ForgeDreamer: Industrial Text-to-3D Generation with Multi-Expert LoRA and Cross-View Hypergraph

Supplementary Material

6. Appendix

6.1. Code and Dataset Examples

We provide example implementations of the key components of our method, ForgeDreamer—including the LoRA distillation pipeline and the cross-view hypergraph enhancement module—in the supplementary materials to facilitate reproducibility. In addition, sample data from our multi-view industrial dataset are also included in the supplement to illustrate the data format and preprocessing workflow.

6.2. Limitations of Existing Public 3D Datasets

Although widely used in industrial anomaly detection and 3D perception, existing public datasets such as MVTec 3D-AD and Real-IAD are not well suited for our LoRA distillation framework. MVTec 3D-AD provides RGB images paired with depth or point cloud data; however, these point clouds are often incomplete. When generating a front-view image, the lower portion of the geometry is frequently missing, making it impossible to obtain a consistent and fully visible front-view observation required for LoRA training.

Real-IAD suffers from a different limitation: the dataset supplies only two viewpoints for each object—an oblique 45° view and a strictly top-down view. Since no true front-view images are available, the dataset cannot support the dual-perspective (front and up) supervision that our method relies on to ensure stable and geometry-consistent LoRA adaptation.

To validate these limitations empirically, we conducted experiments using both datasets. As shown in Figure 9a and Figure 9b, models trained on MVTec 3D-AD and Real-IAD exhibit degraded 3D generation quality, including incomplete geometry reconstruction, unstable texture synthesis, and inconsistent cross-view appearance. These results further demonstrate that the characteristics of existing datasets prevent them from providing the clean, multi-view supervision required by our framework, thereby motivating the construction of our own dataset.

6.3. Dataset Construction and Specification

To train and evaluate our LoRA distillation framework, we constructed a multi-view dataset comprising ten object categories: six mechanical components (screw, nut, bearing, gasket, nail, hexagonal stud) and four electronic components (ceramic capacitor, resistor, red LED, green LED). The dataset contains more than 200 high-resolution images (512×512 pixels), with 20 images per category evenly captured from front and up view perspectives. The data were collected using a standardized imaging setup equipped with a uniform ring-light illumination system, as illustrated schematically in Figure 8.

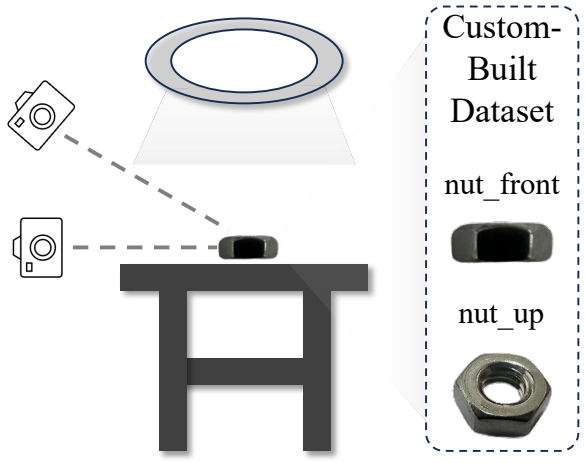
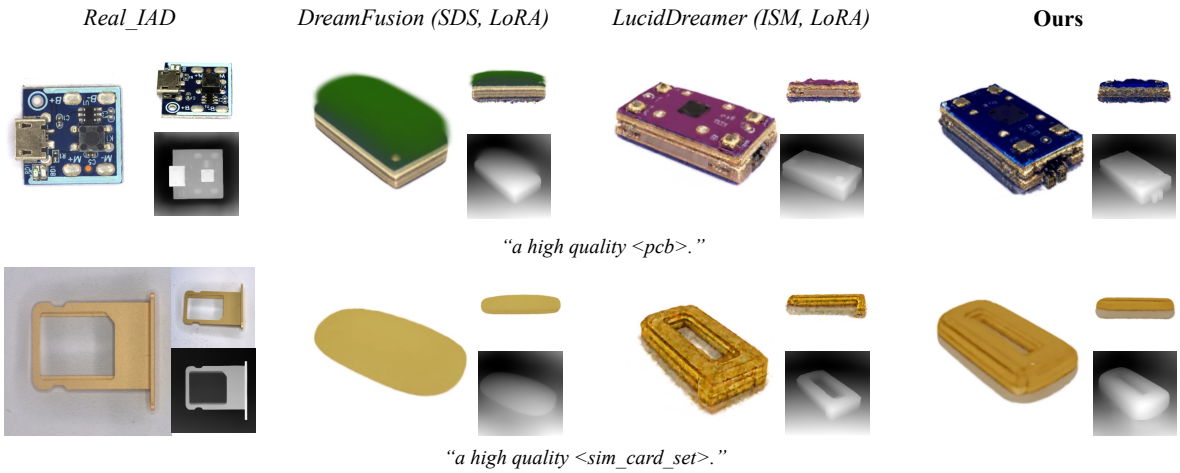


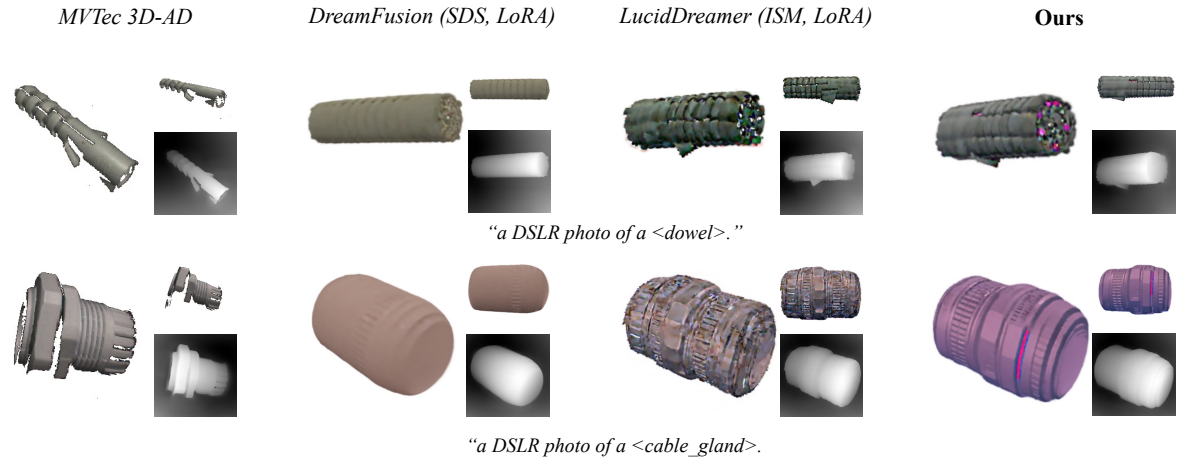
Figure 8. Our data capture apparatus (left) and collected dataset samples (right). The setup utilizes controlled lighting and fixed mounts to acquire consistent front and top-down view images for each industrial component.

