Balancing Fidelity and Efficiency for Efficient Automatic Text-to-3D Asset Generation Pipelines

Nayeon Kim
Media Research Division
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
nyeon@etri.re.kr

Won-Joo Park*
Media Research Division
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
wjpark@etri.re.kr

***Abstract*—Text-to-3D generation has become increasingly important in content creation industries for rapid previsualization workflows. While recent advances in models like TRELLIS have demonstrated remarkable capabilities, the impact of Large Language Model (LLM)-based prompt augmentation on 3D asset quality remains underexplored. This paper presents a comprehensive analysis of how different LLM architectures and model sizes affect 3D generation quality in TRELLIS pipelines. We evaluate multiple LLM variants across Gemma3, Qwen3, Llama3.1, and DeepSeek-R1 architectures using FD_{DINOv2} and dual CLIP metrics. Our findings reveal that architectural compatibility is more critical than model size, with performance variations within size categories often exceeding variations between categories. We demonstrate that smaller LLMs can achieve comparable quality to larger models while offering significant computational cost savings. These results provide practical guidance for industrial deployment where resource efficiency is crucial for large-scale 3D asset generation pipelines.**

***Index Terms*—3D Generation, Text-to-3D, LLM, Prompt Augmentation, Previsualization**

I. Introduction

In contemporary digital content creation pipelines across film, television, advertising, and gaming industries, previsualization serves as a critical workflow wherein creative professionals rapidly transform scripts, storyboards, and sketches into three-dimensional visual representations. Within this context, asset quality is evaluated primarily by how well generated content preserves the creator's original intent rather than photorealistic fidelity or high-resolution texturing.

Automated industrial workflows frequently encounter brief, underspecified textual prompts as inputs. While experienced practitioners can leverage contextual knowledge to interpret such limited descriptions, automated 3D generation systems lack this capability. This creates significant opportunities for Large Language Model (LLM)-based prompt augmentation techniques, which have demonstrated substantial efficacy in text-to-image generation domains. However, systematic investigation of prompt enhancement effects on text-to-3D asset generation remains limited.

Existing research has predominantly focused on architectural improvements for text-to-3D generation models themselves (e.g., TRELLIS, DreamFusion). While these contributions have advanced the field significantly, there exists a notable gap in understanding how LLM-driven prompt augmentation influences actual 3D asset quality metrics. Specifically, the relationship between model scale parameters and their effects on computational efficiency and output fidelity in industrial deployment scenarios has not been comprehensively characterized.

This investigation employs a systematic comparative analysis framework utilizing the TRELLIS-based text-to-3D generation pipeline. Our experimental design encompasses four distinct LLM families—Gemma3, Qwen3, Llama3.1, and DeepSeek-R1—evaluated across multiple parameter scales.

We employ two primary metrics: Fréchet Distance using DINOv2 features (FD_{DINOv2}) for visual fidelity assessment, and dual CLIP scores distinguishing between Augmented CLIP (alignment with LLM-enhanced prompts) and User CLIP (alignment with original prompts). This approach enables assessment of whether improved augmented prompt alignment translates to better preservation of original semantic intent.

Our key contributions include:

- **Architectural Compatibility Over Scale:** Through systematic evaluation of 12 LLM variants, we demonstrate that architectural compatibility with TRELLIS is more critical than model size, with performance variations within size categories often exceeding variations between categories.
- **LLM-3D Generator Mismatch Analysis:** We identify and analyze cases where high prompt augmentation quality (measured by CLIP improvement) does not translate to better 3D generation fidelity, revealing the importance of LLM-generator architectural alignment over simple augmentation capability.
- **Cost-Effective Industrial Deployment Guidelines:** Our comprehensive analysis provides the first empirical evidence that smaller LLMs (0.6B-8B parameters) can achieve comparable 3D generation quality to larger models while offering significant computational cost savings for production pipelines.

These findings enable cost-effective implementation strategies for production environments, demonstrating that smaller language models deliver stable augmentation quality while offering significant computational savings for TRELLIS deployments requiring $\geq$ 16GB GPU memory.

