CREATCTRL: CONTROLLABLE CREATIVE GRAPHIC DESIGN GENERATION FROM REFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Creativity is essential in graphic design. Despite recent advances in graphic design generation, controlling the creativity of generated designs remains unexplored. In this paper, we investigate the problem of creativity control for graphic design generation, and propose a model, *CreatCtrl*, which learns to elevate the creativity of reference designs in a controllable manner. Our model takes as input a textual design intention, a reference design in raster format, and a continuous creativity parameter, and generates more creative, layered graphic designs. Our model enables users to control the creativity degree of the generated designs by adjusting the creativity parameter in a continuous way, which allows them to explore a wide range of creative variants of the reference to find design inspiration. To train our model, we propose a dataset construction method to synthesize training tuples, each consisting of a reference design, its creative variant, and a parameter representing the creativity gap between them, a reliability weight of the creativity parameter. We show the superior performance of our method on design quality, creativity control and creativity inspiration, demonstrating the potential of our creativity-controllable generation paradigm for real-world design applications.

1 INTRODUCTION

Creativity is the driving force behind successful graphic designs, which allows designers to create unique visuals that stand out, resonate with audiences, tell compelling stories and leave a lasting impression. In the design process, to seek inspiration for creative design ideas, designers usually gather visual *references*—mostly existing designs —and explore possible variations around the references to spark their creative thinking. Furthermore, designers would browse a range of variations with various degrees of creativity with respect to the reference before identifying the most ideal one. However, generating such creative design variations itself is expensive and challenging, which often requires navigating through a vast solution space of imagery, color, layout, and typography, and demands considerable technical expertise and artistic sense. Therefore, generative models with abilities to synthesize designs with *controllable creativity* could be of great value.

Recent advances in generative models have enabled remarkable progress in automatic graphic design generation. However, existing design generation methods primarily focus on unconditional generation Yamaguchi (2021); Yang & Cao (2025) or conditional generation from text descriptions Song et al. (2026); Jia et al. (2024); Inoue et al. (2024); Chen et al. (2025); Wang et al. (2025); Qu et al. (2025), lacking the ability to control the creativity of generated designs. Some image editing methods Ye et al. (2023); Ruiz et al. (2023); Parihar et al. (2026); Xu et al. (2025); Gal et al. (2022); Meng et al. (2022) support generating variations of an input reference image, but offer no mechanism to control the creativity of the variations.

In this paper, we study creativity-controllable generation of graphic designs, and present a model, named *CreatCtrl*, for transforming a reference design into more creative alternatives in a creativity-controllable manner. As shown in Figure 1, given a reference design, a textual prompt, and a creativity control parameter, our model can produce design variations that remain semantically aligned with the reference while exhibiting progressively enhanced creativity as the control parameter increases. Our model allows users to navigate through the space of creativity scales, exploring a rich set of creative variants to find inspirational design ideas. For training our model, we develop a dataset construction pipeline, which leverages a vision-language model (VLM), fine-tuned on rating

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.