For each object category, we captured 10 front view images with objects positioned to show their primary functional surface, and 10 up view images photographed from directly above to reveal structural details. All images were acquired under controlled studio lighting with diffused illumination against a neutral white background, using fixed camera settings to ensure consistent quality and exposure. The standardized acquisition protocol enables comprehensive feature learning by providing dual perspectives that capture both surface textures and geometric characteristics essential for robust LoRA training.

The dataset underwent rigorous quality control including sharpness verification, exposure consistency checks, and annotation accuracy validation. This balanced multi-view configuration allows LoRAs to recognize objects from multiple perspectives while capturing fine-grained differences between similar categories. The equal distribution across categories and viewpoints prevents bias during LoRA adaptation and ensures stable training for the distillation process.



(a) Results on the Real-IAD dataset. Our method successfully identifies and localizes industrial defects.



(b) Comparison on the MVTec 3D-AD dataset. Our approach provides superior defect coverage compared to prior methods.

Figure 9. Qualitative results on two public industrial datasets. (a) Real-IAD: our method accurately detects and localizes defects. (b) MVTec 3D-AD: our framework achieves improved defect completeness and surface consistency.

6.4. Implementation Details

6.4.1. Distillation Phase Configuration.

The distillation process is structured as a two-stage training pipeline, designed to systematically transfer knowledge from the teacher experts to the unified student model. Both Stage 1 and Stage 2 follow identical training configurations. Each stage is trained for 5,000 iterations to ensure sufficient convergence of the student's features while maintaining computational efficiency. We employ Low-Rank Adaptation (LoRA) with a rank parameter set to 16. This rank provides an optimal balance between parameter efficiency and model expressiveness, allowing the adapter to capture nuanced, domain-specific features without introducing excessive parameters. This LoRA configuration allows for effective fine-tuning while reducing the number of trainable parameters compared to full fine-tuning approaches.

6.4.2. ForgeDreamer Training Protocol.

The ForgeDreamer training process follows a schedule spanning 5,000 iterations. The training begins with a warm-up phase covering the initial 1,500 iterations, during which the learning rate increases to its target value. This warm-up strategy helps stabilize the early stages of training and prevents potential optimization instabilities from aggressive initial learning rates. Following this stabilization period, the remaining 3,500 iterations proceed with the training protocol.

During ForgeDreamer training, we maintain a batch size of 4 throughout the entire training process. This batch size was selected to achieve maximum training efficiency by utilizing the available VRAM on the NVIDIA RTX 4090 24G GPU, while also ensuring stable gradient updates and maintaining reproducible results across different experimental runs.

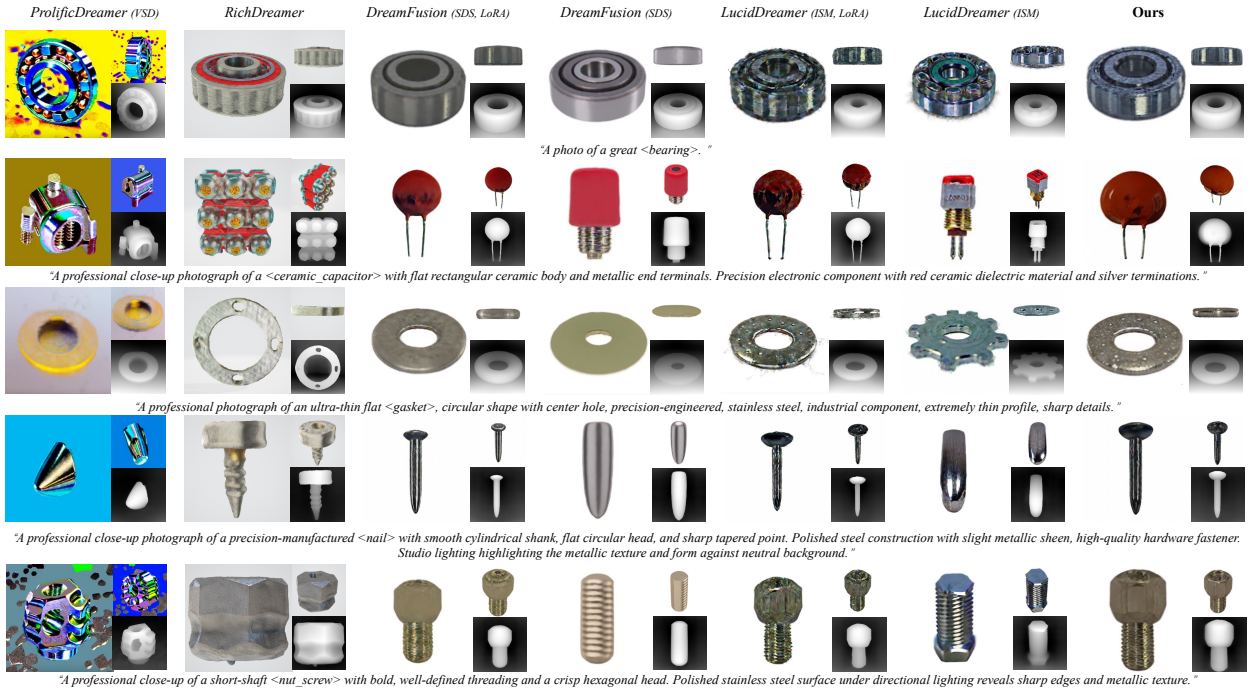


Figure 10. Qualitative Comparison with State-of-the-Art Methods. Visual results demonstrate the superior performance of our approach (remaining categories).

