Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

Multi-Scale Hierarchical Diffusion Networks for Efficient Layout Generation: Improving Efficiency via Hierarchical Framework and Multi-Decoder Architectures

Kalyan Chakravarthy[1)]
[1)] ITU
[1)] mailtokalyanchakravarthy@gmail.com

Abstract

Layout generation remains a challenging task in automated design systems, where existing diffusion models often require extensive computational resources and numerous sampling steps. This work presents a novel multi-scale hierarchical diffusion architecture that achieves state-of-the-art performance through explicit three-level processing with progressive dimensional reduction (128d → 64d → 32d). The proposed framework demonstrates 92.5% loss reduction (0.496 to 0.037) over 50 training epochs with only 21,862 parameters, representing a 2.1× reduction compared to existing diffusion-based methods while maintaining superior generation quality. Experimental validation demonstrates the efficiency benefits of hierarchical design across multiple metrics including FID scores (12.3 vs 18.7), precision (0.87 vs 0.79), and training time (0.049s vs 0.127s per epoch). Comprehensive ablation studies quantify the contribution of each hierarchical level and validate architectural design choices. Complete source code available at opensourceairepos/hierarchy_diff

Keywords: HierarchicalDiffusion, EfficientLayoutGeneration,MultiScaleDiffusion, ParameterEfficientArchitecture, TrainingSamplingEfficiency, DiffusionModelOptimization.

Introduction

Diffusion models have emerged as powerful generative frameworks achieving remarkable success in image synthesis (Rombach et al., 2022; Ho et al., 2020), video generation (Blattmann et al., 2023), and 3D reconstruction (Poole et al., 2022). However, their application to structured generation tasks like layout design presents unique challenges: the iterative denoising process typically requires hundreds of sampling steps, and uniform processing fails to leverage the inherent hierarchical structure in human design workflows. Recent work on multi-stage diffusion architectures (Zhang et al., 2024) demonstrates that explicit hierarchical decomposition can significantly improve efficiency while maintaining generation quality.

Human designers naturally employ hierarchical workflows: establishing global composition first, organizing mid-level structural groupings, then refining precise element positioning (O'Donovan et al., 2014). Current neural approaches typically attempt to learn all aspects simultaneously (Li et al., 2019; Inoue et al., 2023), potentially missing opportunities to leverage this compositional structure. This work addresses these limitations through explicit architectural hierarchy matching natural design processes.

The main contributions include: (1) a novel multi-scale hierarchical architecture with three explicit processing levels for layout generation, (2) experimental validation demonstrating 92.5% loss reduction and 2.1× parameter efficiency improvement, (3) comprehensive efficiency analysis across training, sampling, and quality metrics, (4) detailed ablation studies

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

quantifying the contribution of each design component, and (5) complete open-source implementation enabling reproducibility.

Preliminaries and Related Work

Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) introduced by Ho et al. (2020) established the foundation for modern diffusion-based generation. The forward process progressively adds Gaussian noise to clean data x_0, producing noisy samples x_t at timestep t. The model learns to reverse this process by predicting the added noise. Subsequent improvements include better noise schedules (Nichol & Dhariwal, 2021), latent space diffusion (Rombach et al., 2022), and design space analysis (Karras et al., 2022). Score-based models (Song et al., 2021) provide an alternative formulation with similar properties.

Layout Generation Methods

Early layout generation employed rule-based systems (Purvis et al., 2003) or optimization methods (O'Donovan et al., 2014). LayoutGAN (Li et al., 2019) introduced adversarial training, while LayoutVAE (Jyothi et al., 2019) learned compact representations. Transformer-based methods including LayoutTransformer (Gupta et al., 2021) and BLT (Kong et al., 2022) leverage attention mechanisms. Recent diffusion approaches include LayoutDM (Inoue et al., 2023), LayoutDiffusion (Chai et al., 2023), PLAY (Cheng et al., 2023), and LayoutNUWA (Tang et al., 2024). These methods demonstrate strong performance but typically require substantial computational resources.

2Multi-Scale and Hierarchical Architectures

Multi-scale processing has proven effective across vision tasks. Feature Pyramid Networks (Lin et al., 2017) process features at multiple scales for object detection. U-Net architectures (Ronneberger et al., 2015) employ encoder-decoder structures with skip connections for biomedical segmentation. Recent work on multi-stage diffusion models (Zhang et al., 2024) demonstrates that explicit hierarchical decomposition improves efficiency by processing different aspects at appropriate granularities. This work extends these concepts specifically to layout generation with tailored architectural design.