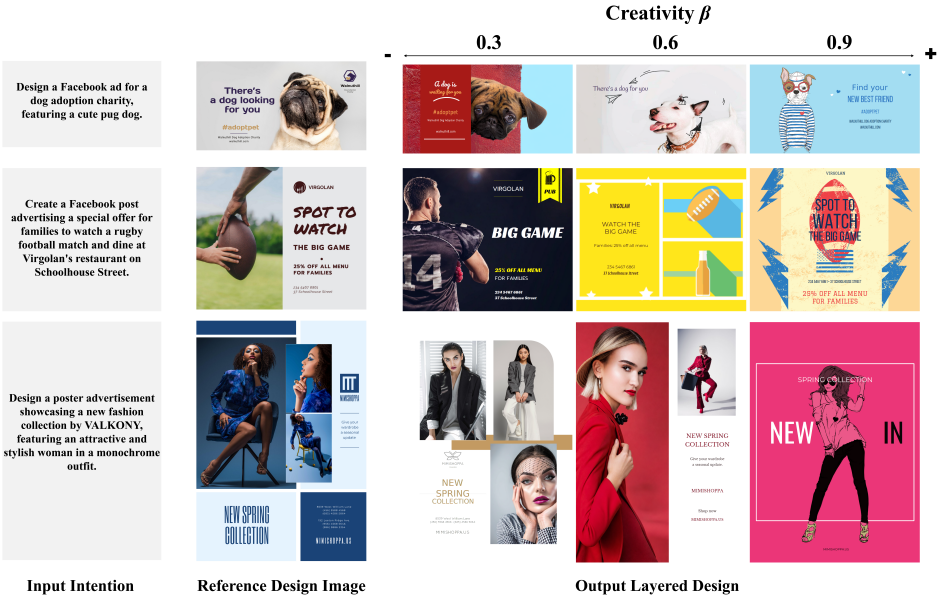


Figure 1: Given an input design intention, a reference design image, our method can generate layered graphic designs that are related to the reference and more creative than it. Notably, the creativity degree of generated designs can be controlled through a continuous parameter $\beta \in [0, 1]$. A higher β indicates higher creativity.

graphic design creativity, to generate synthetic training samples. Each sample comprises: an input reference design, an output creative variant, and a continuous parameter that reflects the discrepancy between the input and output in creativity. We conduct experiments, comparing our model with several baselines. The results show that our model generates high-quality design variations with good intention alignment and offers great controllability of creativity. Furthermore, our model can produce creative variations that better spark human design creativity compared to the baselines.

Our contributions are summarized as follows: 1) we make the first attempt to investigate controllability of creativity in graphic design generation; 2) we propose a model to generate layered variants of a reference design image in a creativity-controllable manner; 3) we build a pipeline to synthesize data for training reference-based, creativity-controllable generative models of graphic designs.

2 Related Work

2.1 Graphic Design Generation

The community has seen a rising interest in building generative models for graphic design. CanvasVAE Yamaguchi (2021) learns a distribution over layered designs described as sets of canvas and element attributes. GOL Yang & Cao (2025) learns the order of design elements to optimize the performance of models that generate a design as a sequence of discrete element attribute tokens. Recent efforts have proposed different paradigms for text-to-design generation. For example, several works utilize fine-tuned text-to-image diffusion model to generate pixel-based designs that, unfortunately, lack editability Wang et al. (2025); Chen et al. (2025). Some systems such as COLE and OpenCOLE Jia et al. (2024); Inoue et al. (2024) decompose the generation process into a set of sub-tasks, each of which is tackled by a specialized model (large language model, large multimodal model or image diffusion model). ART Pu et al. (2025) formulates design generation as a task of multi-layer transparent image generation and learns to generate all layers simultaneously. Recently, some works try to first generate a design image using a powerful pre-trained text-to-image model, and then decompose the image into layers to enable editability. IGD Qu et al. (2025) combines a large language model with an image diffusion model generate graphic designs as text-image docu-

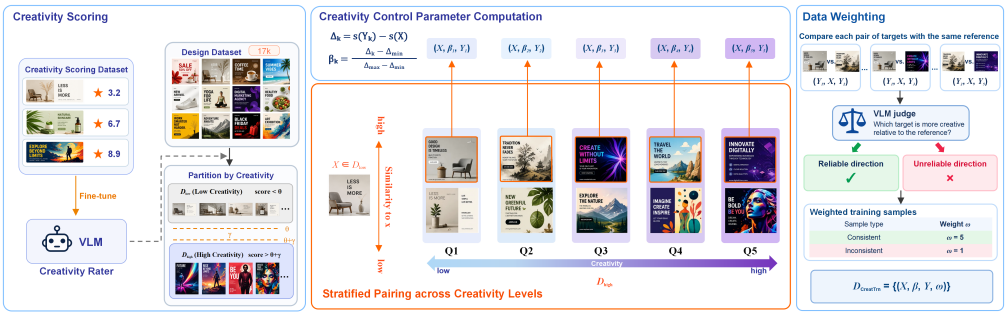


Figure 2: Overview of the training data generation.

ments. However, to our best knowledge, there is no work discussing how to control the creativity of generated design, which is investigated in our paper.

2.2 Image Variation

There is a line of work on image variation Meng et al. (2022); Ruiz et al. (2023); Gal et al. (2022); Parihar et al. (2026); Xu et al. (2025), which can generate variations of a reference image that share the important content with the reference, while being different in other aspects. While such image variations perceptually differ from the reference, there is no guarantee that such difference monotonically maps to creativity degree. Thus, the previous image variation methods are not well suited for creativity-controllable design generation, where intuitive control of creativity level is needed.

3 Method

We aim to train a generative model that, given a reference design, a textual design intention, and a continuous creativity control parameter, produces a more creative design whose content is semantically related to the reference and whose creativity is aligned with the creativity parameter. We first describe our method for synthesizing training data, and then introduce our generative model.