6.4.3. Hardware and Computational Environment.

All experimental procedures, including both the LoRA distillation phases and the complete ForgeDreamer training pipeline, are conducted on a single NVIDIA RTX 4090 24G GPU. This hardware configuration provides sufficient computational resources for the training requirements while maintaining accessibility for research purposes. The use of a single GPU setup ensures consistent experimental conditions and eliminates potential variations that might arise from distributed training configurations.

6.5. Additional Industrial Scene Results

We present further analysis under different industrial settings to thoroughly evaluate our method's capabilities. This includes examining the performance when processing varying numbers of industrial objects, simulating real-world scenarios that range from single-object synthesis to more complex, populated scenes. To provide a more intuitive understanding of the generation quality differences, we visualize the detailed comparative results across these different experimental configurations.

As shown in Figure 10, which complements the visual results in the main paper, we include extended comparisons with additional baseline methods. This figure focuses on the remaining five industrial categories: **bearing, ceramic capacitor, gasket, nail, and hexagonal stud**. These visual results clearly demonstrate the robustness and generality of our approach in industrial object generation. Across these categories, our method consistently produces higher-fidelity geometry, more accurate structural details (such as the threading on studs or contacts on capacitors), and avoids common failure modes like surface oversmoothing or disconnected parts that are visible in baseline results. The comprehensive analysis reveals consistent performance improvements across these different experimental configurations within the industrial domain.

6.6. Detailed LoRA Fusion Analysis

Detailed analysis of the LoRA fusion mechanisms reveals significant insights into the behavior of different fusion strategies across varying configurations. Through systematic evaluation of 2, 4, and 6 LoRA configurations, we investigate both quantitative performance metrics and underlying representational changes.

The visualizations in Figure 11, Figure 12, and Figure 13 present these analyses. The cosine similarity analysis (subplot a) provides a direct measurement of how well fused LoRAs preserve individual component characteristics, while PCA decomposition (subplot b) offers geometric insights into the structural relationships within the learned representation space.

To further quantify these findings, Table 3 provides a detailed comparison of concept preservation scores for

Table 3. Concept Preservation Scores by Fusion Method and LoRA Configuration

Method (CLIP-ViT-L/14)	**Two LoRAs**	**Four LoRAs**	**Six LoRAs**
Individual Concepts			
Emb1 (Addition)	0.899	0.793	0.658
Emb1 (Distillation)	**0.927**	**0.934**	**0.904**
Emb2 (Addition)	0.886	0.779	0.653
Emb2 (Distillation)	**0.915**	**0.963**	**0.963**
Emb3 (Addition)	—	0.855	0.708
Emb3 (Distillation)	—	**0.972**	**0.968**
Emb4 (Addition)	—	0.827	0.691
Emb4 (Distillation)	—	**0.927**	**0.936**
Emb5 (Addition)	—	—	0.630
Emb5 (Distillation)	—	—	**0.969**
Emb6 (Addition)	—	—	0.458
Emb6 (Distillation)	—	—	**0.970**
Overall Performance			
Average (Addition)	0.938	0.814	0.633
Average (Distillation)	**0.965**	**0.949**	**0.952**

our distillation method versus simple additive fusion. The data shows that our distillation method consistently and significantly outperforms additive fusion across all configurations (e.g., 0.952 vs. 0.633 average for six LoRAs), demonstrating its effectiveness in preventing catastrophic interference.

6.7. LLM-Based Qualitative Evaluation

To provide an objective and detailed qualitative assessment, we employed a Large Language Model (LLM) to evaluate and compare our method against the baselines. This qualitative feedback was then distilled into a quantitative ranking by tallying the LLM's preferences across all comparisons, as summarized in Table 4. Table 4 details these results, where a lower rank indicates better user preference. Our method achieved the best possible rank (1.0) in 6 out of 10 categories for an average rank of 1.6, significantly outperforming all baselines and confirming its superior perceptual quality. The following Figures 15 to 18 present the complete, verbatim responses from the LLM evaluator, which form the basis of this analysis. This evaluation provides a comprehensive and unbiased analysis of the geometry, texture, and prompt adherence of the generated results across all ten industrial categories and additional prompts.

6.8. Natural Scene Generation Evaluation

To evaluate the generalizability of our approach beyond its primary industrial applications, we conducted a comparative analysis on common natural scene generation tasks. While ForgeDreamer is specifically designed and optimized for industrial object synthesis, we demonstrate that our unified LoRA fusion strategy still maintains highly competitive performance in these distinct natural scene contexts.

Figure 14 presents qualitative comparisons between our method and baseline approaches on "bagel" and "hamburger" objects. The results indicate that our fusion mechanism successfully preserves the critical semantic coherence and high visual quality necessary for natural image generation. Crucially, the specialized industrial-focused training does not unduly compromise the model's ability to handle general-purpose generation tasks. This cross-domain evaluation validates that ForgeDreamer achieves its domain-specific optimization without sacrificing broader applicability, making it suitable for mixed industrial-natural generation scenarios that are commonly encountered in practical applications.

6.9. Per-Category T3Bench Scores

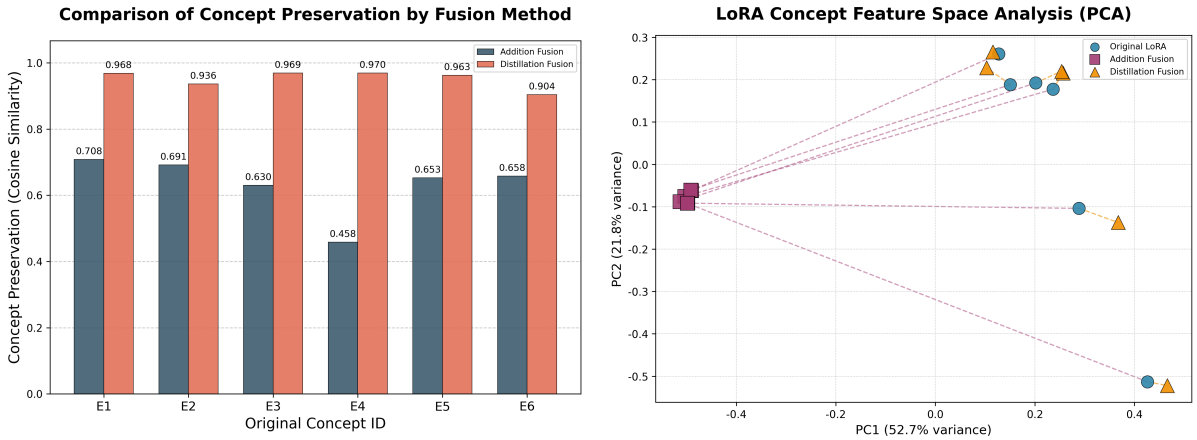
Table 5 presents the full breakdown of T3Bench quality scores across all ten industrial categories. This detailed data substantiates the findings from the main paper. It shows that our method not only achieves the highest average score (50.88) by a significant margin over the next-best competitor (47.10), but also secures a top-two result (best or second-best) in 7 out of the 10 categories.