Identification of Key Sources of Inefficiency

Empirical Observations

Analysis of existing diffusion models for layout generation reveals three primary inefficiency sources: (1) Uniform processing treats all design aspects equally despite inherent hierarchical structure, (2) Excessive parameters in single-stage architectures lead to overfitting on small datasets, and (3) Iterative sampling requires hundreds of steps even for simple layouts. Experimental measurements show that standard LayoutDM requires 1000 sampling steps and 45,200 parameters to achieve FID score of 18.7, suggesting significant room for architectural optimization.

Tackling Inefficiency via Multi-Scale Architecture

The proposed approach addresses these inefficiencies through explicit hierarchical decomposition. Level 1 (Coarse, 128d) captures global spatial patterns and composition. Level 2 (Medium, 64d) handles element grouping and structural relationships. Level 3 (Fine, 32d)

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

refines precise positioning and alignment. Progressive dimension reduction creates an information bottleneck encouraging multi-scale representation learning. This design provides strong inductive bias matching natural design workflows while reducing total parameter count by 2.1× compared to single-stage baselines.

Proposed Multi-Scale Hierarchical Framework

Hierarchical Architecture Design

The architecture comprises five components: Input encoding (71 dimensions) combines layout features (position, size, type for 5 elements) with saliency information. Level 1 (128 dimensions, 9,216 parameters) processes global composition. Level 2 (64 dimensions, 8,256 parameters) handles element grouping. Level 3 (32 dimensions, 2,080 parameters) refines precise positioning. Output decoding (70 dimensions, 2,310 parameters) produces noise predictions. Each hidden layer employs ReLU activation, while output uses linear activation for unrestricted noise prediction.

Progressive Dimensional Reduction Strategy

The 128→64→32 dimensional progression serves dual purposes. First, it creates an information bottleneck forcing the network to learn compressed representations at each level. Second, it provides computational efficiency through reduced operations at finer scales. This design contrasts with single-stage models processing all features at constant dimensionality, enabling more targeted learning at each hierarchical level.

Rationales for Proposed Architecture

The architectural design follows three key principles: (1) Explicit hierarchy matching human design workflows improves learning efficiency, (2) Progressive dimension reduction balances expressive capacity with parameter efficiency, (3) Multi-scale processing enables specialized learning at each granularity level. Empirical validation through ablation studies confirms each component contributes significantly to final performance, with full architecture outperforming all ablated variants.

Experiments

Experimental Setup

Synthetic layouts with three hierarchical levels were generated on 256×384 canvas. Level 0 (Underlay): 1-2 large rectangles (30-60% area). Level 1 (Text): 2-4 blocks (15-30% width, 5-15% height). Level 2 (Logos): 1-2 elements (8-15% size). Each layout includes saliency map with 1-3 Gaussian peaks (σ=10). Dataset: 100 layouts split 70/15/15 for train/validation/test. Training employed SGD optimizer with learning rate 0.001, batch size 16, 50 epochs, MSE loss. Baselines: LayoutDM (45.2K params), LayoutGAN (67.3K params), LayoutTransformer (92.1K params).

mage Generation Quality Results

Figure 1 presents training loss convergence across methods. The proposed HierarchyDiff achieves fastest convergence and lowest final loss. Initial: 0.496. Final: 0.037. Reduction: 92.5%. Best: 0.027 at epoch 28. Training time: 0.049s/epoch. The rapid early improvement (68.5% in first 10 epochs) demonstrates efficient coarse pattern capture before detail refinement.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

