JEWEL-T2I: LEARNING JEWELRY GENERATION FROM EXPERT SEED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

T2I generation can support jewelry customization by visualizing design requirements before physical prototyping, but general models often confuse craft structures or fail to preserve the requested motifs and materials. We present JEWEL-T2I, which constructs training data from expert jewelry assets for domain adaptation. Its data pipeline, JEWELGEN, expands 500 seed assets through semantic and viewpoint editing, while visual craft exemplars guide the synthesis of plain, filigree, and enamel designs. MLLM screening and designer verification select 20K image-text pairs for adapting the diffusion model. The resulting generator is combined with a prompt model trained on a separate curated jewelry corpus, allowing users to specify design requirements without supplying reference images at inference. We also introduce JEWELBENCH, a benchmark of 2K entries that evaluates AF, CF, and AQ across the three crafts. With identical prompts, the adapted generator improves over Qwen-Image by 6.3, 8.6, and 5.0 points on these criteria, respectively. The complete system achieves the highest mean scores among the compared systems across all three crafts under both automated and human evaluation. These results show that the constructed training data improves the ability of T2I models to follow jewelry design requirements on JEWELBENCH.

1 INTRODUCTION

Jewelry customization translates personal preferences into designs that must satisfy both visual and craft requirements. A customer's request may combine meaningful motifs, a preferred material, and a particular technique without specifying how these elements should be arranged. Developing such a request into a design requires communication with skilled designers through sketches, revisions, and physical sampling Ahde (2007); Moreau et al. (2020); Pozzobon (2021). Because design revisions require specialist effort and physical prototypes incur material and production costs, early visual feedback can help customers compare alternatives before committing to fabrication.

T2I generation offers a way to provide this feedback by visualizing variations in motifs, materials, and viewpoints directly from a description. For jewelry, however, the usefulness of a generated proposal depends on more than its visual appeal. Similar silhouettes can represent different craft structures, as in a filigree bracelet with fine wire openwork and a solid bracelet decorated with engraving. Enamel designs introduce further constraints on the boundaries between colored regions and metal, while motifs must remain recognizable when arranged across curved surfaces. A convincing proposal therefore needs to preserve the requested attributes and composition as well as produce an aesthetically coherent design.

General T2I models do not consistently satisfy these requirements, as the failures in Figure 1 illustrate. We evaluate these requirements through attribute fidelity (AF), compositional fidelity (CF), and aesthetic quality (AQ). AF covers craft, material, view, and lighting, while CF concerns the requested motifs and their relationships, and AQ measures overall appearance. This distinction matters because an attractive image can still depict the wrong craft or omit a design element. Adapting models to these requirements calls for suitable jewelry training data, yet collecting and validating such data requires expertise in both design and craft. Synthetic data can expand the available examples, provided that the construction process preserves the intended requirements and excludes incorrect outputs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Jewelry generation failures of Nano-Banana Fortin et al. (2025): errors in motif composition, craft and material attributes, and visual appearance. These examples motivate separate evaluation of CF, AF, and AQ.

Recent studies use synthetic data and feedback to improve generative models Zhang et al. (2025); Qu et al. (2025); Jha et al. (2025). In jewelry generation, expert assets provide an additional source of supervision by showing craft details that can be difficult to specify through text alone. The challenge is to expand design diversity while retaining these details, so that the resulting training data supports generation from text without requiring users to provide craft exemplars at inference.

We address this challenge with JEWELGEN, which constructs training data from 500 expert jewelry assets. Semantic editing varies motifs and materials, and viewpoint editing provides alternative views of the designs. To guide craft synthesis, the pipeline conditions generation on visual exemplars of the requested technique together with the design description. MLLM screening and designer verification then select the generated image-text pairs for supervised fine-tuning. The resulting image generator, JEWELDIFFUSION, forms JEWELCRAFT together with JEWELEXPERT, a domain prompt model that translates structured requirements into descriptions. This approach uses expert visual examples to guide data construction and adapts the generator to produce jewelry designs from text.