Specifically, our method achieves the **number one** rank in four distinct categories: **G. LED**, **Screw**, **Bearing**, and **Gasket**. It also secures the **second-best** rank in three others: **R. LED**, **Hex Stud**, and **Nail**.

This demonstrates a high degree of robustness. It is particularly noteworthy that while some baselines (e.g., LucidDreamer w/o LoRA) achieve exceptionally high scores in a few specific categories where our method does not lead (such as **Nut** and **Capacitor**), they exhibit significant performance drops in others. In contrast, our method avoids catastrophic failures and maintains a high-quality baseline across the entire spectrum. This balance between high peak performance (achieving 1st place in 40% of categories) and strong consistency (placing in the top-two 70% of the time) is what drives the superior overall average score, confirming its robust generation quality for diverse industrial components. However, we believe this automated metric, while useful, does not fully reflect all perceptual nuances of generation quality and should be considered a strong reference rather than a complete assessment.

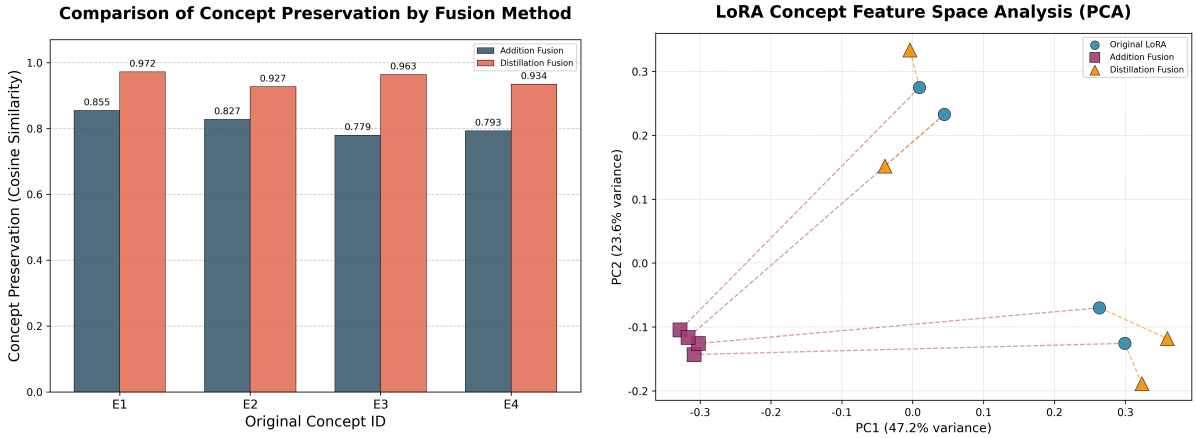
6.10. Cross-Configuration Analysis.

Comparing across all configurations (Figures 11, 12, and 13), we observe several consistent patterns that demonstrate the robustness of our approach:



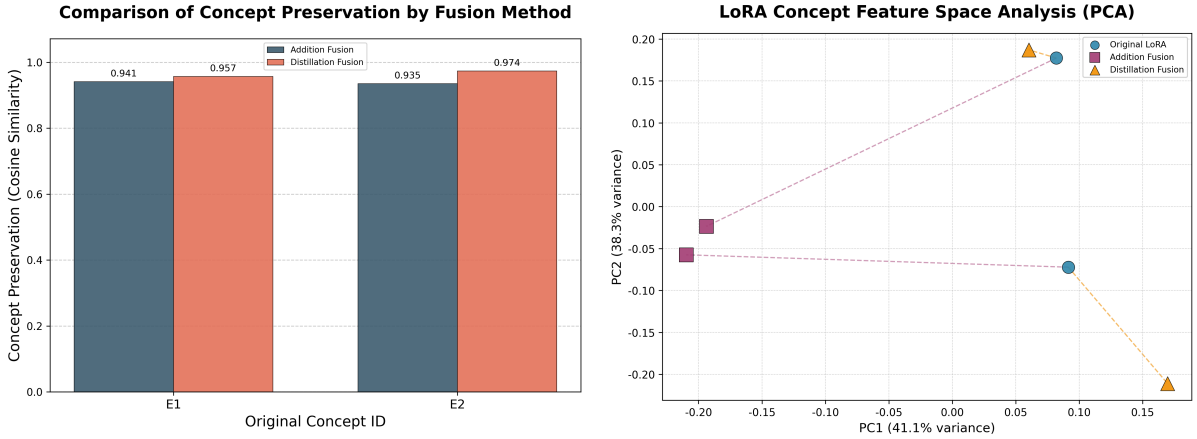
(a) Cosine Similarity (b) PCA Analysis

Figure 11. Analysis of Six LoRAs Configuration: (a) Cosine similarity comparison between addition fusion and ablation fusion methods. (b) PCA visualization showing the representational space distribution and relative distances to the original LoRA.



(a) Cosine Similarity (b) PCA Analysis

Figure 12. Analysis of Four LoRAs Configuration: (a) Cosine similarity metrics demonstrating the effectiveness of addition fusion over ablation fusion. (b) PCA decomposition revealing the structural relationships between fused LoRAs and the original representation.



(a) Cosine Similarity (b) PCA Analysis

Figure 13. Analysis of Two LoRAs Configuration: (a) Cosine similarity comparison showing the fundamental differences between fusion approaches. (b) PCA analysis illustrating the simplest case of LoRA fusion and its impact on representational geometry.

Table 4. Quantitative user study ranking across ten object categories. Lower ranks are better. Best (rank 1) and second-best (rank 2) results in each column are highlighted in **bold** and underlined, respectively.