*Corresponding author

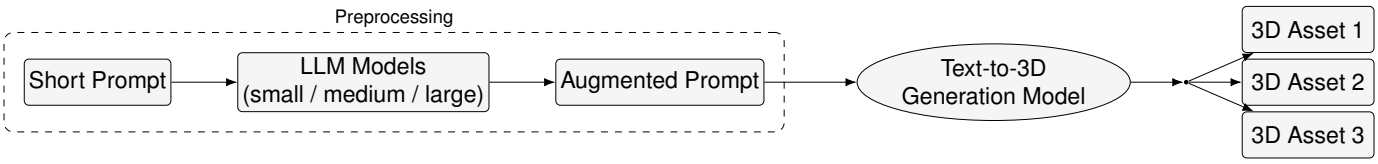

Fig. 1: Automatic Text-to-3D Asset Generation Pipeline from short prompt to generated 3D assets.

II. Related Work

A. Prompt Augmentation in Text-to-Image Generation

In the text-to-image (T2I) domain, prior studies have consistently shown that prompt quality plays a crucial role in aligning generated outputs with user intent. Richer and more descriptive prompts often lead to more visually coherent and aesthetically pleasing results compared to shorter, underspecified inputs [1], [2]. Building on this observation, several works have proposed LLM-based prompt augmentation, where pre-trained or fine-tuned large language models are used to expand or rewrite prompts. These methods enhance descriptions with additional attributes such as style, context, or scene details, thereby improving image quality and faithfulness to the intended semantics [3], [4]. Such findings validate that LLM-augmented prompts can significantly improve generation quality in the T2I setting, motivating their application to more complex modalities.

B. LLM Scale and Industrial Efficiency Considerations

Another dimension that has received little attention is the trade-off between LLM size and efficiency in prompt augmentation workflows. In principle, larger LLMs may offer more nuanced and detailed prompt expansions, but their high computational cost poses challenges for industrial deployment. In practice, researchers and practitioners often resort to smaller open-source LLMs to balance performance with efficiency. For instance, Yeh et al. [5] introduce a lightweight text-to-image prompt optimizer (TIPO) that deliberately avoids large proprietary LLMs, arguing that small-scale open-source models provide sufficient improvements with negligible runtime overhead relative to the generation process. Nevertheless, a systematic comparison of LLM size (small vs. large) and architecture differences in prompt augmentation remains absent, particularly in the context of text-to-3D pipelines. Thus, the relationship between LLM scale, augmentation effectiveness, and industrial efficiency is still an open and underexplored area.

III. Methodology

A. Experimental Setup

We systematically evaluate the impact of LLM-based prompt augmentation on TRELLIS text-to-3D generation across multiple model architectures and scales. Our framework employs the TRELLIS-text-large model (1.1B parameters) trained on 500K+ 3D objects from Objaverse(XL) [6], ABO [7], 3D-FUTURE [8], and HSSD [9], with GPT-4o captioning.

B. LLM Configuration and Prompt Augmentation

We evaluate 12 LLM variants across four architectural families: Gemma3 (12B, 27B-it-q8_0), Qwen3 (0.6B, 14B, 32B-q8_0), Llama3.1 (8B, 70B-instruct-q4_0), DeepSeek-R1 (1.5B, 14B, 32B variants), and GPT-OSS (20B). This spans small (0.6B–1.5B), medium (8B–20B), and large (27B–70B) parameter scales.

Each LLM augments short user prompts ($\leq$6 words) into detailed captions (40 words maximum), following the original TRELLIS protocol. Short prompts are selected from GPT-4o generated descriptions in the original TRELLIS work.

C. Evaluation Metrics

Fidelity Assessment: We use FD_{DINOv2} as the primary metric, measuring distributional similarity between generated and reference 3D assets using DINOv2 features [10]. Lower scores indicate better performance.

Alignment Evaluation: We employ dual CLIP scores (scaled $\times$100): (1) *Augmented CLIP* measures similarity between LLM-augmented prompts and generated assets, (2) *User CLIP* evaluates preservation of original user intent. Higher scores indicate better alignment.