We evaluate the approach using JEWELBENCH, which contains 2K entries with structured design tuples, curated prompts, and reference images across plain, filigree, and enamel jewelry. Automated scoring and human assessment measure AF, CF, and AQ, while comparisons of prompt sources and image generators examine the behavior of the two model components. Our contributions are summarized as follows:

- We propose JEWELGEN, a data construction pipeline that combines design transformations, visual craft exemplars, and automated and expert filtering to expand 500 expert assets into training examples for jewelry generation.
- We develop JEWELCRAFT, which combines a diffusion model adapted on the generated image-text pairs with a prompt model adapted on 10K manually curated jewelry text samples, and introduce JEWELBENCH for evaluation across three crafts.
- We demonstrate improvements over Qwen-Image in AF, CF, and AQ on JEWELBENCH under identical prompts, and analyze the complete system across crafts using automated and human assessment.

2 RELATED WORK

Domain adaptation for image generation. Diffusion models Ho et al. (2020), latent diffusion Rombach et al. (2022), and diffusion transformers Peebles & Xie (2023) support T2I generation at scale, with models such as SDXL Podell et al. (2023) and FLUX.1-dev Labs (2024) providing

general image backbones. Specialized generation has been studied in virtual try-on Zhu et al. (2023); Yang et al. (2024c), medical imaging Chambon et al. (2022), avatar generation Cao et al. (2024), and interior design Gan et al. (2023). For jewelry, Magee (2024) compares generative tools and documents structural errors in their outputs, while Meng et al. (2025) constructs a targeted dataset for Stable Diffusion to generate designs incorporating tie-dye motifs. Our work focuses on constructing supervision from expert jewelry assets and craft exemplars, with evaluation that distinguishes attributes, composition, and aesthetics across three crafts.

Learning from generated data. Selecting synthetic examples before training is an established approach to improving generation. GORS fine-tunes a generator using reward-based sample selection Huang et al. (2023), while RSIDiff studies recursive improvement through prompt construction, preference sampling, and weighting based on the data distribution Zhang et al. (2025). SILMM uses iterative self-feedback and preference optimization to improve compositional alignment in multimodal models Qu et al. (2025), and SPRO optimizes prompt and image generation through self-play with synthetic reward signals Jha et al. (2025). Our approach uses expert assets, visual craft exemplars, and designer verification to construct jewelry data for supervised adaptation. The visual exemplars condition data synthesis, while the adapted generator receives text at inference.

Evaluating T2I generation. Existing benchmarks assess compositional alignment and other aspects of generation Huang et al. (2023); Ghosh et al. (2023); Li et al. (2025); Meng et al. (2024). FID Heusel et al. (2017) measures distributional similarity, and CLIP scores Radford et al. (2021) measure semantic alignment, but neither directly evaluates whether an image depicts the requested jewelry craft. MLLM evaluation supports more detailed assessment Wu et al. (2025b); Yu et al. (2025); Huang et al. (2025), while GenEval diagnoses object occurrence, position, count, and color and validates its automated assessments against human judgments Ghosh et al. (2023). JEWELBENCH complements these evaluations with structured jewelry requirements and craft references, using automated and human assessment to examine errors that broad semantic or aesthetic scores may overlook.

3 LEARNING FROM EXPERT SEED DATA

JEWEL-T2I incorporates expert knowledge through data construction and model adaptation. JEWELGEN expands a collection of expert assets into image-text pairs, using visual exemplars to guide craft synthesis and quality filtering to select training examples. The generation system, JEWELCRAFT, combines JEWELEXPERT for prompt generation with JEWELDIFFUSION for image synthesis, so that users provide structured design requirements at inference.

A design is represented by a tuple d specifying the jewelry type, craft, material, viewpoint, and motifs. Its description p expresses these requirements and the relationships between the elements, forming a candidate (d, p, x) with the generated image x. This representation maintains the correspondence between requirements and supervision during augmentation. When an edit changes a motif or material, the tuple and description are updated accordingly, allowing the generated image to be evaluated against the intended design.

