POLYOMINOGEN: A Controlled Testbed for Understanding Memorization and Compositional Generalization in Conditional Diffusion Models

Aishi Huang
aishuang1999@gmail.com
Individual Researcher

Abstract

Deep generative models can produce samples that appear novel, yet in natural-image domains it is often unclear whether such samples reflect memorization, interpolation, or systematic generalization beyond the observed training support. We introduce POLYOMINOGEN, a controlled testbed for studying memorization and compositional generalization in conditional diffusion models. Each image is rendered from an exact symbolic tuple specifying shape, color, orientation, and position, enabling train–test splits in which specific attribute combinations or geometric transformations are withheld by construction. Across three U-Net DDPM runs, generated objects remain largely valid under held-out conditions, while tuple accuracy drops substantially, particularly for withheld geometric transformations. Pretrained initialization also outperforms training from scratch across all tested few-shot settings, and the novel-valid held-out metric remains stable across a range of nearest-neighbor thresholds. POLYOMINOGEN provides a lightweight diagnostic framework for probing when generative models memorize, when they generalize, and when they fail under controlled distribution shift.

1. Introduction

Deep generative models now produce high-quality images, videos, language, and scientific artifacts. Yet the apparent novelty of a generated sample is often difficult to interpret. A sample that appears new may be a near-copy of a training example, a local interpolation among observed examples, or a systematic recombination of learned factors into a genuinely held-out configuration. Distinguishing among these possibilities is central to understanding when generative models memorize, when they generalize, and when they support structured composition.

This distinction is difficult to study in natural-image domains because large generative models are typically trained on massive corpora whose full contents are unknown. When a model successfully generates a requested composition, it is usually impossible to know whether that exact combination was absent from training. Conversely, when a model fails, it is difficult to determine whether the failure reflects missing concepts, weak conditioning, insufficient data coverage, or a failure to bind otherwise learned factors.

Recent benchmarks have improved the evaluation of compositional text-to-image generation. T2I-CompBench evaluates attribute binding, object relationships, and complex compositions (Huang et al., 2023); GenEval evaluates object co-occurrence, position, count, and color (Ghosh et al., 2023); and ConceptMix introduces controllable compositional difficulty through automatically generated prompts (Wu et al., 2024). These benchmarks are valuable for measuring whether pretrained systems satisfy prompts, but they are less suited for explaining *why* a model succeeds or fails because the relevant training support is not controlled.

We introduce POLYOMINOGEN, a controlled visual testbed for studying memorization and compositional generalization in conditional generative models. Each image is generated from a symbolic tuple specifying a polyomino shape, color, orientation, and position. Because the rendering process is deterministic and the latent factors are known, we can construct train–test splits in which particular attribute combinations or geometric transformations are absent during

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

training while remaining fully evaluable at test time. This makes it possible to ask whether a model can generate a valid object matching a held-out tuple, and whether that object is distinct from all training examples.

The testbed is motivated by prior controlled work on Polyomino World, which studied neural networks performing shape, size, and color classification (Huang et al., 2022). That work showed a gap between immediate transfer and later transfer learning: models often failed to classify novel cases immediately, yet pretrained models learned them faster than randomly initialized models. POLYOMINOGEN brings this distinction into the generative setting by asking whether diffusion models that fail zero-shot held-out generation nevertheless show faster few-shot adaptation, indicating reusable factors that were not compositionally expressed.

We instantiate POLYOMINOGEN with a conditional U-Net DDPM and evaluate it across three random seeds under fixed attribute-composition and geometric-transformation splits. In addition to the main ID–OOD evaluation, we test sensitivity to the nearest-neighbor threshold and compare few-shot adaptation from pretrained and scratch initialization. Our contributions are:

- We introduce POLYOMINOGEN, a controlled visual testbed with enumerable training support and exact symbolic ground truth over shape, color, orientation, and position.
- We define known-support train–test splits that isolate held-out attribute composition and held-out geometric transformation.
- We propose a memorization-aware protocol that separates validity, condition following, training-set proximity, and novel-valid held-out generation.
- Across three random seeds, we characterize the ID–OOD generalization gap, evaluate sensitivity to the nearest-neighbor threshold, and test whether pretrained initialization facilitates adaptation to held-out tuples.