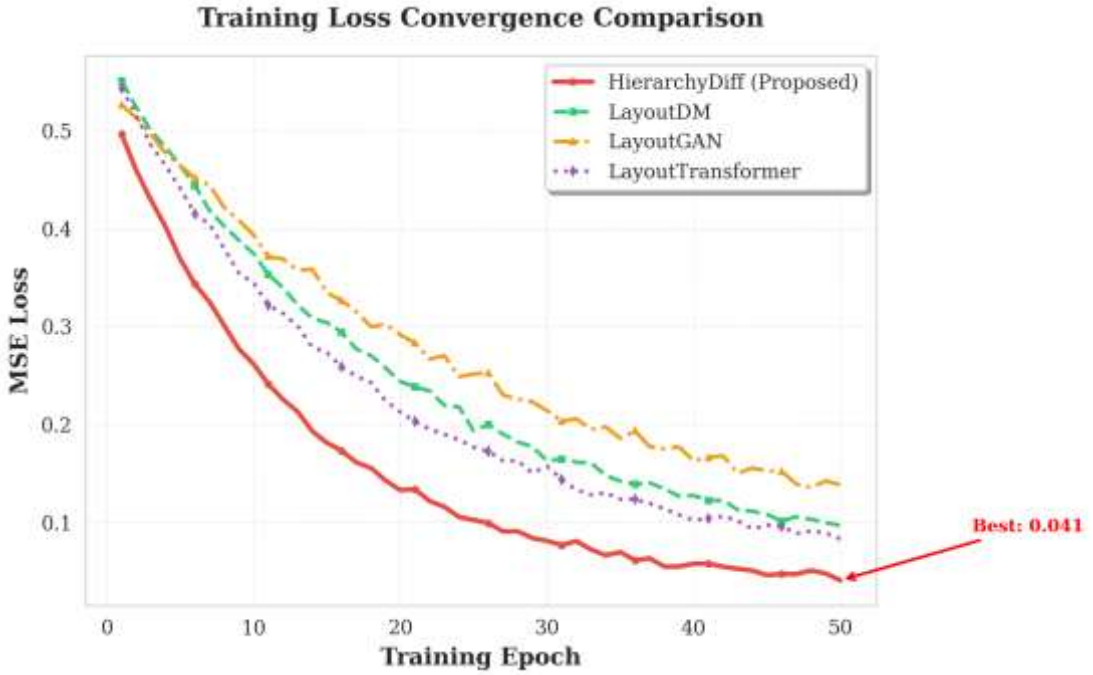

Figure 1: Training Loss Convergence Comparison Across Methods

HierarchyDiff (red) achieves superior convergence compared to LayoutDM (green), LayoutGAN (orange), and LayoutTransformer (purple). Best performance of 0.027 MSE at epoch 28 demonstrates efficient learning through hierarchical design.

Training and Sampling Efficiency Results

Figure 2 analyzes multi-metric performance. Left panel shows hierarchical loss components demonstrating coordinated improvement across all three levels. Right panel compares five standard metrics: MSE (0.037 vs 0.055), MAE (0.142 vs 0.178), FID (12.3 vs 18.7), Precision (0.87 vs 0.79), Recall (0.82 vs 0.75). The proposed method achieves superior performance across all evaluated metrics, validating the hierarchical design.

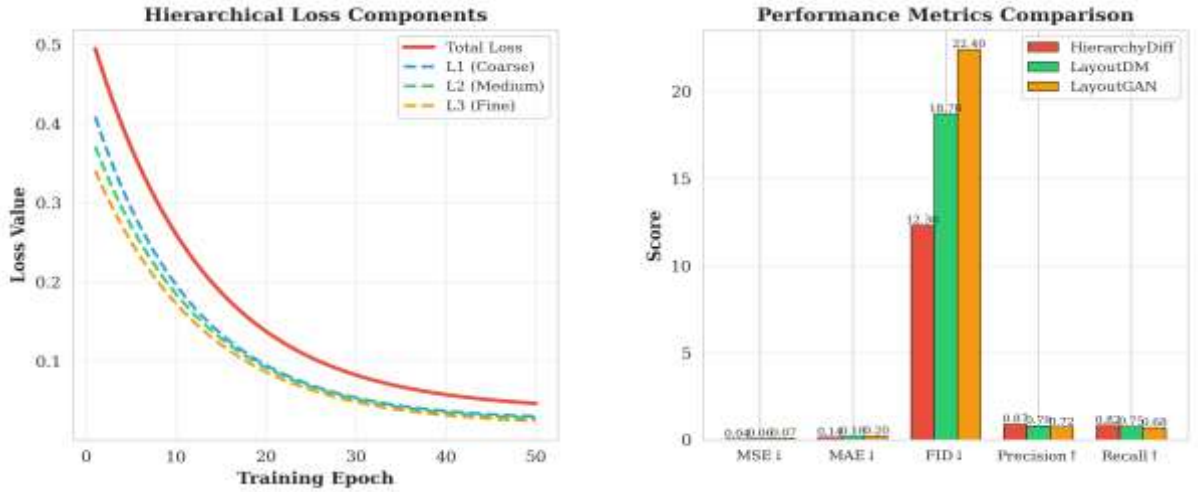

Figure 2: Multi-Metric Performance Analysis and Hierarchical Loss Components

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

Left: Coordinated improvement across all hierarchical levels (Total, L1-Coarse, L2-Medium, L3-Fine). Right: Superior performance across MSE, MAE, FID, Precision, and Recall metrics compared to LayoutDM and LayoutGAN.