3.1 DATA CONSTRUCTION WITH JEWELGEN

JEWELGEN begins with 500 seed assets comprising expert designs and photographs of physical jewelry. These assets serve as both source images for augmentation and visual exemplars for craft synthesis. As shown in Figure 2, we use Nano-Banana Fortin et al. (2025) to perform semantic editing, viewpoint editing, and synthesis conditioned on craft examples. Semantic and viewpoint editing diversify the requested designs, while visual exemplars provide guidance on craft structures during synthesis. Automated screening and designer review then select candidates according to the design requirements and craft appearance.

Semantic and viewpoint editing. Semantic editing expands the design space by changing selected motifs or materials, such as replacing a dragon with a phoenix or changing gold to silver. Viewpoint editing provides front and perspective views of the designs, allowing the training data to represent the requested appearance from different angles. Each operation updates the relevant tuple field

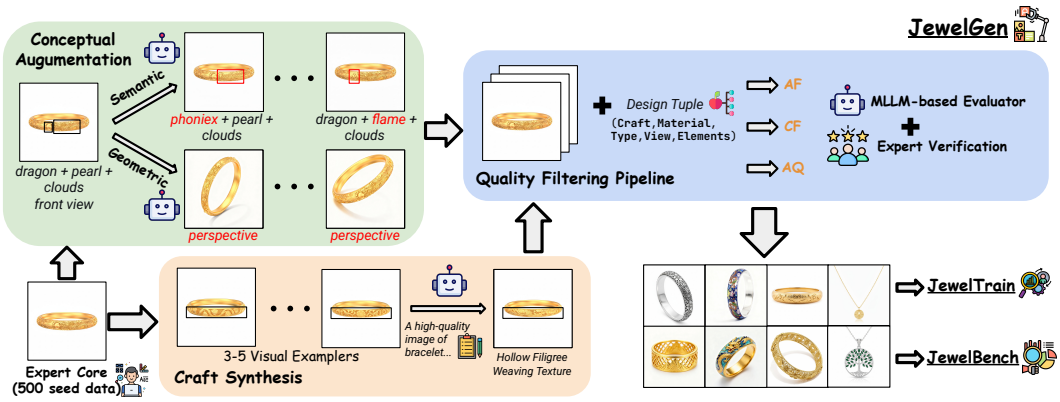


Figure 2: JEWELGEN: semantic and viewpoint editing expand the seed designs, while visual exemplars guide craft synthesis. MLLM screening and designer verification select the resulting candidates for training and evaluation.

and description while retaining the other requirements. For example, changing a motif to a phoenix should preserve the material and craft specified by the original design. The candidate is subsequently checked against this updated specification to reject unintended changes to other elements or the jewelry structure.

Craft synthesis with visual exemplars. Craft terminology does not always provide sufficient guidance for image synthesis, particularly when a model interprets filigree as engraving on a solid surface. We therefore condition synthesis on both the design description and three to five visual exemplars randomly sampled from the seed assets of the requested craft. Plain examples depict continuous metal surfaces, filigree examples provide openwork structures, and enamel examples illustrate colored regions bounded by metal. The exemplars supply visual information about the craft, while the description specifies the motifs, material, and viewpoint. The resulting images are paired with the design descriptions to train a generator that receives only text at inference.

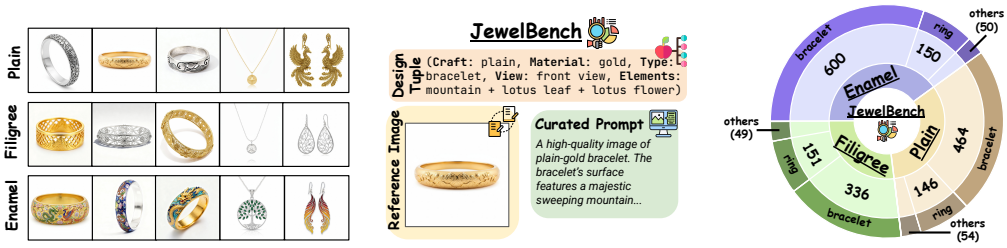
Quality filtering. Generated candidates are first evaluated by Gemini 2.5-pro Comanici et al. (2025) using AF, CF, and AQ to assess the requested attributes, element composition, and visual quality. Candidates that pass this screening are reviewed by jewelry designers for design consistency, craft appearance, and manufacturing plausibility. This combination allows explicit requirements to be checked across the candidate pool and reserves expert review for retained images that may contain less obvious craft or composition errors. A visually plausible pattern on a solid surface, for example, may still fail to represent the requested filigree structure. Appendix C provides the scoring prompt, and Appendix A presents rejected examples.