2. Related Work

Diffusion models and compositional generation. Diffusion models have become a dominant framework for image synthesis, with latent diffusion enabling efficient high-resolution generation and transformer-based diffusion models providing a modern scalable architecture class (Rombach et al., 2022; Peebles & Xie, 2023; Esser et al., 2024). Despite their strong sample quality, text-to-image models continue to struggle with compositional generation, including attribute binding, spatial relationships, and multi-object reasoning (Huang et al., 2023; Ghosh et al., 2023). POLYOMINOGEN does not aim to match the visual diversity of open-world benchmarks; instead, it provides a controlled setting where the training support is known exactly.

Memorization in generative models. A growing body of work shows that diffusion models can reproduce or closely resemble training examples. Somepalli et al. analyze data replication in diffusion-generated images, while Carlini et al. demonstrate extraction of training examples from diffusion models (Somepalli et al., 2023a; Carlini et al., 2023). Subsequent work studies how copying behavior depends on conditioning, captions, dataset properties, and training dynamics (Somepalli et al., 2023b; Gu et al., 2023; Bonnaire et al., 2025). These findings motivate evaluation protocols that compare generated samples directly against the training set. POLYOMINOGEN makes this comparison explicit because the complete training set is known and enumerable.

Compositional evaluation benchmarks. Recent benchmarks evaluate whether generated images satisfy complex prompts. T2I-CompBench covers attribute binding, object relationships, and complex compositions (Huang et al., 2023). GenEval proposes an automated object-focused framework for evaluating text-to-image alignment on co-occurrence, count, position, and color (Ghosh et al., 2023). ConceptMix automatically constructs prompts with controllable compositional difficulty and evaluates whether specified concepts appear in the generated image (Wu et al., 2024). These benchmarks primarily measure prompt satisfaction for pretrained systems. In contrast, POLYOMINOGEN measures known-support generalization: whether a model trained without a specific composition can generate that composition later.

Relationship to controlled diffusion studies. Okawa et al. (2023) study how compositional capabilities emerge as a function of concept frequency and the structure of the underlying data-generating process. Park et al. (2024) examine

Table 1. Core factors controlled in POLYOMINOGEN.

Factor	Symbol	Examples
Shape	s	domino, tromino, L-tetromino
Color	c	red, green, blue, yellow
Orientation	o	rotations and flips valid for s
Position	p	legal grid location

capabilities that may be acquired internally before they can be elicited through ordinary conditioning, while Yang et al. (2025) develop a theoretical abstraction of these learning dynamics. POLYOMINOGEN is complementary to these settings. Its primary contribution is a memorization-aware evaluation protocol over a spatially explicit visual domain: the complete training support is enumerable, geometric transformations can be withheld by construction, generated outputs can be decoded symbolically, and conditionally correct samples can be separated into training-neighbor-like and novel-valid generations. The few-shot transfer diagnostic additionally tests whether zero-shot failure coexists with faster adaptation than learning the same held-out tuples from scratch.

3. The POLYOMINOGEN Testbed

POLYOMINOGEN is a controlled visual testbed for studying when conditional generative models memorize, when they generalize, and when they fail to compose learned factors under distribution shift. The testbed intentionally sacrifices open-world visual diversity for exact control over the data-generating process. This makes it possible to know which compositions were present during training, which were withheld, and whether generated samples match or copy observed examples.

3.1. Symbolic Image Generator

Each image is rendered from a tuple

$$z = (s, c, o, p), \tag{1}$$

where s is a polyomino shape, c is a discrete color, o is a valid orientation, and p is a legal grid position. A deterministic renderer maps the tuple to an image:

$$x = R(z). \tag{2}$$

The shape set contains simple polyominoes composed of one to four connected cells, following the controlled visual structure of Polyomino World (Huang et al., 2022). Each image contains a single colored object on a neutral background. Since all factors are discrete and known, generated images can be decoded by rule-based procedures rather than learned classifiers.