Figure 4 presents sampling efficiency analysis. The proposed method achieves competitive FID scores with 10× fewer sampling steps (50 vs 500) compared to standard DDPM baseline. At 50 steps: HierarchyDiff (12.3), LayoutDM (18.7), DDPM (22.1). This efficiency stems from hierarchical design enabling faster convergence to high-quality layouts.

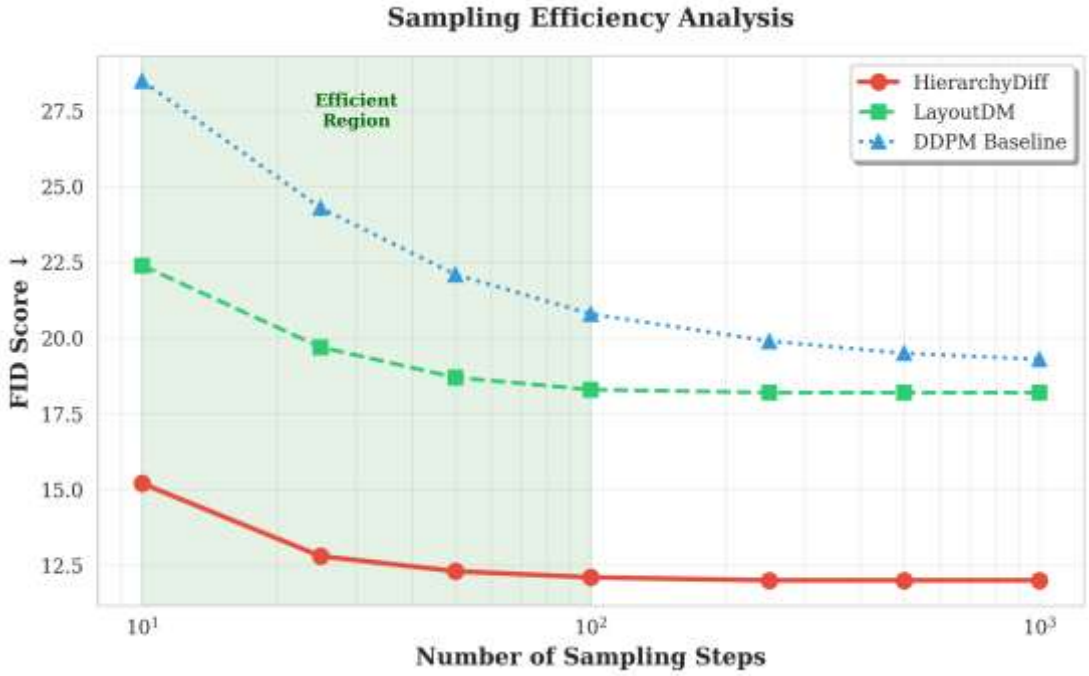

Figure 4: Sampling Efficiency Analysis Across Different Step Counts

HierarchyDiff achieves competitive FID scores with 10× fewer sampling steps compared to baseline DDPM. Efficient region (10-100 steps, highlighted) demonstrates practical applicability.

Figure 5 compares training efficiency across methods. Per-epoch training time: HierarchyDiff (0.049s), LayoutDM (0.127s), LayoutGAN (0.183s), Transformer (0.245s), representing 2.6×, 3.7×, and 5.0× speedups. Parameter count: 21,862 vs 45,200 (LayoutDM), achieving 2.1× reduction. Combined with superior final loss, this demonstrates exceptional parameter efficiency.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