The accepted images are randomly partitioned after augmentation into 20K training pairs and 2K benchmark entries. The training pairs associate each description with an image selected through automated screening and designer review. Appendix A.3 summarizes the data construction operations and their outputs.

3.2 MODEL ADAPTATION AND INFERENCE

Image generation. We adapt Qwen-Image Wu et al. (2025a) on the 20K image-text pairs using LoRA to obtain JEWELDIFFUSION. These pairs associate descriptions of jewelry attributes and composition with the corresponding images, training the model to render designs from text without additional craft exemplars.

Prompt generation. JEWELEXPERT translates structured requirements into descriptions for JEWELDIFFUSION, specifying how the requested motifs, materials, crafts, and viewpoints should appear together. We obtain the model by adapting Qwen2.5-1.5B Yang et al. (2024a); Team (2024) with LoRA on a separate corpus of 10K manually curated jewelry text samples. The corpus includes design terminology, common motif combinations, and relationships between elements, such as a



(a) Examples sampled from JEWELBENCH.

(b) Benchmark entry.

(c) Benchmark statistics across craft subsets.

Figure 3: JEWELBENCH entries combine a design tuple, a curated prompt, and a reference image. The benchmark contains 664 plain, 536 filigree, and 800 enamel examples, covering multiple jewelry types, materials, and viewpoints.

dragon accompanied by clouds or a flaming pearl and relief decoration contained within an object's boundary. These examples support descriptions that express spatial relationships and craft details rather than simply enumerate the requested elements. At inference, JEWELEXPERT converts the user's requirements into a prompt, which JEWELDIFFUSION renders as a visual proposal for subsequent design work.

4 JEWELBENCH

JEWELBENCH contains 2K entries, each comprising a design tuple, a curated prompt, and a reference image. The tuple specifies the design requirements, while the prompt describes their arrangement and appearance. The reference image provides a visual example of the specified craft and design elements, complementing their textual description. By retaining both structured requirements and visual examples, the benchmark supports assessment of individual attributes and relationships that may be overlooked when jewelry is treated as a single object category.

The benchmark covers 664 plain, 536 filigree, and 800 enamel designs, with bracelets and rings forming the main groups and earrings and necklaces providing additional jewelry types. Materials include gold and silver, and the requested views include front and perspective views. Plain designs depict continuous metal surfaces without colored enamel and may include relief motifs or different surface finishes. Filigree designs emphasize fine wire openwork, whereas enamel designs combine colored regions with metal boundaries. These categories characterize the depicted craft appearance, and the scoring rubric distinguishes front and perspective views through level and angled placement, respectively. Figure 3 summarizes the examples, entry structure, and craft distribution.

AF measures agreement with the requested craft, material, view, and lighting, while CF measures the presence, quantity, and arrangement of design elements. AQ assesses overall appearance and design coherence, allowing visual appeal to be considered alongside fidelity to the specified design. An image may depict the requested dragon and flowers but substitute engraving for filigree, or it may render convincing filigree while omitting a flower. Both images can remain visually appealing despite failing different design requirements.

We report results by craft using automated and human assessment of the same design requirements and criteria. Appendix C provides the detailed rubric, and Appendix A.2 illustrates its application to an example entry.

5 EXPERIMENTS

We evaluate JEWELCRAFT against existing generation systems and compare prompt sources and image generators to examine the two model components. Automated and human assessment provide complementary views of performance across the three jewelry crafts.

Table 1: Automated evaluation on JEWELBENCH. Overall averages AF, CF, and AQ across all 2K entries, with craft subsets weighted by their sizes. Bold denotes the highest mean in each column.