D. Dataset and Implementation

Experiments use 100 randomly sampled instances from Toys4k [11], which was excluded from TRELLIS training data. All experiments run on four NVIDIA RTX 4090 GPUs using default TRELLIS inference settings. Generated assets are rendered with standardized camera parameters for consistent evaluation.

IV. Experimental Results

A. Overall Performance Analysis

Our evaluation across all model configurations demonstrates consistent performance characteristics in both prompt alignment and generation fidelity metrics. Table I presents comprehensive results across 12 LLM variants spanning small (0.6B-8B), medium (12B-20B), and large (27B-70B) parameter scales.

The overall results show a mean FD_{DINOv2} score of 545.27 with a standard deviation of 238.10, where lower values indicate better visual fidelity. For prompt alignment, we observe mean CLIP scores of 29.80 for augmented prompts and 29.33 for original user prompts, indicating that LLM-based augmentation provides modest but consistent improvements in text-to-3D alignment.

The relatively high standard deviation in FD scores (as shown in the Std FD_{DINOv2} column of Table I) suggests

TABLE I: Evaluation results across model sizes. Metrics include CLIP scores and FD_{DINOv2} statistics.

Model Size	Model	CLIP_{user}	CLIP_{aug}	FD_{DINOv2}	Std FD_{DINOv2}	Min FD_{DINOv2}	Max FD_{DINOv2}
Small	qwen3:0.6b	29.88	30.20	532.86	252.77	60.67	1053.95
Small	gemma3:1b	29.60	29.90	570.51	236.44	32.12	1062.42
Small	llama3.1:8b	29.19	29.64	581.75	240.81	45.26	1072.31
Small	deepseek-r1:1.5b	27.35	29.41	624.98	234.07	43.92	1093.52
Medium	gemma3:12b	30.04	29.96	523.27	235.37	40.19	1147.45
Medium	deepseek-r1:14b-qwen	29.41	29.46	525.82	227.65	64.24	1157.65
Medium	qwen3:14b	29.62	30.20	526.54	226.29	39.24	1016.11
Medium	gpt-oss:20b	29.50	29.57	532.40	229.98	48.70	986.56
Large	gemma3:27b-it-q8_0	29.52	29.54	509.22	217.81	53.67	1080.16
Large	llama3.1:70b-instruct-q4	29.76	30.02	532.37	265.67	58.63	1133.60
Large	deepseek-r1:32b-qwen	29.08	30.06	532.78	241.78	33.69	1067.25
Large	qwen3:32b-q8_0	28.97	29.66	549.88	243.35	41.60	1156.45

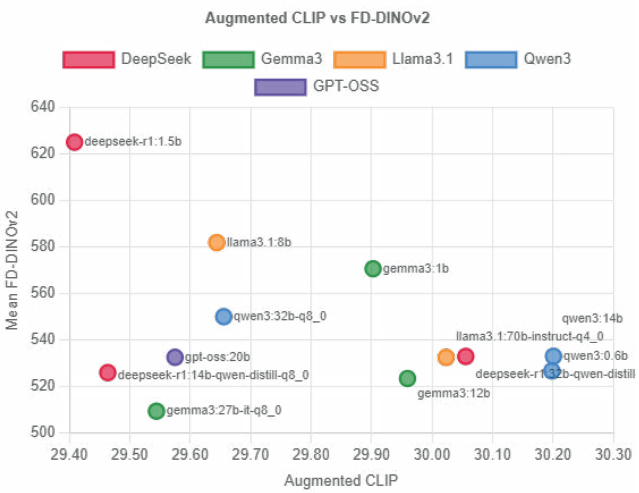

Fig. 2: Scatter plot showing the relationship between Augmented CLIP scores and FD_{DINOv2} values across different LLM model families.