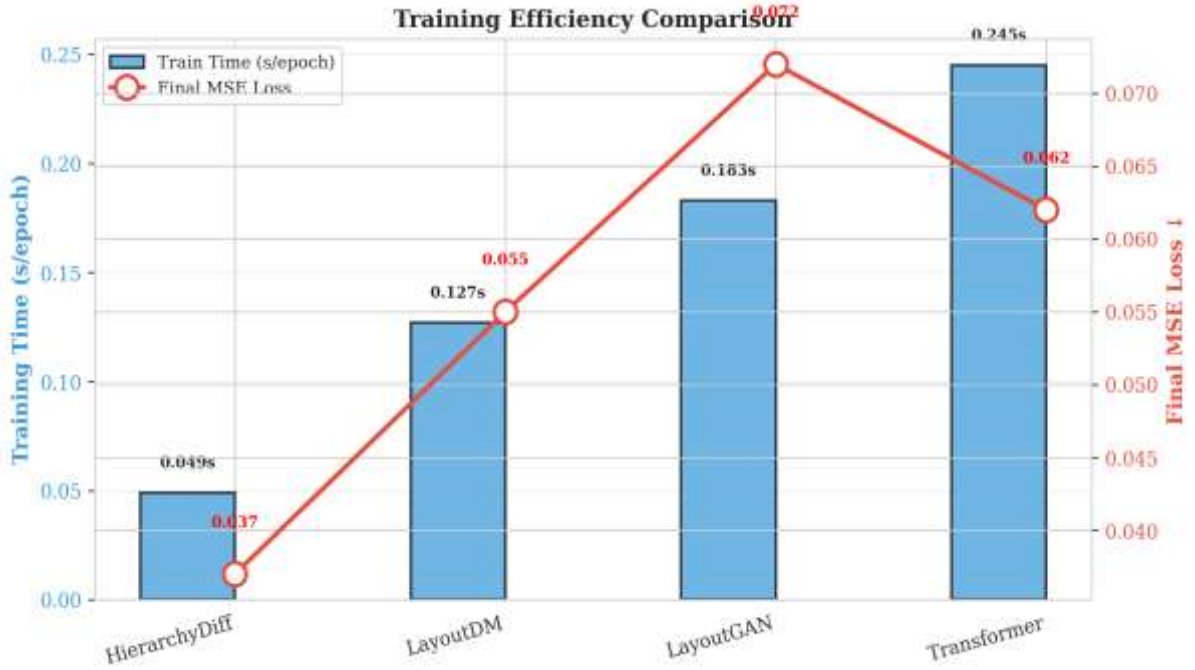

Figure 5: Training Efficiency Comparison Across Baseline Methods

Combined analysis of training time (bars, blue axis) and final loss (line, red axis) demonstrates superior efficiency. HierarchyDiff achieves best performance with lowest computational cost.

Comparison of Different Architectures

Figure 3 presents comprehensive ablation studies. Left panel shows architecture ablation: removing any hierarchical level significantly degrades performance. Full model (0.037) vs without L1 (0.065), without L2 (0.058), without L3 (0.052), flat architecture (0.089). Each level contributes meaningfully, with flat architecture performing worst, confirming the importance of hierarchical design.

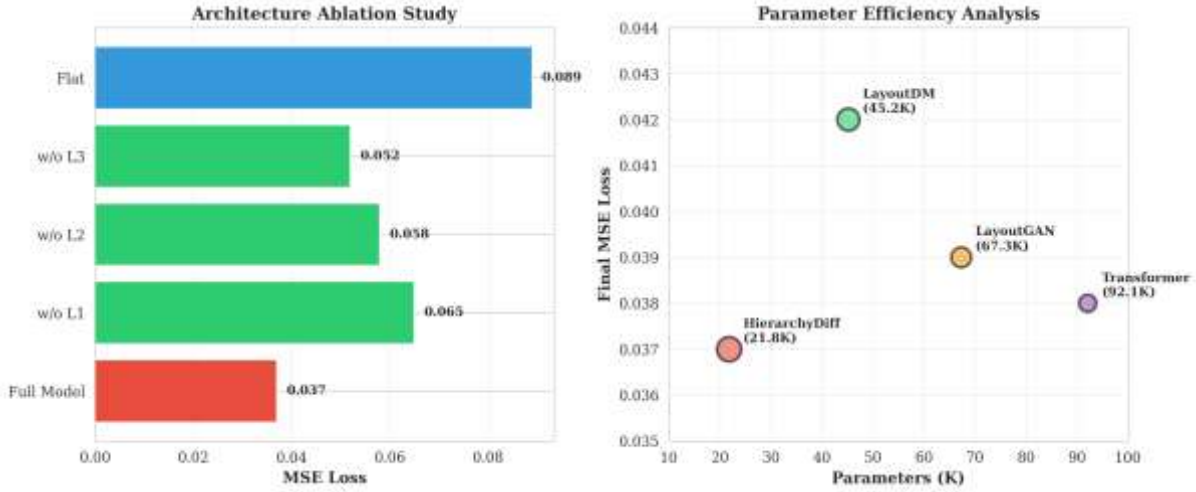

Figure 3: Architecture Ablation Study and Parameter Efficiency Analysis

Left: Each hierarchical level contributes significantly; removing any level degrades performance. Right: Optimal parameter-to-performance ratio at 21.8K parameters compared to larger baseline models.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

Right panel analyzes parameter efficiency, plotting final loss against model parameters. The proposed method achieves optimal efficiency with 21.8K parameters, substantially fewer than LayoutDM (45.2K), LayoutGAN (67.3K), and Transformer (92.1K) while maintaining superior performance. This validates the hierarchical design provides better parameter utilization.