3.1 Training Data Generation

Each sample in the generated training data takes the form (X, β, Y, ω), where X is a reference design, Y is the creative target, β is a continuous creativity control parameter that reflects how creative Y relative to X, and ω indicates the reliability of β. Figure 2 illustrates our data generation pipeline.

Creativity-based partitioning. We first develop a model to predict the creativity ratings of designs, and use it to partition a dataset of graphic designs into a low-creativity set and a high-creativity set. In particular, we collect a dataset of 6k graphic designs from *Awwwards*[1] , with creativity ratings from human designers on a scale of 0 to 10. We split the dataset into 5,400 training samples and 600 test samples, and train our creativity scoring model on the training set. We build the creativity scoring model by adding a lightweight regression head on top of the output embedding of a vision-language model (VLM), Qwen3-VL-30B Bai et al. (2025), which takes as input a design image and a text prompt designed to instruct the VLM to predict a creativity score for the input design. We use a L_1 loss for training, where the VLM parameters are fine-tuned and the regression head is trained from scratch. Our model yields a mean absolute error of 0.50 on the test set.

The scoring model is then applied to a large corpus of 17,369 graphic designs, 8,610 filtered designs from Crello Yamaguchi (2021) and 8,759 designs from two additional sources: 2,200 designs from PSDesigner Shuai et al. (2026) and 6,559 designs from *Freepik*[2]. This yields a creativity score

[1] https://www.awwwards.com
[2] https://www.freepik.com

$s(X) \in \mathbb{R}$ for every design X. We denote this rated corpus, with 17k+ graphic designs and corresponding creativity scores, as $\mathcal{D}$. To test the performance of our model on $\mathcal{D}$, we apply our model to all the designs, and rank them in descending order of the predicted creativity scores. Then, we take the top 20%, middle 20% and bottom 20% of the ranked designs as high-creativity set $\mathbf{S}_{\text{high}}$, medium-creativity set $\mathbf{S}_{\text{medium}}$, low-creativity set $\mathbf{S}_{\text{low}}$, respectively. Finally, we construct design triplets of the form $(x_{\text{h}}, x_{\text{m}}, x_{\text{l}})$, where $x_{\text{h}} \in \mathbf{S}_{\text{high}}$, $x_{\text{m}} \in \mathbf{S}_{\text{medium}}$, $x_{\text{l}} \in \mathbf{S}_{\text{low}}$. That is, the creativity ranking of x_{h}, x_{m}, x_{l} according to the model is: $x_{\text{h}} > x_{\text{m}} > x_{\text{l}}$. We conduct a human study, where subjects are asked to rank three designs in a triplet according to creativity. Our model achieves a model-human consistency rate of 0.71, confirming its effectiveness in rating creativity of designs in $\mathcal{D}$.

We then sort designs in $\mathcal{D}$ in descending order of creativity scores, and split the ranked designs into three groups with a split ratio of 5:3:1. We take the top $\frac{5}{9}$ as a creative target pool $\mathcal{D}_{\text{high}}$, and the bottom $\frac{1}{9}$ as a reference pool $\mathcal{D}_{\text{low}}$, while discarding the middle $\frac{3}{9}$. This creates a large creativity gap between designs in $\mathcal{D}_{\text{low}}$ and $\mathcal{D}_{\text{high}}$.

Stratified pairing across creativity levels. We further partition the ranked designs in $\mathcal{D}_{\text{high}}$ into 5 groups $\{\mathcal{Q}_k\}_{k=1}^{5}$, so that $\mathcal{Q}_k$ contains designs whose scores fall within the top $20(k-1)\%$ to $20k\%$ of all the ranked scores.

For each reference $X \in \mathcal{D}_{\text{low}}$, we form one training pair per group, by retrieving its most similar design from each $\mathcal{Q}_k$, based on cosine similarity between image embeddings from the vision encoder of the fine-tuned Qwen-VL in our creativity scoring model. The fine-tuned vision encoder can extract visual features more relevant to graphic designs, thus allowing us to better measure design similarity, compared to the pre-trained one. Each reference is thus associated with targets $\{Y_1, \ldots, Y_5\}$, resulting in 5 reference-target pairs $\{(X, Y_k)\}_{k=1}^{5}$. Note that our stratified pairing strategy ensures that the targets in the resulting pairs share similar content with the reference, while spanning a large range of creativity with respect to the reference.