Method	G. LED	Nut	R. LED	Resistor	Screw	Bearing	Capacitor	Hex Stud	Nail	Gasket	Average
ProlificDreamer [43] (w/o LoRA)	7	7	7	7	7	7	7	7	7	6	6.9
RichDreamer [34] (w/o LoRA)	6	6	6	6	6	4	6	6	4	3	5.3
DreamFusion [33] (w/ LoRA)	3	5	2	2	2	2	4	2	**1**	**1**	2.4
DreamFusion [33] (w/o LoRA)	5	4	5	5	5	**1**	**1**	5	5	4	4.0
LucidDreamer [20] (w/LoRA)	2	3	3	3	4	6	5	4	3	5	3.8
LucidDreamer [20] (w/o LoRA)	4	2	4	4	3	5	2	3	6	7	4.0
Ours	**1**	**1**	**1**	**1**	**1**	3	3	**1**	2	2	**1.6**

Table 5. Quantitative comparison across ten object categories on T3Bench quality scores. Best and second-best results in each column are highlighted in **bold** and underlined, respectively.

Method	G. LED	Nut	R. LED	Resistor	Screw	Bearing	Capacitor	Hex Stud	Nail	Gasket	Average
ProlificDreamer [43] (w/o LoRA)	27.19	22.34	24.49	26.60	34.85	31.64	20.48	7.85	26.39	29.45	25.13
RichDreamer [34] (w/o LoRA)	17.21	42.26	15.58	42.31	38.33	21.24	17.83	24.38	39.90	23.70	28.27
DreamFusion [33] (w/o LoRA)	32.62	69.72	29.92	28.49	46.59	28.38	35.45	36.66	52.27	59.02	41.91
DreamFusion [33] (w/ LoRA)	48.27	63.31	31.98	38.40	56.29	22.76	21.12	50.61	**62.27**	53.27	44.83
LucidDreamer [20] (w/o LoRA)	44.46	**71.37**	31.24	**53.23**	53.97	36.86	**45.51**	**58.93**	39.66	35.72	47.10
LucidDreamer [20] (w/ LoRA)	48.45	64.62	**53.98**	30.16	57.31	20.60	22.45	54.23	57.39	58.30	46.75
Ours	**51.68**	65.13	47.23	38.59	**58.42**	**45.15**	28.63	57.12	57.55	**59.25**	**50.88**

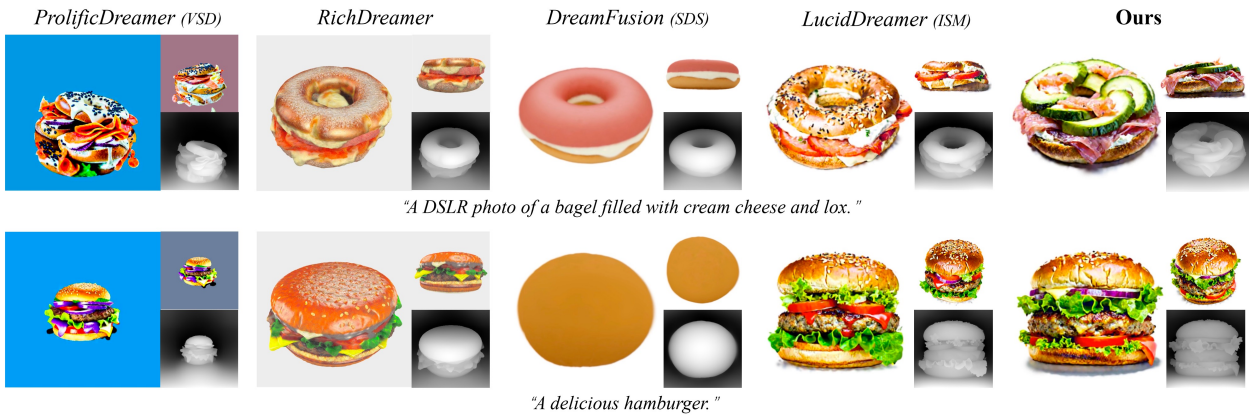
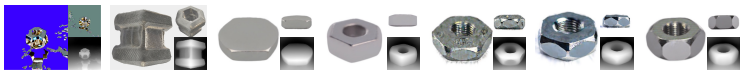


Figure 14. Qualitative comparison with state-of-the-art methods on natural scenes. Visual results demonstrate the superior performance of our approach.

- **Scalability**: Our distillation fusion (labeled "Addition" in the charts) consistently outperforms ablation fusion across all LoRA quantities, with performance advantages becoming more pronounced as complexity increases. Our method effectively handles multi-LoRA integration without suffering from interference effects that typically plague naive fusion strategies.
- **Stability**: The cosine similarity scores remain stable as the number of LoRAs increases, indicating robust fusion behavior. Unlike traditional methods that exhibit degradation with increased complexity, our approach maintains consistent performance even when integrating six different LoRAs.
- **Representational Coherence**: PCA analysis shows that our fusion maintains better alignment with the original LoRA space. The geometric analysis reveals that our fusion strategy preserves essential directional characteristics while creating meaningful combinations, evidenced by tighter clustering patterns and reduced variance in principal components.

The cross-configuration analysis reveals our fusion method's performance advantage over ablation fusion becomes pronounced as the number of LoRAs increases. This suggests the benefits our approach are not merely additive, but are amplified when tackling complex, high-demand fusion scenarios. These quantitative findings, when combined with the comprehensive qualitative comparisons presented in Figure 10, provide strong evidence for the effectiveness and the superior generalizability of our proposed LoRA fusion strategy.



Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional product photo of a precision hex <nut> with bright stainless steel surface, polished metallic finish, clean hexagonal edges.". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

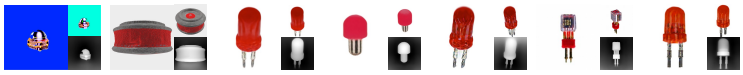
A: Based on the visual quality, geometric accuracy, and surface detail of these text-to-3D generated hex nuts, I would rank them as follows from best to worst: **Image 7, Image 6, Image 5, Image 4, Image 3, Image 2, Image 1**. Image 7 demonstrates the highest quality with clean hexagonal geometry, convincing metallic surface rendering, and proper threading visible in the center hole—it closely resembles a professional product photograph. Image 6 follows closely with similar geometric precision and good surface reflections, though slightly less refined in the threading detail. Image 5 shows reasonable structural accuracy with visible threading, but the surface appears somewhat weathered or oxidized rather than the requested polished finish. Image 4 maintains basic hexagonal shape but lacks threading detail and presents a more matte, simplified appearance. Image 3 exhibits smooth geometry but completely omits the threaded hole, presenting instead as a simplified rounded hexagonal cap with poor adherence to the prompt. Image 2 shows severe geometric degradation with rough, artifact-laden surfaces that appear more like failed 3D printing than polished metal, though it attempts to show threading structure. Image 1 is clearly the worst, displaying catastrophic reconstruction failure with fragmented geometry, floating artifacts, and complete loss of structural coherence—it barely resembles a nut at all and suggests fundamental failure in the 3D generation pipeline.