	Plain			Filigree			Enamel			Overall
Model	AF	CF	AQ	AF	CF	AQ	AF	CF	AQ	
RPG	50.2	43.4	58.0	47.8	50.8	58.7	53.8	44.2	58.4	51.7
Personalized PR	43.4	42.8	58.6	48.7	53.3	61.4	41.7	43.4	59.1	49.9
GPT-4o + Nano-Banana	73.3	62.2	75.0	71.9	58.6	74.8	70.8	64.4	74.9	69.7
JEWELCRAFT	**74.8**	**70.8**	**79.9**	**72.0**	**70.4**	**79.9**	**78.9**	**67.2**	**79.8**	**74.9**

Table 2: Human evaluation on JEWELBENCH. Overall averages AF, CF, and AQ across all 2K entries, with craft subsets weighted by their sizes. Bold denotes the highest mean in each column.

	Plain			Filigree			Enamel			Overall
Model	AF	CF	AQ	AF	CF	AQ	AF	CF	AQ	
RPG	44.2	42.2	41.7	45.9	45.1	44.7	44.5	43.4	43.1	43.8
Personalized PR	44.3	41.8	40.5	43.7	42.9	40.8	41.2	39.6	38.1	41.2
GPT-4o + Nano-Banana	74.5	72.2	66.1	75.1	70.7	65.3	66.7	66.3	55.2	67.5
JEWELCRAFT	**83.1**	**83.1**	**80.6**	**85.2**	**82.9**	**78.5**	**84.4**	**82.3**	**78.3**	**82.0**

5.1 EXPERIMENTAL SETUP

Training and baselines. We use LoRA to adapt Qwen-Image on the 20K image-text pairs produced by JEWELGEN and Qwen2.5-1.5B on the separate corpus of 10K jewelry text samples. Both models are trained for two epochs, each on a single A100 GPU, with training settings summarized in Appendix B. The complete system is compared with RPG Yang et al. (2024b), Personalized PR Chen et al. (2024), and GPT-4o Hurst et al. (2024) paired with Nano-Banana Fortin et al. (2025). All systems receive the same 2K design tuples without the benchmark reference images, while Personalized PR additionally uses the user's previous jewelry requirements as prompt history. Appendix C.1 specifies the model combinations. For the image generator comparison, we include SDXL Podell et al. (2023), FLUX.1-dev Labs (2024), Qwen-Image Wu et al. (2025a), and Nano-Banana.

Evaluation protocol. Gemini 2.5-pro and human assessors evaluate each generated image against the same original design requirements using AF, CF, and AQ. These requirements remain fixed across methods, while the rewritten descriptions serve as inputs to image generation. Gemini assigns one of five scores, 20, 40, 60, 80, or 100, to each criterion, and reported gains denote differences in mean scores.

For the complete system comparison, five jewelry designers and five users independently rate all outputs for the 2K benchmark entries through an interface that hides model names and randomizes image order. Each image receives ten ratings, which are averaged for each criterion before aggregation across images. Overall scores average AF, CF, and AQ across the benchmark, weighting the three craft subsets by their sizes. Appendix C.3 provides the aggregation details.

5.2 SYSTEM COMPARISON

Tables 1 and 2 show that JEWELCRAFT achieves the highest mean AF, CF, and AQ scores among the compared systems in each craft category under both evaluations. Compared with GPT-4o + Nano-Banana, the overall score increases from 69.7 to 74.9 under automated evaluation and from 67.5 to 82.0 under human assessment. The improvements vary across criteria, with a larger automated gap in filigree CF (70.4 versus 58.6) than in filigree AF (72.0 versus 71.9). For plain designs, the largest difference is also in CF (70.8 versus 62.2), whereas enamel designs show the largest difference in AF (78.9 versus 70.8). Reporting each craft separately therefore reveals differences in attributes and composition that are less apparent from the overall score.

Table 3: Prompt comparison with JewelDiffusion fixed.

