Master's Project Final Report

Jiayi Li
Brown University
Project Advisor: Prof. David H. Laidlaw
May 14, 2026

Project Report and Author Contributions

This report summarizes my contributions to three research efforts completed during this academic period: (1) a collaborative thematic diagram generation project, (2) a collaborative pictorial chart generation project, and (3) an independent project on pixel art multi-view generation. For each project, I list the paper authors, paper status, and my role.

Paper 1: Skeleton-Guided Generation and Editing of Thematic Diagrams
Authors: Shishi Xiao, Jiayi Li, Yuqiao Guan, Yiyang Zhang, David H. Laidlaw.
Status: In review at SIGGRAPH Asia 2026.
My Role: My contributions focused on dataset construction, evaluation design, and supporting the overall system development for structure-aware infographic generation. I designed and implemented a multi-stage dataset pipeline, beginning with the generation of over 7,000 skeleton structures using an LLM-based prompting strategy, followed by filtering to ensure structural validity, including removing distorted layouts, overlapping elements, and invalid node connections. This dataset was later expanded into over 16,000 high-quality skeleton–infographic pairs.

I also contributed to the evaluation pipeline for structural accuracy. I developed a segmentation-based workflow using SAM2 and SAM3 to detect node regions, combined with background removal using BiRefNet and inpainting using LaMa to isolate backbone structures. Based on these components, I implemented a structure-aware F1 metric for evaluating node detection and backbone alignment in generated infographics.

In addition, I contributed to writing sections of the paper, particularly those related to dataset construction and evaluation methodology. I also conducted literature analysis on related generative modeling approaches, including TokenVerse, Mod-Adapter, and XVerse, and helped position the work within the broader research landscape.

Paper 2: ChArtist: Generating Pictorial Charts with Unified Spatial and Subject Control
Authors: Shishi Xiao, Tongyu Zhou, David H. Laidlaw, Gromit Yeuk-Yin Chan.
Status: Accepted to CVPR 2026.
My Role: My contributions focused on designing quantitative evaluation metrics for measuring structural accuracy in generated charts. For pie charts, I proposed a ray-based method to estimate sector angles directly from image data, improving robustness compared with standard geometric approximations. For line charts, I designed a distance-field and sampling-based approach, inspired by point cloud analysis, to evaluate alignment between generated outputs and target chart structures.

Paper 3: SpriteForge: Consistent Multi-View Pixel Art Generation via LoRA-Enhanced Image Conditioning

Authors: Jiayi Li.
Status: Independent research report.
My Role: This project is a fully independent research effort in which I was responsible for problem formulation, dataset construction, model design, training, evaluation, and paper writing. I identified a key mismatch between existing multi-view generation approaches and the pixel art domain, and reformulated the task as learning view-dependent transformations rather than enforcing strict 3D geometric consistency.

I designed and implemented the dataset pipeline, including collecting and curating multi-view pixel art assets, enforcing consistent orientation conventions, standardizing transparent backgrounds and discrete color palettes, and processing spritesheets into aligned multi-view training samples.

On the modeling side, I explored several approaches and iteratively refined the method. I first evaluated ControlNet-based structural guidance and identified its limitations for low-resolution pixel art due to unreliable control signals. Based on this analysis, I pivoted to a conditioning-based LoRA fine-tuning approach on a pre-trained image-conditioned diffusion model. I implemented the training and inference pipeline and evaluated multi-view consistency, focusing on orientation correctness, structural alignment, and color palette preservation.

Overall, this project demonstrates my ability to independently define a research direction, critically evaluate alternative approaches, and carry a system from initial idea through implementation, evaluation, and documentation.