A model is trained conditionally: given a requested symbolic tuple z, it must generate an image matching that tuple. In the main experiments, conditions are represented as one-hot factor embeddings. This removes ambiguity from natural-language prompting and isolates whether the model can bind the requested factors into a valid visual object.

3.2. Known-Support Splits

Let $\mathcal{Z}$ be the set of all legal symbolic tuples. We partition it into disjoint training and held-out supports:

$$\mathcal{Z}_{\text{train}} \cap \mathcal{Z}_{\text{test}} = \emptyset. \tag{3}$$

Training images are generated from $\mathcal{Z}_{\text{train}}$, while evaluation conditions are drawn from $\mathcal{Z}_{\text{test}}$. Thus, a held-out condition is not merely a new image sample; it is a symbolic composition absent from training by construction.

We focus on two split families. *Held-out attribute composition* withholds specific factor combinations while keeping each individual factor observed; for example, a model may observe blue dominoes and red trominoes but never blue trominoes. This tests whether the model can bind known factors into an unseen composition. *Held-out geometric transformation* withholds rotations or flips of known shapes, testing whether the model learns reusable geometric structure or orientation-specific templates. Together, these splits test whether conditional diffusion models generalize beyond observed support in both attribute and geometric factor spaces.

4. Evaluation Protocol

The evaluation protocol separates four behaviors that are often conflated in open-world generation: producing a valid object, satisfying the requested condition, generating a held-out composition, and copying or closely matching a training example. Because each image has an exact symbolic description, all metrics are computed using a rule-based decoder and the known training set.

4.1. Decoding and Conditional Correctness

For each requested tuple z, a model generates K samples $\{\hat{x}_{z,k}\}_{k=1}^{K}$. A deterministic decoder D maps each generated image either to a decoded tuple $\hat{z} = D(\hat{x})$ or to an invalid symbol $\perp$. The decoder checks whether the image contains a single connected object, estimates its mask, color, orientation, and position, and matches the mask to the closest legal polyomino shape. Outputs with no object, multiple objects, disconnected masks, ambiguous colors, or illegal shapes are marked invalid.

We report *validity*, the fraction of generated samples decoded as legal polyominoes:

$$\text{Valid} = \frac{1}{K|\mathcal{Z}_{\text{eval}}|} \sum_{z,k} 1[D(\hat{x}_{z,k}) \neq \perp]\,. \tag{4}$$

We also report *tuple accuracy*, the fraction of generated samples whose decoded tuple exactly matches the requested condition:

$$\text{Acc}_{\text{tuple}} = \frac{1}{K|\mathcal{Z}_{\text{eval}}|} \sum_{z,k} 1[D(\hat{x}_{z,k}) = z]\,. \tag{5}$$

Factor-level accuracies for shape, color, orientation, and position are reported in the same way, with invalid samples counted as incorrect.

4.2. Held-Out Generalization and Memorization

To measure known-support generalization, we compute tuple accuracy separately on in-support conditions $\mathcal{Z}_{\text{ID}}$ and held-out conditions $\mathcal{Z}_{\text{OOD}}$. The generalization gap is

$$\Delta_{\text{gen}} = \text{Acc}_{\text{tuple}}(\mathcal{Z}_{\text{ID}}) - \text{Acc}_{\text{tuple}}(\mathcal{Z}_{\text{OOD}}). \tag{6}$$

Correct held-out generation alone does not rule out copying a nearby training example. For every generated image $\hat{x}$, we compute its nearest-neighbor distance to the training set:

$$d_{\min}(\hat{x}) = \min_{x_i \in \mathcal{D}_{\text{train}}} d(\hat{x}, x_i), \tag{7}$$

using mask-level and pixel-level distances. A sample is training-neighbor-like when $d_{\min}(\hat{x}) \leq \tau$ for a fixed threshold τ.