Prompt source	AF	CF	AQ	Avg.
Human-written	72.0	68.3	78.5	73.0
RPG	56.3	50.7	65.7	57.6
Personalized PR	50.1	50.9	65.3	55.4
GPT-4o	70.8	65.0	76.8	70.9
JEWELEXPERT	**75.7**	**69.2**	**79.9**	**74.9**

Table 4: Generator comparison with JewelExpert prompts.

Image model	AF	CF	AQ	Avg.
SDXL	58.3	53.4	68.0	59.9
FLUX.1-dev	57.8	60.8	68.7	62.4
Qwen-Image	69.4	60.6	74.9	68.3
Nano-Banana	66.2	69.0	77.8	71.0
JEWELDIFFUSION	**75.7**	**69.2**	**79.9**	**74.9**

Automated and human assessment. Both evaluations rank JEWELCRAFT first and GPT-4o + Nano-Banana second across all nine combinations of craft and criterion, supporting the main system comparison. However, they reverse the relative ranking of RPG and Personalized PR in six combinations, namely plain AF and AQ, all three filigree criteria, and enamel AQ. The performance gaps also differ in magnitude, with JEWELCRAFT improving enamel AQ over GPT-4o + Nano-Banana by 23.1 points under human assessment and 4.9 points under automated evaluation. The two evaluations agree on the leading systems while differing in baseline ordering and the magnitude of the performance gaps.

5.3 COMPONENT COMPARISONS

Table 4 evaluates the benefit of adapting the image backbone by fixing JewelExpert prompts and comparing JewelDiffusion with Qwen-Image and other generators. Table 3 examines the choice of prompt source with JewelDiffusion fixed. Both tables report means over the 2K entries, with Avg. denoting the mean of AF, CF, and AQ. The JewelExpert and JewelDiffusion rows reuse the complete system outputs from Table 1.

Prompt generation. We compare JewelExpert with human-written prompts and descriptions produced by RPG, Personalized PR, and GPT-4o. RPG and Personalized PR contribute only their prompt generation components in this comparison, with JewelDiffusion rendering the resulting descriptions. JewelExpert achieves the highest mean scores among the prompt sources, improving AF, CF, and AQ over human-written prompts by 3.7, 0.9, and 1.4 points, respectively. Its average score is 74.9, compared with 70.9 for GPT-4o, indicating that the choice of prompt source affects performance even when the image generator remains fixed.

Image generation. Each generator in Table 4 receives the same JewelExpert prompts. In particular, Qwen-Image and JewelDiffusion use matching resolution, sampling steps, and guidance, with one image generated and evaluated per benchmark entry. JewelDiffusion improves AF from 69.4 to 75.7, CF from 60.6 to 69.2, and AQ from 74.9 to 79.9, increasing the average from 68.3 to 74.9. Adapting Qwen-Image with the constructed training data improves all three criteria, with the largest gain in CF.

Compared with Nano-Banana, JewelDiffusion gains 9.5 points in AF and 2.1 points in AQ, while their CF scores are similar (69.2 versus 69.0). The largest improvement over this baseline is therefore in attribute fidelity.

Figure 4 illustrates several differences in the generated designs. For the plain bracelet, SDXL produces a tilted view despite the requested front view, while Qwen-Image and FLUX.1-dev introduce colored regions into the gold design. In the enamel example, SDXL depicts dark metal relief without the requested pastel enamel. Qwen-Image and FLUX.1-dev both produce openwork in the filigree example, with differences in the shape and arrangement of the motifs.



Figure 4: Qualitative comparison of the image generators in Table 4. The plain, enamel, and filigree rows each show a shared prompt and the corresponding outputs, with JewelDiffusion shown in the column labeled JewelCraft. The complete prompts are provided in Appendix C.4.

6 LIMITATIONS

Our experiments cover three crafts and one round of data construction and adaptation. The split is performed after augmentation without grouping by source seed, so variants of the same design may occur in training and evaluation. Generalization to independent source designs requires further study. The component comparisons evaluate prompt sources and image generators without isolating individual data construction steps or comparing JewelExpert with an unadapted Qwen2.5 control.