Creativity control parameter computation. For each pair (X, Y_k), we define a creativity control parameter β_k as the normalized creativity score gap. Specifically, let $\Delta_{X,Y_k} = s(Y_k) - s(X)$ be the creativity score difference between designs X and Y_k, and $\Delta_{\min}$ and $\Delta_{\max}$ be the minimum and maximum score differences across all the 5 pairs, respectively. We then define β_k as:

$$\beta_k = \frac{\Delta_{X,Y_k} - \Delta_{\min}}{\Delta_{\max} - \Delta_{\min}} \in [0, 1]. \tag{1}$$

Subtracting $\Delta_{\min}$ ensures that β_k's span the full $[0, 1]$ range and that $\beta_k = 0$ corresponds to the smallest creativity gap actually present in training data. Smaller β_k indicates mild change; larger β_k indicates more substantial transformation. At inference, β is exposed to the user as a continuous control parameter. After computing a control parameter β for each pair (X, Y), we obtain a dataset of 10491 triplets $\mathcal{D}_{\text{CreatTrm}} = \{(X, \beta, Y)\}$, representing creative transformations at various degrees.

Data weighting. We further weight each sample (X, β, Y) in $\mathcal{D}_{\text{CreatTrm}}$ according to how well the β can parameterize the creativity degree of Y with respect to X—larger β corresponds to greater creativity and vice versa. In particular, we first create a set of tuples (y_0, x, y_1) from $\mathcal{D}_{\text{CreatTrm}}$, where x is a reference design and (y_0, y_1) are two target designs paired with x in $\mathcal{D}_{\text{CreatTrm}}$. y_0 and y_1 have different β values, and the design with a larger β value is assumed to be more creative than the one with a smaller β. We end up with a set D_{test} of 20k test tuples, where each test tuple in D_{test} involves two triplets in $\mathcal{D}_{\text{CreatTrm}}$. For each test tuple (y_0, x, y_1), a VLM (Qwen3-VL Plus) is given the three design images, and prompted to choose which of y_0 and y_1 is more creative relative to x. The test tuple is labeled as "positive" if the VLM choice agrees with the β-based choice, and "negative" otherwise. A triplet in $\mathcal{D}_{\text{CreatTrm}}$ is labeled as "reliable" only if all the test tuples that it participates in are labeled as "positive". The reliable triplets are given a weight ω of 5, while other triplets are assigned a weight ω of 1. This gives rise to a weighted dataset: $\mathcal{D}_{\text{CreatTrm}} = \{(X, \beta, Y, \omega)\}$. During training, loss terms corresponding to reliable triplets will be multiplied by their weights, to increase the contribution of the reliable triplets to training. $\mathcal{D}_{\text{CreatTrm}}$ is partitioned into a training split $\mathcal{D}_{\text{CreatTrm-train}}$ (9441 samples), and a test split $\mathcal{D}_{\text{CreatTrm-test}}$ (1050 samples).

To test the reliability of the VLM choices, we compute how often VLM choices agree with human choices on a small subset of D_{test}. The subset, containing 200 test tuples, is constructed so that β values in it span the full range between 0 and 1. The VLM achieves 82% agreement with humans.

3.2 REFERENCE-BASED, CREATIVITY-CONTROLLABLE DESIGN GENERATOR

Our design generator is an autoregressive Transformer decoder that produces a target design , represented as a sequence of element attributes, conditioned on the reference preview, the textual intention, and the creativity control parameter β.

Design representation. A graphic design X is composed of a canvas and a set of elements that are categorized into four types: text, image, SVG, and background. A text element is represented as $(c, x, y, w, h, z, f_f, f_c, f_s)$, where c denotes its category, (x, y, w, h) denotes its bounding box with center (x, y) and size (w, h), z is the layer order, and (f_f, f_c, f_s) denotes font family, color, and size. Note that the text content is not included, which is generated by a downstream language model during rendering (introduced later). An image or SVG element is represented as (c, x, y, w, h, z, u), where u denotes visual content encoded by the visual embedding from CLIP-ViT-H-14 Radford et al. (2021), and a background element is represented as (c, x, y, w, h, z, r), where r is the background color. Following prior work on graphic design generation Yang & Cao (2025), we serialize each design into a sequence of discrete tokens. Continuous attributes (x, y, w, h, u) are quantized via K-means. Each element is represented as a subsequence prefixed by its category token ([TXT], [IMG], [SVG], [BG]); the canvas is prefixed by [CAV]. The full sequence is formed by concatenating the subsequences into a sequence, and prepending discretized canvas size tokens to the sequence, and adding [BOS] and [EOS] to the start and end of the resulting sequence, respectively.