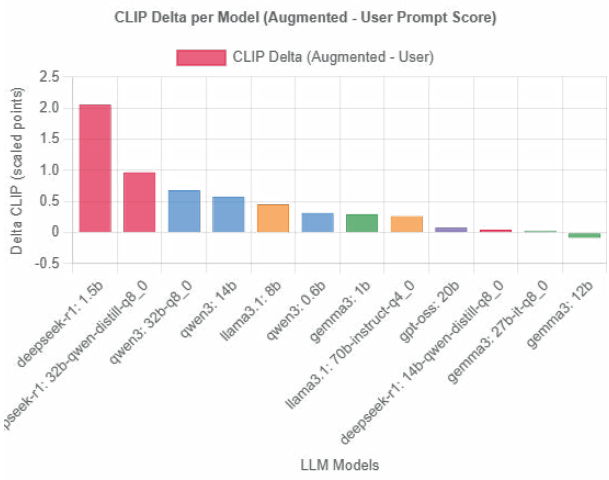

Fig. 3: CLIP Delta per Model (Augmented prompt vs. User prompt).

significant variability in generation quality across different prompt types and content categories, highlighting the complexity of text-to-3D generation tasks and the sensitivity of current evaluation metrics to prompt characteristics.

B. Model Size versus Performance Relationship

Contrary to conventional expectations, our analysis reveals that larger model parameters do not consistently correlate with superior 3D generation performance. As demonstrated in Table I, small models achieve an average FD_{DINOv2} score of 520.02, while medium models perform similarly at 520.00. Notably, large models demonstrate slightly inferior performance with an average score of 526.65.

This unexpected trend suggests that within the context of prompt augmentation for 3D generation, the relationship between model scale and output quality follows a different pattern than typically observed in pure language modeling tasks. The plateau effect observed across model sizes indicates that prompt augmentation quality may be more dependent on the model's ability to understand spatial and visual concepts rather than raw parameter count.

Examining individual model performance in Table I, we observe that the best-performing model overall is gemma3:27b-it-q8_0 (509.22 FD_{DINOv2}), followed by gemma3:12b (523.27), demonstrating that architectural design may be more critical than scale. The considerable variance across model families, as shown by the wide range between minimum and maximum FD values, further supports this hypothesis.

C. Cross-Metric Correlation Analysis

Our correlation analysis reveals important relationships between different evaluation metrics that provide insights into the underlying mechanisms of text-to-3D generation. We observe a strong negative correlation between User CLIP scores and FD_{DINOv2} values ($r \approx -0.80$), indicating that preserving original user intent strongly predicts high-fidelity 3D generation. This relationship suggests that maintaining semantic consistency with the original prompt is more critical than achieving perfect alignment with augmented descriptions.

In contrast, the correlation between Augmented CLIP scores and FD_{DINOv2} values shows a much weaker relationship ($r \approx -0.38$). This finding implies that while augmentation can improve prompt-asset alignment in some cases, excessive elaboration may not necessarily translate to better visual quality in

the final 3D output. Our findings suggest that the **quality of augmentation** (whether the generator can effectively process it) is more important than the **quantity of augmentation**. Verbose augmentation may improve CLIP scores but can deteriorate FD performance if the generator cannot effectively utilize the additional information.

D. Prompt-Generator Mismatch Analysis

Our analysis identifies several cases where prompt augmentation quality does not translate directly to generation performance, revealing important insights about the interaction between LLMs and 3D generation models. A notable case is DeepSeek-R1:1.5B, which shows substantial augmentation improvement (+2.06 CLIP delta) but produces suboptimal generation fidelity (624.98 FD_{DINOv2}), as shown in Table I.

However, closer examination reveals that this apparent contradiction stems primarily from the exceptionally low User CLIP score (27.35) rather than an abnormally high Augmented CLIP score (29.41). This suggests that the model's augmentation may not effectively preserve the original user intent, leading to semantic drift that compromises generation quality. When considering the full distribution statistics (standard deviation: 234.07, minimum: 43.92, maximum: 1093.52), DeepSeek-R1:1.5B does not show extreme outlier behavior across all metrics compared to other models in its size category.