Our primary metric is *novel-valid held-out generation*:

$$\text{NVH}_{\tau} = \frac{1}{K|\mathcal{Z}_{\text{OOD}}|} \sum_{z,k} 1[D(\hat{x}_{z,k}) = z \wedge d_{\min}(\hat{x}_{z,k}) > \tau]\,. \tag{8}$$

This metric counts samples that are valid, conditionally correct, drawn from held-out support, and not nearest-neighbor-like.

We report the training-neighbor rate over all generated OOD samples. In contrast, NVH_{τ} counts only samples that both satisfy the requested held-out tuple and fall outside the nearest-neighbor threshold. The two quantities therefore measure different aspects of model behavior and need not sum to one.

4.3. Few-Shot Transfer Diagnostic

A zero-shot failure on held-out tuples does not necessarily imply that the model failed to learn the relevant factors. It may have learned shape, color, orientation, or position separately, but failed to bind them under a new condition. To

Table 2. Evaluation metrics used in POLYOMINOGEN.

Metric	Definition
Validity	Fraction of samples decoded as legal polyomino images.
Tuple accuracy	Fraction whose decoded tuple matches the requested condition.
Generalization gap	Difference between in-support and held-out tuple accuracy.
Training-neighbor rate	Fraction within threshold τ of a training image.
NVH_τ	Held-out samples that are valid, conditionally correct, and not training-neighbor-like.
Transfer gain	Few-shot accuracy gain from pretrained initialization over scratch training.

test this, we fine-tune the pretrained model on m examples per held-out tuple and compare it to a randomly initialized model trained on the same examples. The transfer gain is

$$\text{TG}(m) = \text{Acc}_{\text{tuple}}^{\text{pretrained}}(m) - \text{Acc}_{\text{tuple}}^{\text{scratch}}(m). \tag{9}$$

A positive transfer gain indicates that pretraining produced reusable structure that helps adapt to held-out compositions, even when zero-shot generation fails.

5. Experimental Setup

We evaluate POLYOMINOGEN using a conditional U-Net DDPM trained from scratch. The objective is not to maximize natural-image fidelity, but to measure memorization and known-support generalization under controlled symbolic splits.

5.1. Dataset Instances

Images are rendered at 32×32 resolution with one colored polyomino on a neutral background. The shape set contains a monomino, domino, two trominoes, and two tetrominoes, with four discrete colors and all shape-valid orientations. We use one fixed held-out support for each split family. The attribute-composition split contains 42 training and 14 held-out shape–color conditions. The geometric-transformation split contains 40 training and 16 held-out shape–orientation conditions. Each individual factor remains represented in training, while the target factor combinations are absent by construction. We repeat training and sampling with random seeds $\{0, 1, 2\}$ and report the mean and standard error across seeds. Thus, the reported variation reflects optimization and sampling randomness rather than variation across split constructions.

5.2. Model and Training

The model is a small conditional U-Net DDPM trained with the standard noise-prediction objective (Ho et al., 2020). Symbolic factors are represented with learned embeddings and injected through the timestep-conditioning blocks. The U-Net uses 24 base channels and is trained for 1,500 updates with batch size 64, learning rate 2×10^{-3}, 50 diffusion steps, and a linear schedule ending at $\beta = 0.12$. At evaluation time, we generate $K = 8$ samples for every ID and OOD condition and decode them using the rule-based evaluator from Section 4. The default nearest-neighbor threshold is $\tau_0 = 0.003$. We additionally evaluate

$$\tau \in \{0.5\tau_0, \tau_0, 2\tau_0, 4\tau_0\}$$

to test whether memorization-aware conclusions depend on this choice.

5.3. Few-Shot Transfer

After zero-shot evaluation, we fine-tune the pretrained model on $m \in \{1, 2, 4, 8\}$ examples per held-out tuple. A randomly initialized U-Net is trained on the same examples for the same 180-update adaptation budget. We compare OOD tuple accuracy under pretrained and scratch initialization. Because the number of updates is fixed across m, these experiments test the benefit of initialization rather than a monotonic sample-scaling law.

6. Results

We report mean performance and standard error across three random seeds. All runs use the same held-out support for each split, so the reported variation reflects optimization and sampling randomness rather than variation across

independently constructed splits. Compared with the initial pilot, the replicated evaluation increases the number of generated samples per condition from $K = 2$ to $K = 8$.