Gemini participates in both screening and evaluation, and AF aggregates craft, material, view, and lighting. Further analysis of craft accuracy and evaluator agreement would complement the reported means. The generated images are visual design proposals whose appearance does not establish physical manufacturability. Future work includes independent design evaluation, controlled studies of data construction, and preference optimization Black et al. (2023); Wallace et al. (2024).

7 CONCLUSION

We presented JEWEL-T2I, which constructs training data from expert jewelry assets for generation from text. JEWELGEN combines semantic and viewpoint editing with visual craft exemplars, followed by MLLM screening and designer verification. The selected data is used to adapt the image generator in JEWELCRAFT, while a domain prompt model converts structured requirements into descriptions. Experiments on JEWELBENCH show improvements over the image backbone in AF, CF, and AQ when prompts are held fixed. Evaluation across crafts and with both automated and human scoring distinguishes fidelity to design requirements from visual appeal. The results support using expert jewelry assets to construct training data for adapting general T2I models to jewelry design.

AI USE STATEMENT

Generative AI tools were used to revise the organization and language of the manuscript, to assist with writing code, and to check consistency in the reported experimental settings and summary scores. Apart from the generative and evaluation models described in the paper, we did not use generative AI tools for formulating research ideas or designing experiments. The authors take responsibility for the final content of this work, including AI-assisted text and code.

REFERENCES

Petra Ahde. Appropriation by adornments: Personalization makes the everyday life more pleasant. In *Proceedings of the 2007 conference on Designing pleasurable products and interfaces*, pp. 148–157, 2007.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 958–968, 2024.

Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7727–7736, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. Introducing Gemini 2.5 Flash Image, our state-of-the-art image model. Google Developers Blog, August 2025. URL `https://developers.googleblog.com/introducing-gemini-2-5-flash-image/`.

Ruyi Gan, Xiaojun Wu, Junyu Lu, Yuanhe Tian, Dixiang Zhang, Ziwei Wu, Renliang Sun, Chang Liu, Jiaxing Zhang, Pingjian Zhang, et al. idesigner: A high-resolution and complex-prompt following text-to-image diffusion model for interior design. *arXiv preprint arXiv:2312.04326*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ritika Jha, Aanisha Bhattacharyya, Yaman Singla, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. SPRO: Improving image generation via self-play. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2734–2744, 2025.

Michael D. Magee. Generative artificial intelligence as a tool for jewelry design. *Gems & Gemology*, 60(3), 2024. URL https://www.gia.edu/gems-gemology/fall-2024-artificial-intelligence-in-jewelry-design.

Fanqing Meng, Wenqi Shao, Lixin Luo, Yahong Wang, Yiran Chen, Quanfeng Lu, Yue Yang, Tianshuo Yang, Kaipeng Zhang, Yu Qiao, et al. Phybench: A physical commonsense benchmark for evaluating text-to-image models. *arXiv preprint arXiv:2406.11802*, 2024.

Yan Meng, Shun Yao, Li Yang, and Yufan Chen. Application of Stable Diffusion based generative artificial intelligence in jewelry design with intangible cultural heritage element: A case study of tie-dye. *Journal of Gems & Gemmology*, 27(5):98–105, 2025. doi: 10.15964/j.cnki.027jgg.2025.05.012.

C Page Moreau, Emanuela Prandelli, Martin Schreier, and Silke Hieke. Customization in luxury brands: can valentino get personal? *Journal of Marketing Research*, 57(5):937–947, 2020.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Anna Pozzobon. Personalization of jewelry products: an added value or loss in recognition for luxury brands? 2021.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. SILMM: Self-improving large multimodal models for compositional text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.05818.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mcleavey, Ilya Sutskever, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024b.

Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7017–7026, 2024c.

Mingzhe Yu, Yunshan Ma, Lei Wu, Changshuo Wang, Xue Li, and Lei Meng. Fashiondpo: Fine-tune fashion outfit generation model using direct preference optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 212–222, 2025.

Xulu Zhang, Xiaoyong Wei, Jinlin Wu, Jiaxin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. Generating on generated: An approach towards self-evolving diffusion models. *arXiv preprint arXiv:2502.09963*, 2025. URL https://arxiv.org/abs/2502.09963.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4606–4615, 2023.