Skeleton-Guided Generation and Editing of Thematic Diagrams

SHISHI XIAO, Brown University, United States of America
JIAYI LI, Brown University, United States of America
YUQIAO GUAN, Brown University, United States of America
YIYANG ZHANG, Brown University, United States of America
DAVID H. LAIDLAW, Brown University, United States of America

Diagrams are widely used to communicate complex ideas through explicit structures, yet their visual appearance often lacks engagement and memorability and remains weakly connected to the themes they represent. We introduce thematic diagrams, which transform abstract structures into stylized diagrams that reflect a given theme, enabling visual storytelling both structurally and semantically. To help users specify desired structures accurately and easily, we propose a skeleton representation that abstracts diagrams into nodes and edges. Based on this representation, our framework unifies two tasks: generating thematic diagrams from skeletons and editing diagram structures while preserving style. Our model introduces semantic-aware positional encoding to bind visual appearance to structural roles, and analogy-aware attention to route information between reference and target conditions while avoiding direct copying. Additionally, we introduce a new dataset of 10k skeleton–thematic diagram pairs. Experiments demonstrate that our method generates visually compelling thematic diagrams, preserves target structures, and supports iterative editing across diverse structures.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Information Visualization, Image Generation and Editing, Visual Analogy

ACM Reference Format:
Shishi Xiao, Jiayi Li, Yuqiao Guan, Yiyang Zhang, and David H. Laidlaw. 2026. Skeleton-Guided Generation and Editing of Thematic Diagrams. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (SIGGRAPH 2026)*. ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

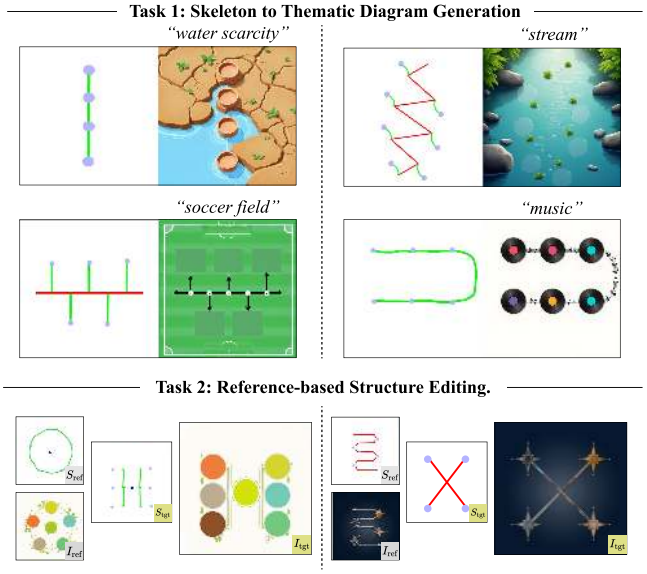


Fig. 1. Our method supports both infographic generation (left) and structure (right) editing using a skeleton-based control representation. The transformation from skeleton to infographic, expressed by a reference pair ($S_{\text{ref}}, I_{\text{ref}}$), is applied to a target skeleton S_{tgt} to produce the edited infographic I_{tgt}.

1 Introduction

Diagrams are powerful visual communication tools that organize and convey complex information through deliberately designed visual structures. Existing works [Chen et al. 2019; Offenwanger et al. 2023; Zala et al. 2024] have made it easier to create diagrams by arranging components and specifying connections. However, they often rely on conventional primitives such as box and arrow, leaving the visual appearance weakly connected to the content being communicated. Inspired by recent advances in pictorial charts [Coelho and Mueller 2020; Xiao et al. 2023], we explore how diagrams can move beyond generic visual primitives toward semantically meaningful visual representations.

Creating thematic diagrams requires a careful balance between structural fidelity and expressive, theme-aware imagery. Beyond generation, diagram creation is inherently iterative, as creators frequently change the underlying structure during the design process. However, such structure-aware editing remains underexplored. Existing design tools often require laborious manual adjustments, while image generative models produce rasterized results that are difficult to edit structurally. Therefore, we focus on two key tasks in thematic diagram creation: generation and editing.

Existing controllable image generation methods provide various forms of spatial guidance, ranging from dense controls such as edge maps [Zhang and Agrawala 2023] to abstract inputs such as sketches [Voynov et al. 2023]. Dense controls are costly for users to specify, while abstract sketches often leave room for structural drift. More importantly, both forms treat visual signals largely uniformly, ignoring that diagram components play distinct structural

Authors' Contact Information: Shishi Xiao, Brown University, Providence, United States of America, shishi_xiao@brown.edu; Jiayi Li, Brown University, Providence, United States of America, jiayi_li6@brown.edu; Yuqiao Guan, Brown University, Providence, United States of America, yuqiao_guan@brown.edu; Yiyang Zhang, Brown University, Providence, United States of America, yiyang_zhang3@brown.edu; David H. Laidlaw, Brown University, Providence, United States of America, david_laidlaw@brown.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGGRAPH 2026, Los Angeles, California
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

roles, such as central nodes, peripheral elements, and backbones. To address this, we propose a role-aware skeleton representation that abstracts diagram elements as nodes and their spatial relationships as edges. Building on this, we introduce a unified framework for thematic diagram generation and structure-aware editing.