This case represents an exceptional scenario where the relationship between original prompts and augmented results requires broader analysis considering both LLM and TRELLIS architectural compatibility. Such findings highlight the importance of **LLM-generator architectural alignment** rather than simple augmentation capability.

E. Qualitative Evaluation

When evaluating the overall generation results through human visual assessment, we observed that quality differences across model sizes were generally not substantial. However, some smaller models occasionally generated results that diverged significantly from user intentions. To illustrate these findings, we present detailed case studies using boat generation examples.

Our visual assessment reveals that within specific architectural families, model size variations do not significantly impact the fundamental visual quality of generated assets. For example, across DeepSeek variants spanning 1.5B to 32B parameters, all models successfully generate recognizable boat structures that align with basic user requirements. This consistency across scales supports our quantitative findings that model size is not the primary determinant of generation quality.

More revealing are the architectural differences within the same size category, as illustrated in Figure 4. Among small models with comparable parameter counts, we observe dramatic variations in output quality and intent preservation. Notably, Gemma3:1b and Qwen3:0.6b produce results that bear little resemblance to the intended boat concept, generating what appears to be abstract vessel-like structures that fail to capture the specified "blue hull with white railing" characteristics. In contrast, DeepSeek-R1:1.5b and Llama3.1:8b successfully generate recognizable boats with appropriate color schemes and structural elements.

This architectural dependency suggests that the relationship between LLM augmentation and TRELLIS generation quality is more complex than simple parameter scaling. The divergent results within the small model category indicate that either: (1) specific architectural features influence compatibility with the TRELLIS generation pipeline, or (2) certain LLM-generated prompt augmentations create semantic drift that the TRELLIS model cannot effectively process, leading to degraded outputs.

Importantly, when comparing results across different architectures within the small LLM category, we found that some models successfully generated outputs aligned with user intent, making it difficult to attribute these variations solely to model size differences. This architectural dependency suggests that the relationship between model scale and generation quality is more nuanced than simple parameter counting would suggest.

A critical limitation of our evaluation framework became apparent when these qualitatively poor generation results were not adequately filtered by our quantitative quality metrics. For instance, while Gemma3:1b and Qwen3:0.6b produced visually problematic outputs as shown in Figure 4, their quantitative scores (FD_{DINOv2}: 570.51 and 532.86 respectively) do not reflect the severity of the semantic deviation from user intent. This discrepancy between human perceptual assessment and automated evaluation highlights the need for more sophisticated evaluation methodologies in future research.

These qualitative observations reinforce our quantitative findings that architectural compatibility may be more critical than model size for effective prompt augmentation in 3D generation pipelines. The visual evidence suggests that practitioners should prioritize architectural selection and LLM-TRELLIS compatibility assessment over simple parameter-based model choices.

F. Limitations of Quantitative Metrics

We observed cases where quantitative metrics showed poor performance despite visually acceptable results. This can be attributed to:

- CLIP's sensitivity to global concepts (texture, color) while being less sensitive to geometric errors
- FD_{DINOv2}'s focus on global distribution metrics, which may miss localized geometric issues

These limitations underscore the importance of complementing automated metrics with human evaluation, particularly in applications where perceptual quality is paramount for industrial deployment.

G. Industrial Implementation Implications

1) Resource Allocation Considerations: Our findings are particularly relevant for industrial deployment where TRELLIS's inherent memory requirements ($\geq$16GB) make resource

allocation a practical concern. By demonstrating that small-scale LLMs deliver stable augmentation quality, this study guides practitioners in selecting cost-effective solutions without sacrificing previsualization fidelity.

The ability to utilize smaller models for prompt augmentation without significant quality degradation provides valuable cost optimization opportunities for studios and content creators, especially in large-scale production environments where computational efficiency directly impacts operational costs.

2) Cost-Effectiveness Analysis: The minimal performance differences between model sizes, combined with the computational and financial overhead of larger models, suggests that smaller LLMs represent a more efficient choice for production pipelines. This is especially valuable for industries requiring large-scale asset generation with balanced quality and computational efficiency.