6.1. Valid Generation Does Not Imply Held-Out Correctness

Table 3 reports validity, tuple accuracy, training-neighbor rate, and novel-valid held-out generation. On the attribute-composition split, ID tuple accuracy is $51.4 \pm 7.8\%$, compared with $31.2 \pm 5.1\%$ OOD accuracy, yielding a 20.1 ± 6.7 percentage-point generalization gap. The corresponding validity rates remain high at $86.7 \pm 3.8\%$ for ID samples and $90.8 \pm 4.8\%$ for OOD samples.

The geometric-transformation split produces a substantially larger separation. ID tuple accuracy reaches $69.2 \pm 3.9\%$, whereas OOD tuple accuracy is only $2.9 \pm 0.7\%$, yielding a 66.3 ± 3.3 percentage-point gap. Nevertheless, OOD validity remains $86.5 \pm 1.6\%$. Thus, the model frequently generates legal polyominoes under held-out geometric conditions while failing to realize the requested orientation. These results demonstrate that structural validity and conditional correctness are distinct properties.

Table 3. U-Net DDPM results as percentages, reported as mean ± standard error across three seeds. NN_τ is computed over all OOD samples using the default threshold $\tau_0 = 0.003$.

Split	ID Val.	ID Acc.	OOD Val.	OOD Acc.	Gap	NN_τ	NVH_τ
Attribute	86.7±3.8	51.4±7.8	90.8±4.8	31.2±5.1	20.1±6.7	7.7±6.4	31.2±5.1
Geometry	91.5±0.6	69.2±3.9	86.5±1.6	2.9±0.7	66.3±3.3	14.8±5.5	2.9±0.7

6.2. Memorization-Aware Conclusions Are Stable Across Thresholds

The absolute training-neighbor rate depends on the nearest-neighbor threshold, as expected. Increasing τ from $0.5\tau_0$ to $4\tau_0$ increases the fraction of OOD outputs classified as training-neighbor-like from 0.0% to $42.3 \pm 2.8\%$ on the attribute split and from 0.0% to $67.2 \pm 4.3\%$ on the geometry split.

In contrast, NVH_τ remains stable over most of the tested range. On the geometry split, NVH_τ remains $2.9 \pm 0.7\%$ for all four thresholds. On the attribute split, it remains $31.2 \pm 5.1\%$ through $2\tau_0$ and decreases only slightly to $29.8 \pm 4.3\%$ at $4\tau_0$. Therefore, the principal conclusion that geometric held-out generation is substantially weaker than attribute recombination is not an artifact of the default threshold. Complete threshold-sensitivity values are reported in Appendix B.