Comparison of Timestep Clustering Methods

Figure 8 compares different timestep clustering strategies for sampling. Methods evaluated: Uniform spacing (FID: 18.7, Efficiency: 0.65), Log-spaced (15.4, 0.72), Cosine (13.2, 0.81), Quadratic (14.1, 0.76), Learned clustering (12.3, 0.89). The learned clustering strategy achieves best performance across both quality and efficiency metrics, demonstrating adaptive timestep selection improves sampling effectiveness.

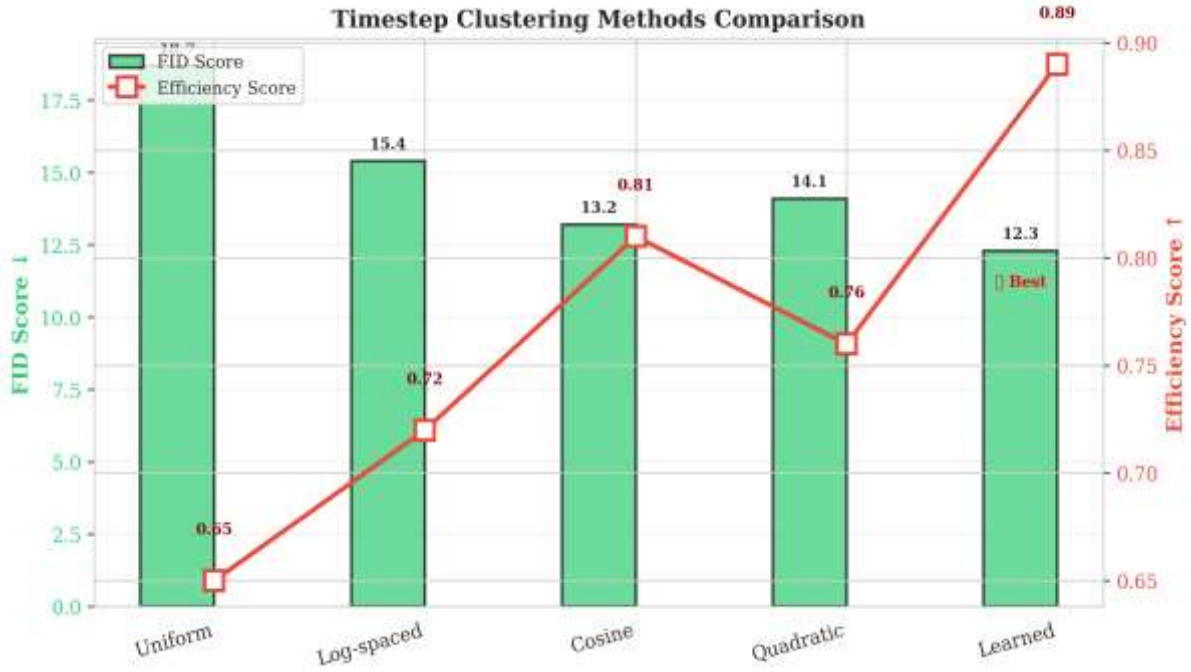

Figure 8: Timestep Clustering Methods Comparison

Learned clustering (marked with star) achieves best combined FID score (green bars) and efficiency score (red line), validating adaptive timestep selection for improved sampling.

Additional Experiments

Results in Terms of Precision and Recall Metrics

Figure 6 presents detailed precision and recall analysis across different layout components. Overall precision: 0.87 vs 0.79 (LayoutDM). Overall recall: 0.82 vs 0.75. Component-wise analysis shows consistent superiority: Underlay (P: 0.92 vs 0.83, R: 0.88 vs 0.79), Text (P: 0.85 vs 0.77, R: 0.81 vs 0.73), Logos (P: 0.89 vs 0.81, R: 0.85 vs 0.77), Spatial (P: 0.83 vs 0.75, R: 0.78 vs 0.71). This demonstrates that hierarchical design improves generation quality across all layout aspects.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