Given that commercial LLM APIs incur per-token costs and TRELLIS itself requires substantial GPU memory, the demonstrated efficiency of smaller models enables more sustainable and scalable deployment strategies for content creation workflows across film, television, advertising, gaming, and extended reality applications.

H. Architectural Dependencies and Future Research Directions

Beyond model size, our results suggest that specific LLM architectures may have varying compatibility with 3D generation pipelines. What initially appears to be a "model size effect" may actually represent "architectural compatibility effects," warranting dedicated research into LLM-3D generator architectural alignment.

The considerable performance variation within size categories indicates that architectural features may be more predictive of 3D generation success than parameter count alone. This finding has significant implications for both model selection in industrial applications and future research directions in multimodal AI systems.

V. Conclusion

This study provides the first comprehensive analysis of LLM-based prompt augmentation effects on TRELLIS 3D asset generation quality. Our key findings demonstrate that:

1) **Model size does not consistently correlate with better 3D generation quality**, with small and medium models achieving comparable performance to larger counterparts.
2) **Original user intent preservation** (User CLIP) shows stronger correlation with generation fidelity than augmented prompt consistency ($r \approx -0.80$ vs $r \approx -0.38$).
3) **Architectural compatibility** appears more critical than model size, as evidenced by significant performance variations within size categories and specific cases like DeepSeek-R1 models.
4) **Small-scale LLMs can provide adequate efficiency** for industrial applications, though careful model selection is required to avoid architectures that produce results divergent from user intent.

Limitations: Our evaluation scope was limited to a specific dataset (Toys4k) and single 3D generation model (TRELLIS). Some small models occasionally generated results that significantly deviated from user intentions, highlighting the need for architectural-aware selection rather than size-based decisions alone.

These findings enable practitioners to make informed decisions about LLM selection for 3D generation pipelines, prioritizing architectural compatibility and cost-efficiency over raw model size. Future work should explore diverse domains, multiple 3D generation models, and systematic analysis of architectural features that enhance LLM-3D generator compatibility.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2025-02263036, Generative AI-based Pre-visualization Technology for Media Production Coordination].

References

[1] A. Fernando, K. Nguyen, M. Wang, and X. Li, "Clipscore++: Advances in reference-free evaluation for image captioning and text-to-image generation," in *ACL*, 2023.

[2] J. Hessel, A. Holtzman, M. Forbes, and Y. Choi, "Clipscore: A reference-free evaluation metric for caption evaluation," in *EMNLP*, 2021.

[3] W. Zhou, X. Sun, and K. Chen, "Prompt engineering with large language models for improved text-to-image generation," *arXiv preprint arXiv:2305.12345*, 2023.

[4] Y. Zhang, H. Li, Y. Xu, and D. Zhao, "Automatic prompt generation for text-to-image models using llms," *arXiv preprint arXiv:2306.09876*, 2023.

[5] R. A. Yeh, H. Chang, Y. Xu, X. Wang, and S. Maji, "Tipo: Lightweight prompt optimization for text-to-image generation," in *CVPR*, 2024.

[6] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] J. Collins, S. Goel, K. Deng, A. Luthra, L. Xu, E. Gundogdu, X. Zhang, T. F. Y. Vicente, T. Dideriksen, H. Arora *et al.*, "Abo: Dataset and benchmarks for real-world 3d object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 126–21 136.

[8] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, and D. Tao, "3d-future: 3d furniture shape with texture," *International Journal of Computer Vision*, pp. 1–25, 2021.

[9] M. Khanna, Y. Mao, H. Jiang, S. Haresh, B. Shacklett, D. Batra, A. Clegg, E. Undersander, A. X. Chang, and M. Savva, "Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation," *arXiv preprint arXiv:2306.11290*, 2023.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[11] S. Stojanov, A. Thai, and J. M. Rehg, "Using shape to categorize: Low-shot learning with an explicit shape bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1798–1808.