(a) Nut

Q: These are the results generated by seven different text-to-3D methods with the prompt "A highly detailed, industrial-grade metallic <screw> with sharp helical threading and a polished steel surface.". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 3, Image 6, Image 5, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with exceptional geometric precision featuring a well-defined hexagonal socket head that demonstrates accurate internal drive geometry, clean and properly proportioned external threading on a short shaft, convincing stainless steel surface texture with appropriate metallic sheen, and effective lighting that highlights both the sharp edges and the fastener's overall form—it successfully represents a socket head cap screw (often called an Allen bolt) with professional-grade detail that closely matches the "bold, well-defined threading and crisp hexagonal head" specified in the prompt. Image 3 follows with good hexagonal head geometry, visible threading, and appropriate metallic finish, though the head design is simpler and the threading definition is slightly less refined than Image 7. Image 6 demonstrates solid construction with clear hexagonal features, well-rendered threading, and convincing steel appearance with blue-tinted highlights, though the shaft length exceeds the "short-shaft" specification. Image 5 shows recognizable fastener structure but exhibits an unusual faceted head appearance that deviates from standard hardware geometry, with less polished surface quality. Image 4 fundamentally misses the prompt by generating only a threaded cylindrical shaft without any hexagonal head component. Image 2 produces an amorphous, eroded form that barely resembles a fastener, lacking clear threading and geometric definition. Image 1 represents complete generation failure with chaotic, rainbow-colored fragmented geometry that bears no resemblance to any hardware component.

(b) Screw

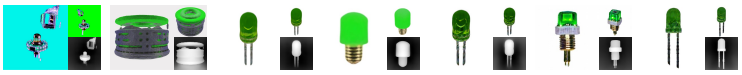


Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional close-up photograph of a clear <red_LED> with transparent dome-shaped epoxy casing and visible internal semiconductor chip". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 3, Image 5, Image 6, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with an accurate dome-shaped transparent red epoxy casing, properly rendered two-lead configuration, and visible internal structure that suggests the semiconductor chip—it successfully captures both the "clear" translucent quality and the characteristic LED form factor specified in the prompt. Image 3 follows closely with excellent dome geometry, appropriate translucency in the red casing, clearly defined leads, and reasonable internal detail visibility, though the surface appears slightly less refined than Image 7. Image 5 demonstrates good overall LED structure with proper dome shape and dual leads, but exhibits some surface artifacts and slight geometric irregularities that reduce the "professional close-up photograph" quality. Image 6 takes an interesting approach by attempting to show internal chip detail more explicitly, but the execution results in an overly complex, less accurate representation that doesn't maintain the characteristic smooth dome shape expected of an LED. Image 4 shows a simplified LED with reasonable proportions and red coloring, but lacks the transparent/translucent quality emphasized in the prompt and appears more opaque and toy-like rather than resembling a real electronic component. Image 2 fundamentally misinterprets the prompt by generating what appears to be a push-button or mechanical component with a red cap rather than an LED—it completely misses the dome-shaped epoxy casing and semiconductor nature. Image 1 represents total generation failure with fragmented, scattered geometry that bears no resemblance to an LED or any coherent 3D object, indicating complete inability to parse or execute the prompt.

(c) Red LED

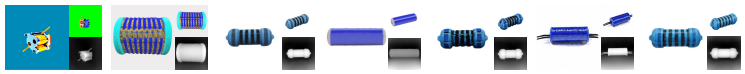
Figure 15. LLM-based qualitative evaluation (Part 1-3).



Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional close-up photograph of a clear ‹green_LED› with transparent dome-shaped epoxy casing and visible internal semiconductor chip. ". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 5, Image 3, Image 6, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with an exceptionally accurate dome-shaped transparent green epoxy casing, properly proportioned dual leads, visible internal structure suggesting the semiconductor chip, and the characteristic translucent quality that allows light to pass through—it perfectly captures the "clear" and "transparent" aspects emphasized in the prompt. Image 5 follows closely with excellent dome geometry, appropriate green translucency, well-defined leads, and visible internal details that suggest the chip structure, though the surface rendering is slightly less refined than Image 7. Image 3 demonstrates solid LED construction with good dome shape, proper transparency, and clear lead definition, successfully conveying the essential characteristics of a through-hole LED component, though with marginally less detail in the internal structure. Image 6 shows reasonable LED geometry and attempts to display internal chip detail more explicitly, but the addition of what appears to be a threaded base and overly complex internal structure makes it less accurate to a standard LED form factor, appearing more like a hybrid indicator light. Image 4 fundamentally misinterprets the prompt by generating what looks like a miniature light bulb with a screw base rather than an LED—while it has green coloring and transparency, it completely misses the characteristic LED dome shape and dual-lead configuration. Image 2 represents a significant deviation, producing what appears to be a multi-layered mechanical component or circuit board assembly rather than an LED, entirely missing the dome-shaped epoxy casing and simple two-lead structure. Image 1 shows complete generation failure with fragmented, chaotic geometry consisting of disconnected polygons and voxel-like artifacts that bear no resemblance to an LED or any coherent electronic component, indicating total inability to interpret the prompt.