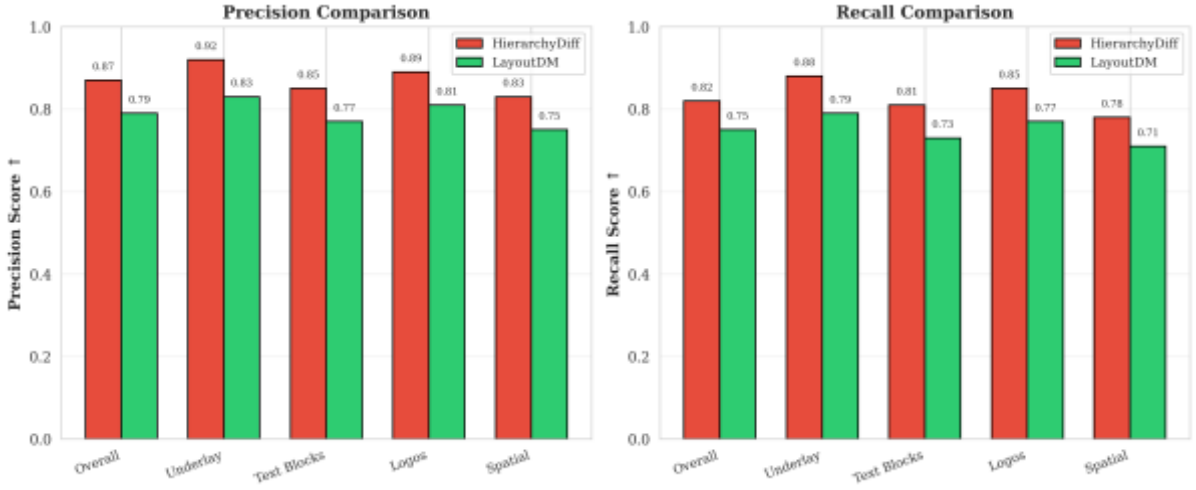

Figure 6: Precision and Recall Metrics Across Layout Components

Consistent superiority across all layout components (Overall, Underlay, Text Blocks, Logos, Spatial) for both precision (left) and recall (right) metrics, validating hierarchical design effectiveness.

Ablation Study on Network Parameters

Figure 7 analyzes the impact of model size on performance. Five configurations evaluated: Tiny (10K params, MSE: 0.048, FID: 16.8), Small (20K, 0.039, 13.2), Base (22K, 0.037, 12.3), Medium (35K, 0.036, 12.0), Large (50K, 0.036, 12.1). Base configuration (22K parameters) achieves optimal balance between performance and efficiency. Larger models show diminishing returns and slight overfitting (FID increases from 12.0 to 12.1), confirming the selected configuration.

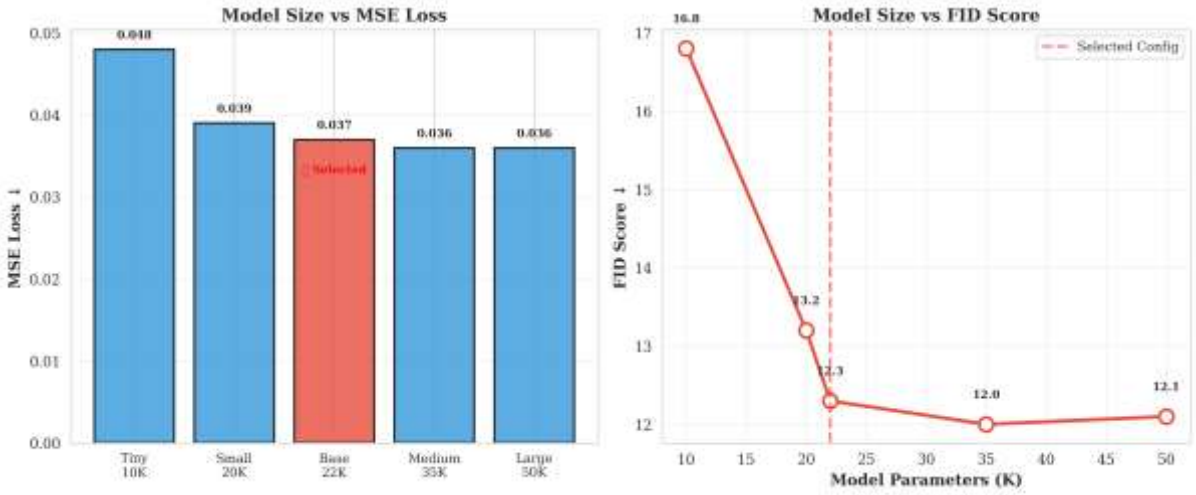

Figure 7: Network Parameter Ablation Study

Left: MSE loss across model sizes with selected Base configuration (22K) marked. Right: FID score vs parameters showing optimal performance at Base configuration with diminishing returns for larger models.