6.3. Qualitative Held-Out Samples

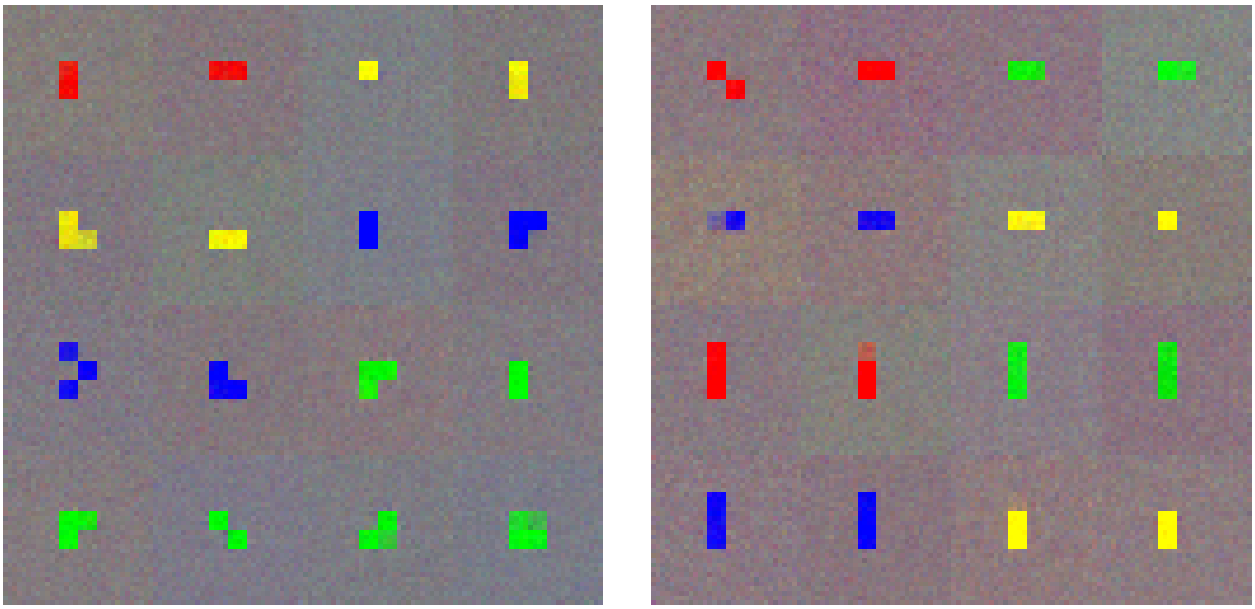
Figure 1 shows representative OOD generations from seed 0. Many outputs are visually valid and preserve individual factors such as color or coarse size, but fail the complete requested tuple. In particular, geometric OOD conditions often produce a valid seen orientation rather than the withheld orientation. This illustrates why qualitative inspection or validity alone is insufficient for evaluating compositional generalization.

6.4. Pretrained Initialization Improves Few-Shot Adaptation

We next test whether zero-shot OOD failures reflect missing factors or failures to bind learned factors compositionally. After zero-shot evaluation, we fine-tune the pretrained U-Net on $m \in \{1, 2, 4, 8\}$ examples per held-out tuple and compare it with a randomly initialized model trained using the same examples and adaptation budget.

Figure 2 shows that pretrained initialization yields higher mean OOD accuracy for both split families at every tested value of m. At $m = 1$, attribute accuracy reaches $54.8 \pm 5.4\%$ after pretrained adaptation, compared with $5.1 \pm 1.1\%$ from scratch. Geometry accuracy reaches $60.7 \pm 1.6\%$, compared with $3.4 \pm 1.4\%$ from scratch.

The curves are not monotonic in m, and each condition uses a fixed 180-update adaptation budget. We therefore interpret the consistent advantage over scratch initialization as evidence that pretraining provides reusable structure, rather than as evidence that accuracy scales monotonically with the number of adaptation examples. Complete numerical results are reported in Appendix C.



Attribute split Geometry split

Figure 1. Representative OOD generations from seed 0. Outputs may be structurally valid while failing the requested held-out tuple, motivating symbolic decoding rather than qualitative inspection alone.

Exploratory difficulty ablation. Appendix D reports a single-seed attribute-composition ablation using two, three, and four colors. The results are non-monotonic and therefore do not support a simple claim that compositional generalization degrades solely as the number of colors increases. Instead, the experiment suggests that difficulty also depends on the selected held-out support and training dynamics.

6.5. Summary

The replicated experiments support three conclusions. First, high validity does not imply correct generation outside the observed symbolic support. Second, held-out geometric transformations are substantially more difficult than held-out shape–color combinations under the current setup. Third, pretrained initialization consistently facilitates adaptation to held-out tuples relative to training from scratch. The threshold analysis further shows that the qualitative distinction between attribute and geometric generalization remains stable across a range of nearest-neighbor definitions.

