diffusion supervisor within our pipeline. As detailed in Table 1, even when guided by a highly performant teacher, the AR student continues to show significant improvements in subject fidelity. This consistent performance gain, which mirrors our findings with large-scale models like FLUX.1, confirms that the weak-to-strong generalization phenomenon is a fundamental characteristic of the AR proxy-tuning paradigm, robust even against powerful supervisors. By operating as a black-box, model-agnostic pipeline, our AR proxy-tuning paradigm strictly requires only the synthesized image outputs from the supervisor. It does not depend on shared attention mechanisms, gradient matching, or architectural similarities between the teacher and the student. This unique property not only facilitates the current results but also ensures that AR models can seamlessly inherit and amplify future advancements in diffusion-



(a) The AR model (Green) significantly reduces feature variance (**336.53 →239.14**) compared to the Diffusion supervisor, indicating superior consistency.



(b) The histogram reveals that the AR model (Blue) successfully **truncates the "long-tail" of outliers** and artifacts present in the Diffusion supervisor's distribution (Red), empirically validating our hypothesis that AR models filter out stochastic noise via mode-seeking.

Figure 1. **Empirical validation of the weak-to-strong generalization phenomenon via feature distribution analysis.** (a) Feature Consistency Comparison: The proxy-tuned AR student exhibits a significantly lower total feature variance (239.14) compared to the diffusion supervisor (336.53), demonstrating superior subject consistency and effective filtering of generative noise. (b) Distribution of Feature Deviations: A histogram of the L2 distances to the subject centroid reveals that the AR model actively truncates the "long-tail" of outliers and stochastic artifacts prevalent in the supervisor's distribution. This mode-seeking behavior enables the AR model to converge on a more faithful and consistent representation of the target subject identity.

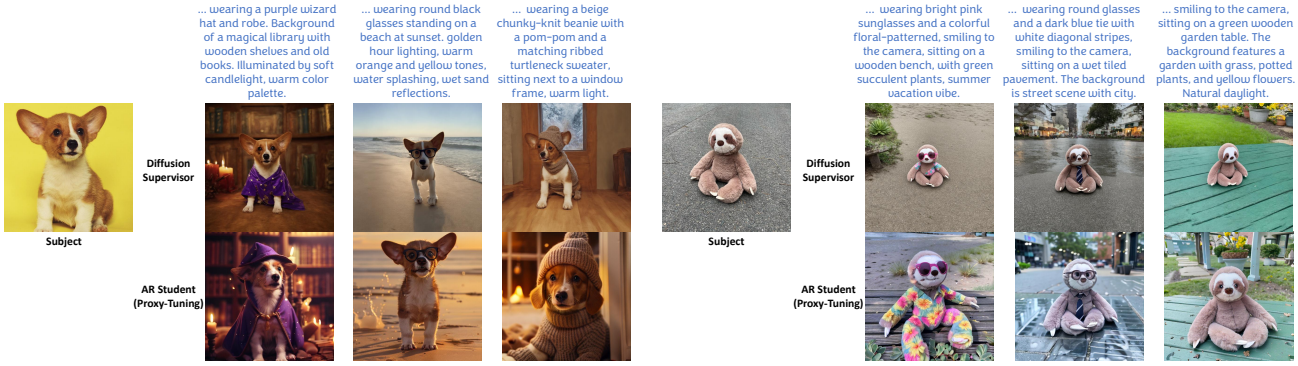


Figure 2. **Qualitative evaluation under complex, open-ended scenarios.**

Table 1. **Compatibility and performance with superior customization supervisors.**

	CLIP-I	CLIP-T	DINO
AttnDreamBooth [8]	0.7938	**0.32**	0.6996
Ours	**0.8267**	0.29	**0.7744**

based customization without requiring architecture-specific redesigns.

3. Details of User Study

To complement our quantitative metrics and provide a rigorous human perception-based evaluation, we conduct a user study comparing four customization approaches: direct subject tuning of the diffusion model (SDXL), direct subject tuning of the AR model (Lumina-mGPT), Proxy-Tuning of the diffusion model, and our proposed Proxy-Tuning of the AR model. The study involves 27 participants, and the evaluation protocol is systematically structured as follows. Each participant is tasked with assessing images generated by all four aforementioned methods. Within the evaluation of each method, the assessment covers 3 distinct subjects. Furthermore, each subject is evaluated under 3 different text prompts, with 2 synthesized images presented per prompt. During the survey, participants are provided with the reference image of the target subject and the specific text prompt, alongside the corresponding generated images. They are

asked to grade the synthesized results on a standard 1-to-5 Likert scale (with 5 being the highest) based on three core criteria: *image quality* (visual appeal and absence of artifacts), *subject fidelity* (how well the images preserve the reference subject's identity and characteristics), and *prompt fidelity* (how well the images align with the given text condition).

References

[1] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. 1

[2] 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. 1

[3] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017. 1

[4] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 1

[5] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 1

[6] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, 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. 1

[7] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023. 1

[8] Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 39869–39900, 2024. 1, 2

[9] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 1

[10] 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. 1

[11] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[12] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[13] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1