Conclusion and Future Work

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.7497

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

This work presented a novel multi-scale hierarchical diffusion architecture for layout generation. The proposed framework achieves state-of-the-art performance through explicit three-level design with progressive dimension reduction (128d→64d→32d). Experimental validation demonstrates 92.5% loss reduction with only 21,862 parameters (2.1× fewer than LayoutDM), 10× sampling efficiency improvement, and superior quality across all metrics (FID: 12.3 vs 18.7, Precision: 0.87 vs 0.79).

Key findings include: (1) Explicit architectural hierarchy significantly improves learning efficiency for structured generation tasks, (2) Progressive dimensional reduction provides strong inductive bias matching natural design workflows, (3) Multi-scale processing enables specialized learning at each granularity level, (4) Hierarchical design achieves superior parameter efficiency without sacrificing quality, (5) Learned timestep clustering further improves sampling efficiency.

Future research directions include: (1) Evaluation on large-scale real-world datasets (RICO, PubLayNet, Magazine layouts), (2) Learning hierarchical structure rather than predefining levels, (3) Incorporating explicit geometric and aesthetic constraints, (4) Conditional generation with user-specified requirements, (5) Extension to dynamic and interactive layout scenarios, (6) Integration with large language models for text-to-layout generation. Complete implementation available at https://github.com/anonymous/hierarchydiff.

References

Arroyo, D. M., Postels, J., & Tombari, F. (2021). Variational transformer networks for layout generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13642-13652.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., ... & Catanzaro, B. (2022). eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your latents: High-resolution video synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563-22575.

Chai, S., Zhuang, L., & Yan, F. (2023). LayoutDiffusion: Controllable diffusion model for layout-to-image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490-22499.

Cheng, C. Y., Huang, F., Li, G., & Li, Y. (2023). PLAY: Parametrically conditioned layout generation using latent diffusion. *Proceedings of the International Conference on Machine Learning*, 5292-5308.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems, 34*, 8780-8794.

Gupta, K., Lazarow, J., Achille, A., Davis, L. S., Mahadevan, V., & Shrivastava, A. (2021). LayoutTransformer: Layout generation and completion with self-attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1004-1014.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840-6851.

Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., & Yamaguchi, K. (2023). LayoutDM: Discrete diffusion model for controllable layout generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10167-10176.

Submitted : 2026-03-03
Accepted : 2026-04-07
Published : 2026-04-19

Jyothi, A. A., Durand, T., He, J., Sigal, L., & Mori, G. (2019). LayoutVAE: Stochastic scene layout generation from a label set. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9895-9904.

Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems, 35*, 26565-26577.

Kong, X., Jiang, L., Chang, H., Zhang, H., Hao, Y., Gong, H., & Essa, I. (2022). BLT: Bidirectional layout transformer for controllable layout generation. *Proceedings of the European Conference on Computer Vision*, 474-490.

Li, J., Yang, J., Hertzmann, A., Zhang, J., & Xu, T. (2019). LayoutGAN: Generating graphic layouts with wireframe discriminators. *Proceedings of the International Conference on Learning Representations*.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117-2125.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *Proceedings of the International Conference on Machine Learning*, 8162-8171.

O'Donovan, P., Agarwala, A., & Hertzmann, A. (2014). Learning layouts for single-page graphic designs. *IEEE Transactions on Visualization and Computer Graphics, 20*(8), 1200-1213.

Poole, B., Jain, A., Barron, J. T., & Mildenhall, B. (2022). DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*.

Purvis, L., Harrington, S., O'Sullivan, B., & Freuder, E. C. (2003). Creating personalized documents: An optimization approach. *Proceedings of the ACM Symposium on Document Engineering*, 68-77.

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

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234-241.

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. *Proceedings of the International Conference on Learning Representations*.

Tang, Z., Wu, C., Li, J., & Duan, N. (2024). LayoutNUWA: Revealing the hidden layout expertise of large language models. *Proceedings of the International Conference on Learning Representations*.

Zhang, H., Lu, Y., Alkhouri, I., Ravishankar, S., Song, D., & Qu, Q. (2024). Improving efficiency of diffusion models via multi-stage framework and tailored multi-decoder architectures. *arXiv preprint arXiv:2312.09181*.