Architecture. The generator takes three input conditions. *Reference design.* The image I of the reference design X is encoded by CLIP-ViT-H-14, and we take the penultimate layer features as $\mathbf{f}_{\text{ref}}$. *Textual intention.* To encode the intention text $\mathcal{P}$ (e.g., "a minimalist poster with warm colors"), we employ a dual-encoder design, concatenating penultimate layer embeddings from CLIP-ViT-L-14 and CLIP-ViT-bigG-14 to form $\mathbf{f}_{\text{intent}}$. *Creativity control parameter.* The scalar $\beta \in [0, 1]$ is first processed by an MLP, and then injected into the generator via adaptive layer normalization (AdaLN), similar to timestep in diffusion transformers (DiTs) Peebles & Xie (2023).

The reference and intention features are fused into a multimodal conditioning vector via an MLP: $\mathbf{c} = \text{MLP}([\mathbf{f}_{\text{ref}}; \mathbf{f}_{\text{intent}}])$. Following IP-Adapter Ye et al. (2023), we inject $\mathbf{c}$ through cross-attention (CrossAttn) inserted into each Transformer block.

Training. We adopt a two-stage training strategy.

Stage 1: Vectorization Pretraining. In the first stage, we pretrain the generator on a layered design dataset with 27k samples, including 10k samples from HQ-Poster100K Chen et al. (2025) and 17k samples from $\mathcal{D}$ introduced in Section 3.1. For each design D, we feed its own image I as the conditioning input and require the model to reconstruct its full structured sequence S^X:

$$\mathcal{L}_{\text{PT}} = -\sum_{X} \sum_{t=2}^{T} \log p_\theta \left(S_t^X \mid S_{1:t-1}^X, I \right). \tag{2}$$

This stage trains the model to output structured, layered design representations effectively.

Stage 2: Creativity-conditioned Fine-tuning. In the second stage, we fine-tune the pretrained generator on the constructed training triplets $\mathcal{D}_{\text{CreatTrm-train}} = \{(X, \beta, Y, \omega)\}$, with the objective:

$$\mathcal{L}_{\text{FT}} = -\sum_{(X,\beta,Y)} \sum_{t=2}^{T} \omega \log p_\theta \left(S_t^Y \mid S_{1:t-1}^Y, I, \mathcal{P}, \beta \right), \tag{3}$$

where I is X's image, and $\mathcal{P}$ is Y_k's design intention. Across each reference's triplets, the targets span various creativity levels parameterized by different β values, which enables the model to learn to move its output along the creativity axis as β varies.

Design rendering. Once the generator outputs a design sequence, we render the elements through element-specific branches and compose the results into a layered, editable output. *Background.* For a background element, we directly fill its bounding box with its RGB color. *Image.* For an image element, we feed its visual embedding (the bin center of its content token) into the pretrained SDXL Podell et al. (2023) via IP-Adapter Ye et al. (2023), to generate an image at the size of the element's bounding box. The generated image is then positioned on the canvas according to the bounding box. However, we empirically found that, when the predicted embedding falls out of the

input distribution of the SDXL, the quality of generated images degrades. To ensure the quality of images, we adopt a *generate-or-retrieve* strategy (similar to Koh et al. (2023)), where we train a decision model on top of the visual embedding to decide whether to generate or retrieve an image. When the model predicts that an image should be retrieved for a visual embedding, we directly retrieve an image from a dataset of existing images. For more details of the decision model, see the supplementary material. *SVG.* Since existing image generation models struggle with generating high-quality, clean SVG images, for an SVG element, we adopt a retrieval-based approach to obtain its image, by matching its visual embedding (the bin center of its content token) against those of SVG images in a curated dataset via cosine similarity. The best-matching SVG image is returned and placed on the canvas according to the bounding box. *Text.* For a text element, our model only predicts where text should appear and how it should be stylized. We leverage an LLM to generate its text content. See the supplementary material for more details.