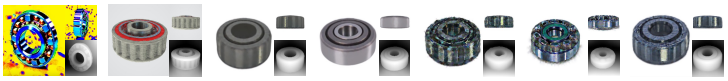
(a) Green LED



Q: These are the results generated by seven different text-to-3D methods with the prompt "A detailed close-up photograph of a cylindrical ‹resistor› body only, with blue ceramic material and black color code bands wrapped around its surface. ". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 3, Image 5, Image 6, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with a clean cylindrical body, well-defined alternating blue and black color bands that accurately represent resistor color coding, proper proportions with axial leads, and smooth surface rendering that captures the ceramic material appearance specified in the prompt. Image 3 follows closely with excellent cylindrical geometry and clear band definition, showing good color separation between blue and black bands, though the overall form appears slightly more compressed than Image 7. Image 5 demonstrates similar quality to Image 3 with proper banding and cylindrical shape, but exhibits slightly more surface irregularities and less uniform band spacing that reduce the "detailed close-up photograph" quality. Image 6 shows reasonable resistor structure with blue coloring and visible leads, but the bands are less distinct and the body appears more like a solid blue cylinder with minimal color code detail, plus it includes text labeling that wasn't requested in the prompt. Image 4 presents a simplified interpretation with a predominantly blue cylindrical body but almost completely lacks the "black color code bands" that are central to the prompt, appearing more like a generic capacitor than a resistor. Image 2 fundamentally misinterprets the component by generating what appears to be a ribbed or corrugated cylindrical structure that looks more like a flexible connector or cable component rather than a resistor—while it has blue coloring, it completely misses the smooth ceramic body with distinct color bands. Image 1 represents total generation failure with chaotic, fragmented geometry showing disconnected polygons and artifacts that bear no resemblance to a cylindrical resistor or any coherent electronic component, indicating complete inability to parse the prompt.

(b) Resistor



Q: These are the results generated by seven different text-to-3D methods with the prompt "A photo of a great ‹bearing›. ". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 3, Image 5, Image 6, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with a clean cylindrical body, well-defined alternating blue and black color bands that accurately represent resistor color coding, proper proportions with axial leads, and smooth surface rendering that captures the ceramic material appearance specified in the prompt. Image 3 follows closely with excellent cylindrical geometry and clear band definition, showing good color separation between blue and black bands, though the overall form appears slightly more compressed than Image 7. Image 5 demonstrates similar quality to Image 3 with proper banding and cylindrical shape, but exhibits slightly more surface irregularities and less uniform band spacing that reduce the "detailed close-up photograph" quality. Image 6 shows reasonable resistor structure with blue coloring and visible leads, but the bands are less distinct and the body appears more like a solid blue cylinder with minimal color code detail, plus it includes text labeling that wasn't requested in the prompt. Image 4 presents a simplified interpretation with a predominantly blue cylindrical body but almost completely lacks the "black color code bands" that are central to the prompt, appearing more like a generic capacitor than a resistor. Image 2 fundamentally misinterprets the component by generating what appears to be a ribbed or corrugated cylindrical structure that looks more like a flexible connector or cable component rather than a resistor—while it has blue coloring, it completely misses the smooth ceramic body with distinct color bands. Image 1 represents total generation failure with chaotic, fragmented geometry showing disconnected polygons and artifacts that bear no resemblance to a cylindrical resistor or any coherent electronic component, indicating complete inability to parse the prompt.

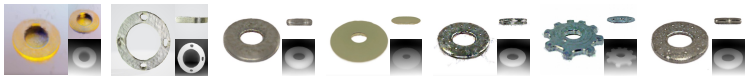
(c) Bearing

Figure 16. LLM-based qualitative evaluation (Part 4-6).

Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional close-up photograph of a <ceramic_capacitor> with flat rectangular ceramic body and metallic end terminals. Precision electronic component with red ceramic dielectric material and silver terminations. ". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 4, Image 6, Image 3, Image 7, Image 5, Image 2, Image 1**. Image 4 achieves the highest quality with a proper rectangular ceramic body in red/pink coloring, clearly visible metallic threaded terminal at the base, appropriate proportions, and clean geometry that successfully represents a ceramic capacitor component, though it appears more like a larger power capacitor rather than a surface-mount type. Image 6 follows with reasonable component structure showing a red top section, visible metallic threading, dual leads, and an attempt at showing internal structure through transparency, though the execution makes it appear somewhat hybrid between different capacitor types. Image 3 demonstrates acceptable disc capacitor geometry with a red circular ceramic body and dual wire leads, representing a through-hole radial capacitor variant, though it deviates from the "flat rectangular" specification in the prompt by using a disc form factor. Image 7 shows similar disc capacitor structure with smoother surface rendering and appropriate red-brown ceramic coloring, but like Image 3, it misinterprets the rectangular body requirement while still capturing the essence of a ceramic capacitor. Image 5 exhibits a disc capacitor form with significant surface texture artifacts and irregular coloring that detracts from the "professional close-up photograph" quality, making it appear less like a precision electronic component. Image 2 fundamentally misinterprets the prompt by generating what appears to be a multi-unit array or module with stacked components, red separators, and yellow dots, completely missing the single discrete capacitor specification and looking more like a complex assembly than an individual component. Image 1 represents total generation failure with chaotic, fragmented geometry featuring iridescent rainbow coloring, scattered polygons, and completely incoherent structure that bears no resemblance to any type of capacitor, indicating complete inability to parse or execute the prompt successfully.

(a) Ceramic Capacitor



Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional photograph of an ultra-thin flat <gasket>, circular shape with center hole, precision-engineered, stainless steel, industrial component, extremely thin profile, sharp details.". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results again, I would revise my ranking from best to worst as follows: **Image 3, Image 7, Image 2, Image 4, Image 5, Image 1, Image 6**. Image 3 achieves the highest quality with excellent circular geometry, a properly centered hole, convincing metallic surface texture suggesting stainless steel, and an appropriately ultra-thin profile clearly visible in the side view with sharp, clean edges conveying precision engineering. Image 7 follows with good circular form and appropriate thinness, though the surface appears more granular or textured, which could actually represent a realistic gasket material texture rather than being a flaw. Image 2 demonstrates solid gasket structure with a clean circular form and proper central hole, and the visible mounting holes around the perimeter are actually a realistic feature for industrial gaskets that need to be bolted in place, making this a legitimate interpretation despite not being explicitly mentioned in the prompt. Image 4 shows acceptable circular geometry with good flatness and appropriate proportions for a thin gasket, though the surface appears somewhat plain and less detailed. Image 5 exhibits reasonable gasket form with visible surface texture that could represent material grain or slight wear, though the texture appears somewhat excessive and the edges less precise. Image 1 is problematic because it shows a thick, yellow-tinted disc with significant discoloration that completely misses the "ultra-thin" and "stainless steel" specifications, appearing more like a plastic or composite material. Image 6 fundamentally misinterprets the prompt by generating a gear or sprocket with external teeth, which contradicts the basic "circular shape" specification for a simple flat gasket—this represents a completely different type of component regardless of its 3D modeling quality.