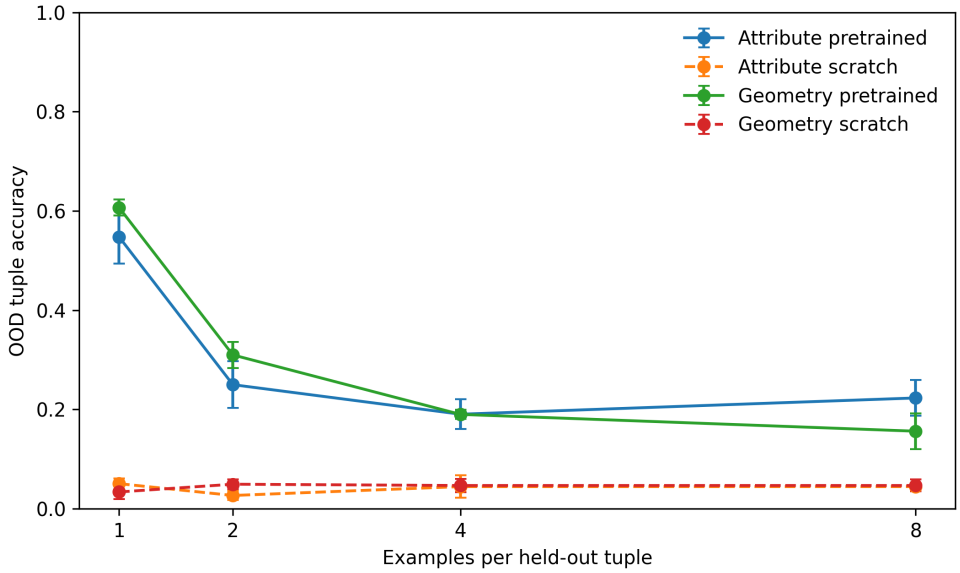


Figure 2. Few-shot OOD tuple accuracy, reported as mean ± standard error across three seeds. Pretrained initialization outperforms scratch initialization at every tested adaptation-set size, although performance is non-monotonic in m under the fixed update budget.

7. Discussion, Limitations, and Conclusion

POLYOMINOGEN makes a narrow but important question experimentally tractable: when a conditional generative model produces a sample that appears novel, is it reproducing observed structure or composing factors beyond the training support? Across three seeds, the U-Net DDPM produces predominantly valid objects on both ID and OOD conditions, but tuple correctness deteriorates under held-out support. The especially large geometric gap indicates that learning to render a legal shape is distinct from learning a transformation-aware representation that can satisfy an unseen orientation request.

The few-shot analysis offers a complementary view. Pretrained initialization consistently outperforms scratch initialization, suggesting that zero-shot failure can coexist with representations that facilitate subsequent adaptation. Because the curves are non-monotonic under a fixed update budget, this result should be interpreted as an initialization effect rather than a sample-efficiency scaling law.

The nearest-neighbor sensitivity analysis also clarifies the scope of the memorization claim. The fraction of all outputs classified as training-neighbor-like depends strongly on τ, as expected. In contrast, NVH changes little over the tested range, indicating that the qualitative comparison between attribute and geometric generalization is not an artifact of the default threshold.

Several limitations remain. The study evaluates one U-Net DDPM architecture, one fixed held-out support per split, and simple single-object scenes. Results therefore should not be generalized directly to latent diffusion, Diffusion Transformers, or web-scale text-to-image systems. Although three seeds measure optimization and sampling variability, future work should additionally average over independently constructed held-out supports. The exploratory color-count ablation is single-seed and does not establish a monotonic relation between factor count and difficulty. Extending the testbed to multi-object scenes, relational conditions, richer renderers, and additional generative architectures remains important future work.

We introduced POLYOMINOGEN, a controlled visual testbed for separating valid generation, held-out condition following, training-set proximity, and few-shot transfer. The replicated experiments preserve the original conclusion while providing stronger evidence that validity and compositional correctness can diverge sharply under known distribution shift.

References

Bonnaire, T., Urfin, R., Biroli, G., and Mézard, M. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *Advances in Neural Information Processing Systems*, 2025.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, 2023.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024.

Ghosh, D., Hajishirzi, H., and Schmidt, L. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Huang, A., Huebner, P. A., and Willits, J. A. Generalization and transfer learning in neural networks performing shape, size, and color classification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

Huang, K., Sun, K., Xie, E., Li, Z., and Liu, X. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Okawa, M., Lubana, E. S., Dick, R. P., and Tanaka, H. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. In *Advances in Neural Information Processing Systems*, 2023.