4 Experiments

4.1 Experimental Setup

Implementation Details. For pre-training in Stage 1, we train our generator for up to 100 epochs, with learning rate 4×10^{-5}, weight decay 0.2, dropout 0.3, and batch size 128. For fine-tuning in Stage 2, the generator is trained for up to 200 epochs, with learning rate 5×10^{-5}, weight decay 0.1, dropout 0.2, and batch size 144. For creativity parameter adherence, we enable classifier-free guidance with guidance scale 3.0 Sanchez et al. (2023), by dropping the input creativity parameter 15% of the time during training. Both stages use the AdamW with 10 warmup epochs followed by cosine decay, gradient clipping at 1.0, label smoothing 0.1, an EMA of the weights with decay 0.999. We evaluate our method on the test split $\mathcal{D}_{\text{CreatTrm-test}}$ of our generated dataset in Section 3.1. For each input reference design, three creative variations are generated using three different β values: 0.3, 0.6, 0.9.

Metrics. We employ the following metrics:

- *FID.* We render each generated design as an image and extract features using a Masked Autoencoder (MAE) trained on the Crello dataset to compute FID (against the *target* designs in the test set), which measures overall design quality.
- *Density and Coverage.* Following Yang & Cao (2025), we use Density and Coverage Naeem et al. (2020) to measure design fidelity and diversity, respectively. The metrics use the same design feature space as the FID above. The two metrics are computed against the *target* designs in the test set.
- *CLIP-Score.* We compute alignment between designs and their textual intentions using DesignCLIP in Song et al. (2026).
- *C-FID$_\beta$.* To measure creativity control, we compute FID between the original *reference* designs in the test set and the generated designs using β as C-FID$_\beta$. Note that C-FID$_\beta$ needs to be interpreted differently from traditional FID: *higher* C-FID$_\beta$ indicates better creativity as we aim to generate creative variants that deviate from the original references.

Following COLE Jia et al. (2024), we also use a multimodal large language model (*MLLM*), GPT-6 Astra, as a proxy for human evaluation of design quality, to rate generated designs on a scale of 1 to 10 across 5 different dimensions.

Baselines. We compare against three baselines:

- *Nearest Neighbor Search (NNS).* Given the input reference image, we retrieve a design image whose similarity to the reference is closest to $1 - \beta$, from a database of designs (the creative target pool $\mathcal{D}_{\text{high}}$ introduced in Section 3.1). To compute image similarity, we take the fine-tuned OpenCLIP image encoder Radford et al. (2021) in DreamSim Fu et al. (2023), and further fine-tune it on design-intention pairs from Crello, by optimizing a contrastive loss while keeping the text encoder frozen. Cosine similarity between embeddings from the resulting image encoder is used to measure image similarity.
- *Text-based Image Editing (TIE).* We first feed the input reference image and the design intention into Qwen3-VL Bai et al. (2025), and query it to generate a prompt describing image

Table 1: Quantitative comparison of different methods. The MLLM-based evaluation considers 5 dimensions: (1) layout, (2) content relevance, (3) typography and color, (4) graphics and images, (5) innovation. The best and second best results are in bold and underlined, respectively. "Ref."refers to the original reference designs.

	Design Quality			Text Align	Creativity Control			MLLM				
					C-FID$_\beta$↑							
Method	FID↓	Dens.↑	Cov.↑	CLIP↑	$\beta = 0.3$	$\beta = 0.6$	$\beta = 0.9$	(1)	(2)	(3)	(4)	(5)
Ref.	–	–	–	21.26	–	–	–	–	–	–	–	–
NNS	<u>27.95</u>	**1.09**	<u>0.52</u>	17.47	<u>47.83</u>	55.37	99.14	**7.44**	3.07	**7.35**	**7.45**	<u>5.68</u>
TIE	52.33	<u>0.99</u>	0.44	<u>17.72</u>	18.12	50.81	68.50	5.04	3.35	3.64	5.39	4.07
LMM	48.54	0.94	0.49	17.70	16.23	<u>65.54</u>	**132.36**	5.13	<u>3.73</u>	4.27	6.19	4.75
Ours	**27.87**	0.97	**0.73**	**19.82**	**55.44**	**90.12**	<u>115.36</u>	<u>5.92</u>	**5.43**	<u>5.76</u>	<u>7.05</u>	**6.16**

Table 2: Results of the user study comparing our method against baselines in terms of visual quality, reference relevance, creativity control, and creativity inspiration. The best and second best results are in bold and underlined, respectively.