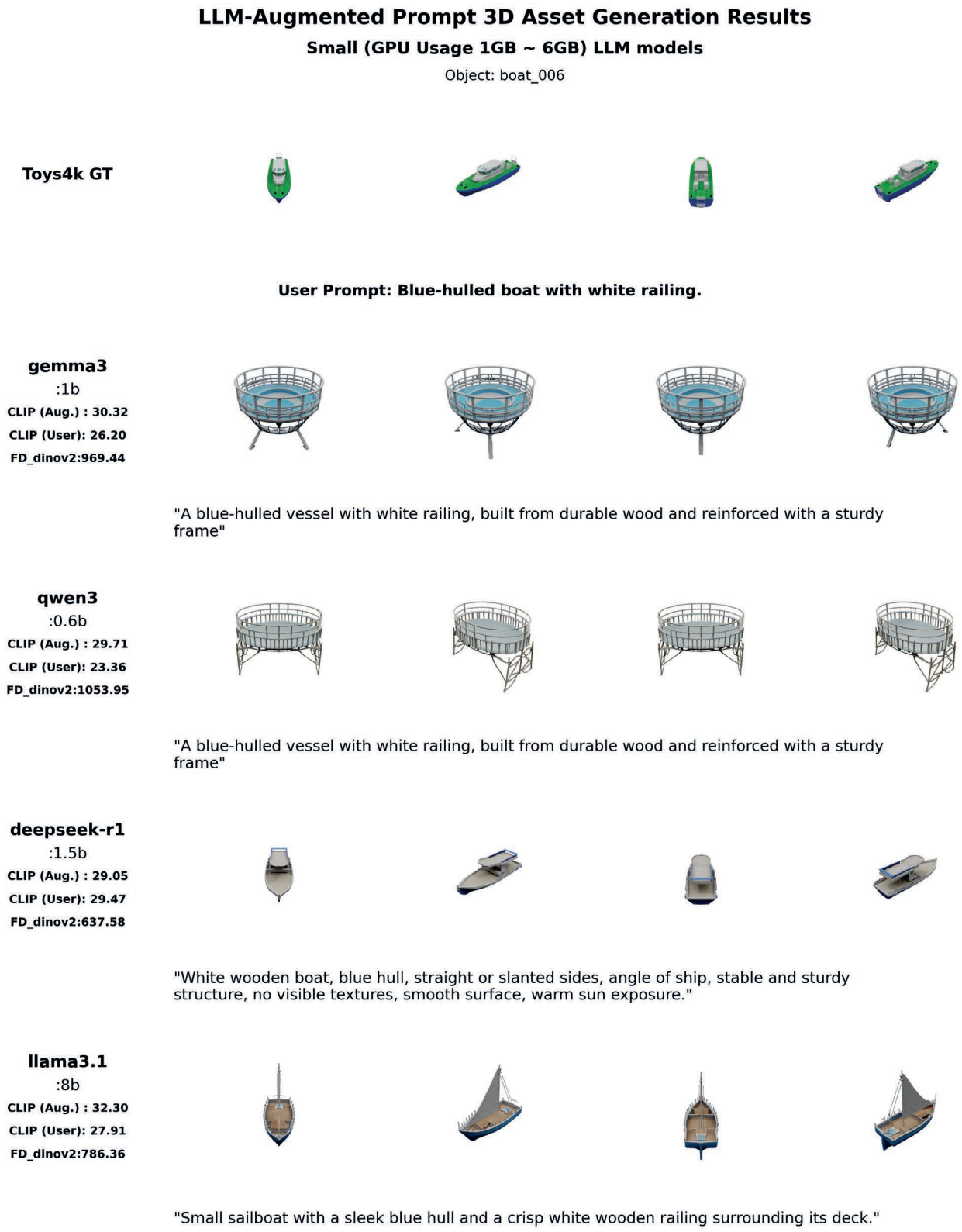

Fig. 4: Architectural variation within small LLM models (1GB-6GB) for boat generation. The top two models (Gemma3:1b and Qwen3:0.6b) generate results that significantly deviate from the intended boat concept, while the bottom two models (DeepSeek-R1:1.5b and Llama3.1:8b) successfully capture the boat structure and color scheme specified in the user prompt.