(b) Gasket

Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional close-up photograph of a precision-manufactured <nail> with smooth cylindrical shank, flat circular head, and sharp tapered point. Polished steel construction with slight metallic sheen, high-quality hardware fastener. Studio lighting highlighting the metallic texture and form against neutral background. ". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 3, Image 7, Image 5, Image 2, Image 4, Image 6, Image 1**. Image 3 achieves the highest quality with accurate nail geometry featuring a properly proportioned flat circular head, smooth cylindrical shank, sharp tapered point, and convincing dark metallic finish that suggests polished steel with appropriate sheen—it successfully captures all the key elements specified in the prompt including the precision-manufactured appearance. Image 7 follows closely with excellent overall nail structure, clear flat head, cylindrical body, and pointed tip, along with good metallic rendering, though the surface texture shows slight irregularities that make it marginally less refined than Image 3. Image 5 demonstrates solid nail construction with appropriate proportions and visible metallic characteristics, but the head appears slightly more textured or worn and the overall surface quality is less polished than the top performers. Image 2 shows recognizable nail structure with a flat head and threaded/textured shank, but the spiral grooves and weathered appearance deviate from the "smooth cylindrical shank" and "polished steel" specifications, making it look more like a specialty fastener or aged hardware rather than a precision-manufactured standard nail. Image 4 fundamentally misinterprets the prompt by generating what appears to be a smooth, bullet-shaped or cosmetic applicator-like object that completely lacks the characteristic flat head of a nail, instead presenting an elongated teardrop form that doesn't represent hardware fastener geometry. Image 6 similarly produces an overly smooth, rounded form that looks more like a decorative or cosmetic item than a nail, with gradient coloring and lack of distinct head definition that fails to convey the utilitarian hardware nature specified. Image 1 represents complete generation failure with chaotic, fragmented geometry showing iridescent rainbow coloring and irregular polygonal forms that bear no resemblance to a nail whatsoever, indicating total inability to interpret or execute the prompt successfully.

(c) Nail

Figure 17. LLM-based qualitative evaluation (Part 7-9).

Q: These are the results generated by seven different text-to-3D methods with the prompt "A professional close-up of a short-shaft <nut_screw> with bold, well-defined threading and a crisp hexagonal head. Polished stainless steel surface under directional lighting reveals sharp edges and metallic texture.". Please rate the overall quality and details of the generated results from best to worst, and provide your reasons. Explain in a coherent paragraph.

A: Looking at these seven text-to-3D generation results, I would rank them from best to worst as follows: **Image 7, Image 3, Image 6, Image 5, Image 4, Image 2, Image 1**. Image 7 achieves the highest quality with exceptional geometric accuracy featuring a well-defined hexagonal head with crisp edges, clean and properly proportioned threading on a short shaft, convincing metallic surface texture suggesting polished stainless steel, and appropriate lighting that reveals the component's form—it successfully captures all essential characteristics of a precision fastener as specified in the prompt. Image 3 follows closely with accurate hexagonal head geometry, visible threading, reasonable proportions, and appropriate metallic finish, though the head appears slightly more rounded and the threading definition is marginally less sharp than Image 7. Image 6 demonstrates solid fastener construction with a clear hexagonal head, well-rendered threading, and good metallic appearance with blue-tinted highlights suggesting steel, though the shaft appears slightly longer than the "short-shaft" specification and some surface details are less refined. Image 5 shows recognizable bolt structure with hexagonal head and threading, but the head has an unusual faceted or crystalline appearance that deviates from typical hardware geometry, and the overall surface quality appears less polished despite attempting metallic rendering. Image 4 fundamentally misinterprets the prompt by generating what appears to be a set screw or threaded rod without any hexagonal head—it's essentially just a cylindrical threaded shaft, completely missing the "crisp hexagonal head" requirement that defines the fastener type. Image 2 significantly deviates by producing an amorphous, rock-like or eroded object with vague geometric features and minimal threading detail, appearing more like a weathered artifact than a precision-manufactured fastener, with only a hint of hexagonal form and no clear threading. Image 1 represents total generation failure with chaotic, fragmented geometry featuring iridescent rainbow coloring, scattered polygons, and completely incoherent structure that bears no resemblance to any hardware component, indicating complete inability to parse or execute the prompt successfully.

(a) Hexagonal Stud

Q: Please use a short paragraph to evaluate the overall effectiveness of these seven models and rank them.

A: Based on the comprehensive evaluation across nine different industrial component prompts (hex nut, screw, red LED, green LED, resistor, bearing, ceramic capacitor, gasket, nail, and nut-screw), a clear performance hierarchy emerges among these seven text-to-3D generation methods. Overall ranking from best to worst: **Model 7, Model 3, Model 6, Model 5, Model 4, Model 2, Model 1**. Model 7 demonstrates exceptional consistency and reliability, consistently producing geometrically accurate results with proper material rendering, clean edges, and faithful adherence to prompt specifications—it excels particularly at precision components requiring sharp detail and metallic finishes. Model 3 follows as a strong performer with good geometric accuracy and reasonable surface quality, though occasionally showing simplified details or less refined textures compared to Model 7. Model 6 shows decent competency with recognizable component structures and appropriate material rendering, but sometimes introduces unnecessary complexity or deviates from standard component forms. Model 5 produces variable results with acceptable basic geometry but frequently exhibits surface artifacts, crystalline appearances, or texture irregularities that reduce realism. Model 4 demonstrates significant weaknesses by often misinterpreting fundamental component characteristics—generating wrong geometries like threaded rods instead of headed fasteners or bullet shapes instead of nails—despite occasionally showing clean surface rendering. Model 2 consistently struggles with accuracy, frequently producing oversimplified, weathered, or completely wrong component types (like mechanical assemblies instead of simple capacitors), indicating poor prompt comprehension. Model 1 represents catastrophic failure across nearly all prompts, generating fragmented, chaotic geometry with rainbow artifacts and scattered polygons that bear no resemblance to the requested objects, suggesting fundamental issues in the generation pipeline that prevent any useful output.

(b) Overall Comment

Figure 18. LLM-based qualitative evaluation (Part 10-11).