Method	Visual Quality↑	Ref. Relevance↑	Creativity Control↑	Inspiration↑
NNS	<u>34.7%</u>	8.0%	0.173	4.3%
TIE	6.7%	<u>96.0%</u>	0.227	15.7%
LMM	16.0%	**98.0%**	**0.867**	<u>20.4%</u>
Ours	**42.7%**	92.0%	<u>0.307</u>	**59.6%**

editing operations that transform the reference design to be more creative. The prompt is then input into a text-based image editing model, Kontinuous Kontext Parihar et al. (2026), to generate creative outputs. Kontinuous Kontext exposes a continuous parameter between 0 and 1, which controls edit strength. The parameter is regarded as β: larger edit strength leads to more creative outputs.

- *Large Multimodal Model (LMM).* We leverage Qwen-Image-Edit-2511 Wu et al. (2025), whose inputs include: the reference design image, a text prompt including the design intention and a task instruction conditioned on β (a smaller β requests a slightly more creative variant, while a larger β requests a substantially reimagined design that preserves the core message). The design image generated by Qwen-Image-Edit-2511 is then processed by Qwen-Image-Layered Yin et al. (2025) to obtain a layered representation.

4.2 Quantitative Results

Table 1 reports quantitative results. Our method achieves the best FID and Coverage scores, suggesting that our method is able to generate high-quality and diverse output designs. Furthermore, the CLIP-Score value of our method is better than those of the baselines, and is on par with that of the reference designs. This indicates that our method can well maintain alignment with textual design intentions when creating creative variations. Compared to other methods, our method yields significantly higher C-FID$_\beta$ at $\beta = 0.3$ and $\beta = 0.6$, and the second highest C-FID$_\beta$ at $\beta = 0.9$, which confirms its superiority in creating creative outputs at different creativity scales. We note that, C-FID$_\beta$ of our method monotonically increases with higher β values, which highlights the effectiveness of β in controlling creativity degree. In the MLLM-based evaluation, our method yields the best results on content relevance and innovation, which again underscores the strong ability of our method in design variations that are both intention-aligned and creative.

4.3 User Study

We conduct a user study with 10 professional graphic designers to evaluate different methods for visual quality, reference relevance, creativity control, and creativity inspiration. We use 100 reference designs sampled from the test set. For each design, one of the four methods (NNS, TIE, LMM, Ours) generates three output designs at $\beta = 0.3, 0.6, 0.9$.

The results are shown in Table 2. *Visual quality.* The participants are shown four designs generated by our method and the three baselines from the same input reference and identical β, and asked to select which one has the best visual quality. Our method receives the highest preference rate, surpassing all the baselines. Notably, our method surpasses NNS, even though NNS directly returns real designs, which confirms the high visual quality of our generations. *Reference relevance.* The participants are given a reference design and three designs generated by one method at three different β values, and asked to determine if all the generated designs are relevant to the reference. Our method yields a high reference relevance rate of 92.0%, on par with LMM (98.0%) and TIE (96.0%), suggesting that our creative generations are still related to the input reference. *Creativity control.* The participants are presented with a reference design, along with a set of three designs generated by one method at three different β values, and asked to rank the designs by perceived creativity gap from the reference. Each set is ranked by an average of 10 participants, with an average inter-participant consistency rate of 0.82. We measure controllability by agreement between human and β-based rankings, measured by Kendall rank correlation coefficient (or Kendall's $\tau \in [-1, 1]$). Our method attains the second highest τ among all methods, confirming that β provides effective control over creativity. LMM obtains the highest τ, because it applies increasingly aggressive stylization as β grows (shown later in Section 4.4), which makes it easier for the participants to order the perceived creativity gaps. *Creativity inspiration.* The participants are shown a reference design, and four sets of generated designs. Each set is generated by one of the four methods, containing three creative variations at $\beta = 0.3, 0.6, 0.9$. The participants are asked to choose which set is the most useful for inspiring their design creativity and provide a detailed explanation for the choice. Our method is selected 59.6% of the time, significantly outperforming the other methods. This indicates that our method is able to offer the most inspiring design variations for designers.