Park, C. F., Okawa, M., Lee, A., Tanaka, H., and Lubana, E. S. Emergence of hidden capabilities: Exploring learning dynamics in concept space. In *Advances in Neural Information Processing Systems*, 2024.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, 2023b.

Wu, X., Yu, D., Huang, Y., Russakovsky, O., and Arora, S. ConceptMix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024.

Yang, Y., Park, C. F., Lubana, E. S., Okawa, M., Hu, W., and Tanaka, H. Swing-by dynamics in concept learning and compositional generalization. In *The Thirteenth International Conference on Learning Representations*, 2025.

A. Additional Experimental Details

All replicated runs use seeds $\{0, 1, 2\}$ and the same symbolic held-out support. The model uses 24 base U-Net channels, batch size 64, learning rate 2×10^{-3}, 1,500 base-training updates, 50 diffusion steps, and a linear noise schedule ending at $\beta = 0.12$. Evaluation uses eight generated samples per symbolic condition.

The attribute split contains 42 training and 14 held-out conditions, producing 336 ID and 112 OOD generated samples per seed. The geometry split contains 40 training and 16 held-out conditions, producing 320 ID and 128 OOD generated samples per seed. Few-shot adaptation uses 180 updates for both pretrained and scratch initialization.

B. Nearest-Neighbor Threshold Sensitivity

Table 4. Sensitivity of memorization-aware OOD metrics to the nearest-neighbor threshold. Values are percentages, reported as mean $\pm$ standard error across three seeds.

Split and threshold	OOD Acc.	NN_τ	NVH_τ
Attribute, $0.5\tau_0$	31.2 ± 5.1	0.0 ± 0.0	31.2 ± 5.1
Attribute, τ_0	31.2 ± 5.1	7.7 ± 6.4	31.2 ± 5.1
Attribute, $2\tau_0$	31.2 ± 5.1	28.0 ± 6.4	31.2 ± 5.1
Attribute, $4\tau_0$	31.2 ± 5.1	42.3 ± 2.8	29.8 ± 4.3
Geometry, $0.5\tau_0$	2.9 ± 0.7	0.0 ± 0.0	2.9 ± 0.7
Geometry, τ_0	2.9 ± 0.7	14.8 ± 5.5	2.9 ± 0.7
Geometry, $2\tau_0$	2.9 ± 0.7	49.5 ± 7.8	2.9 ± 0.7
Geometry, $4\tau_0$	2.9 ± 0.7	67.2 ± 4.3	2.9 ± 0.7

Here, $\tau_0 = 0.003$. The all-sample nearest-neighbor rate increases with τ, whereas NVH remains stable for geometry and changes only slightly for attribute composition at the largest threshold.

C. Complete Few-Shot Results

Table 5. Few-shot OOD tuple accuracy as mean $\pm$ standard error across three seeds.

Split	Initialization	$m = 1$	$m = 2$	$m = 4$	$m = 8$
Attribute	Pretrained	54.8 ± 5.4	25.0 ± 4.7	19.0 ± 3.0	22.3 ± 3.6
Attribute	Scratch	5.1 ± 1.1	2.7 ± 0.9	4.5 ± 2.2	4.5 ± 0.5
Geometry	Pretrained	60.7 ± 1.6	31.0 ± 2.6	19.0 ± 0.9	15.6 ± 3.6
Geometry	Scratch	3.4 ± 1.4	4.9 ± 0.9	4.7 ± 1.4	4.7 ± 1.2

D. Exploratory Color-Count Ablation

We vary the number of colors in the attribute-composition task while approximately preserving the proportion of held-out shape–color pairs. This experiment uses one seed and is intended only as an exploratory difficulty check.

Table 6. Single-seed attribute-composition ablation. Values are tuple-accuracy percentages.

Number of colors	ID Acc.	OOD Acc.
2	60.9	27.5
3	46.3	9.4
4	66.4	33.9

The results are non-monotonic and therefore do not establish a simple scaling relation between the number of colors and compositional difficulty. Additional seeds and independently sampled supports are required for such a claim.