4.4 Qualitative Results

Figure 3 shows some examples of the generated designs used in our user study (Section 4.3). NNS is limited by its retrieval pool. Its generated variants drift from the reference theme at larger β: The designers commented its results as *"high-quality but irrelevant designs"* that *"look more like randomly retrieved templates than creative variants of the reference."*. TIE and LMM are frequently criticized by the designers for distorted text that *"severely degrades visual quality"*. Furthermore, the generated variations of TIE at different β often look similar, with small differences (e.g., in the bottom-left and bottom-right examples), which reveals its inferior creativity control capability. LMM enables effective creativity control. Its results focus on altering the visual style of the main subject in the reference design, with limited changes in layout and color scheme (e.g., in the top-right and bottom-right examples). The designers found LMM's changes to the reference superficial, which limits its practical usage for creativity inspiration: in the middle-right example, the generated variations are *"more like random visual perturbation than effective changes that inspires designers"*; in the bottom-left example, *"creativity mostly adds borders, flowers, and handwritten fonts"*; in the bottom-right example, *"innovation mainly involves style change, without forming inspiring layouts"*. In contrast, our method can *"turn the reference into executable new design directions"* that vary numerous aspects including imagery, layout, typography and color, while preserving the core message.

5 Conclusion

In this paper, we make an attempt to study controllable creativity for graphic design generation, and introduce a novel method for generating creative variants of a reference design whose creativity can be adjusted in a continuous manner. Our experiments show that our method generates high-quality design variations that are well aligned with the design intention the design intention, while offering effective control over creativity. We also show that our generated outputs can effectively spark the design creativity of designers. We hope our work can inspire future work on generative creativity modeling in the design domain.



Figure 3: Visual comparison of different methods.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

SiXiang Chen, Jianyu Lai, Jialin Gao, Tian Ye, Haoyu Chen, Hengyu Shi, Shitong Shao, Yunlong Lin, Song Fei, Zhaohu Xing, Yeying Jin, Junfeng Luo, Xiaoming Wei, and Lei Zhu. Postercraft: Rethinking high-quality aesthetic poster generation in a unified framework, 2025. URL https://arxiv.org/abs/2506.10741.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. URL https://arxiv.org/abs/2306.09344.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. Opencole: Towards reproducible automatic graphic design generation, 2024. URL https://arxiv.org/abs/2406.08232.

Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, Shanghang Zhang, and Baining Guo. Cole: A hierarchical generation framework for multi-layered and editable graphic design, 2024. URL https://arxiv.org/abs/2311.16974.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models, 2023. URL https://arxiv.org/abs/2305.17216.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. URL https://arxiv.org/abs/2108.01073.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020. URL https://arxiv.org/abs/2002.09797.

Rishubh Parihar, Or Patashnik, Daniil Ostashev, R. Venkatesh Babu, Daniel Cohen-Or, and Kuan-Chieh Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing, 2026. URL https://arxiv.org/abs/2510.08532.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

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, 2023. URL https://arxiv.org/abs/2307.01952.

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, Lin Liang, Lijuan Wang, Ji Li, Xiu Li, Zhouhui Lian, Gao Huang, and Baining Guo. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7952–7962, June 2025.

Yadong Qu, Shancheng Fang, Yuxin Wang, Xiaorui Wang, Zhineng Chen, Hongtao Xie, and Yongdong Zhang. Igd: Instructional graphic design with multimodal layer generation, 2025. URL https://arxiv.org/abs/2507.09910.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance, 2023. URL https://arxiv.org/abs/2306.17806.

Xincheng Shuai, Song Tang, Yutong Huang, Henghui Ding, and Dacheng Tao. Psdesigner: Automated graphic design with a human-like creative workflow, 2026. URL https://arxiv.org/abs/2603.25738.

Xinya Song, Bo Yang, and Ying Cao. Harnessing layered graphic designs with real intentions for text-to-design generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4729–4738, 2026.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models, 2025. URL https://arxiv.org/abs/2503.01645.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Zhenyu Xu, Xiaoqi Shen, Haotian Nan, and Xinyu Zhang. Numerikontrol: Adding numeric control to diffusion transformers for instruction-based image editing, 2025. URL https://arxiv.org/abs/2511.23105.

Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents, 2021. URL https://arxiv.org/abs/2108.01249.

Bo Yang and Ying Cao. Order matters: Learning element ordering for graphic design generation. *ACM Trans. Graph.*, 44(4), July 2025. ISSN 0730-0301. doi: 10.1145/3730858. URL https://doi.org/10.1145/3730858.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Shengming Yin, Zekai Zhang, Zecheng Tang, Kaiyuan Gao, Xiao Xu, Kun Yan, Jiahao Li, Yilei Chen, Yuxiang Chen, Heung-Yeung Shum, Lionel M. Ni, Jingren Zhou, Junyang Lin, and Chenfei Wu. Qwen-image-layered: Towards inherent editability via layer decomposition, 2025. URL https://arxiv.org/abs/2512.15603.