For generation, we aim to generate a thematic diagram I_{ref} from skeleton S_{ref}. skeleton conditional tokens are jointly processed with noise tokens through multimodal attention. To connect visual appearance with diagram structure, we further introduce *role-aware positional encoding*, which extends positional encoding from spatial coordinates to a joint encoding of location and structural role. This enables feature sharing across spatially separated components with the same structural role, while decoupling appearance modeling across different roles.

For editing, our goal is to adapt a diagram to a new structure while preserving its visual style. our method takes a reference skeleton–diagram pair and a new target skeleton, and generates a diagram that follows the target structure while preserving the reference style. We formulate this task as a visual analogy problem, $S_{\text{ref}} : I_{\text{ref}} :: S_{\text{tgt}} : I_{\text{tgt}}$, where the transformation from skeleton to visual realization is inferred in a role-specific manner and transferred to the target skeleton. Compared with prior visual analogy methods [Gu et al. 2024; Hertzmann et al. 2023], which often assume spatially aligned inputs, diagram editing poses a unique challenge: transferring style across structurally different layouts. To address this, we introduce *analogy-aware attention*, which regulates token interactions between reference and target conditions through a Boolean mask to preserve element-wise style correspondence.

To support this work, we introduce Diagrams10K, a large-scale dataset of paired diagrams and skeleton representations. Constructing such data at scale is challenging because real-world thematic diagrams are scarce and often noisy, containing redundant text and decorative elements that make structure extraction difficult. To scale dataset construction, we adopt a dual data-engine strategy. The first engine extracts skeletons from real-world thematic diagrams: we train a detector to recognize node and edge elements, remove redundant visual content, and abstract the detected components into our skeleton representation. The second engine starts from synthesized skeletons and generates corresponding thematic diagrams through a self-consuming generative training loop, where outputs from the latest model are used to further expand the dataset. We evaluate the proposed skeleton representation and framework on skeleton-to-diagram synthesis and reference-based structure editing, with extensive experiments demonstrating strong structural faithfulness and style consistency.

2 Related Works

Creative Visualization Creation. Visualizations are increasingly expected not only to encode information, but also to communicate ideas in expressive and memorable ways. Existing methods can be broadly grouped into *data-driven* and *structure-driven* approaches. Data-driven methods focus on binding data to visual elements, either by augmenting charts with embellishment icons [Wang et al. 2018; Wang and Lu 2024] or by mapping data values to graphical attributes of customized elements [Kim et al. 2016; Xia et al. 2018; Zhang et al. 2020]. In contrast, structure-driven methods focus on organizing visual elements through explicit structures to support storytelling [Chen et al. 2019; Offenwanger et al. 2023, 2024; Tsandilas 2020]. For example, Tyagi et al. [Tyagi et al. 2022] proposed a semi-automated framework for structured infographic design, where templates are retrieved from a database and arranged according to a user-sketched layout. These interactive authoring systems make structure specification and revision tractable, but their reliance on predefined primitives or reusable templates limits visual expressiveness. Recent diffusion-based methods improve the ability to synthesize richer visual appearances [Wu et al. 2023; Xiao et al. 2023; Zala et al. 2024], yet their outputs remain difficult to revise at the level of diagram components and relationships. Our work targets this trade-off by combining editable structural control with generative visual synthesis, enabling thematic diagrams that support both expressive visuals and flexible structural variation.

Controllable Image Generation. Controllable image generation aims to guide the synthesis process using user-specified conditions, This covers a broad range of controllable image generation methods, ranging from spatially aligned synthesis guided by explicit structural condition [Bhat et al. 2024; Isola et al. 2017; Luo et al. 2024; Tan et al. 2025; Zhang and Agrawala 2023; Zhu et al. 2016], to subject-driven generation [Garibi et al. 2025; Kumari et al. 2023; Ruiz et al. 2023; Sohn et al. 2023; Wei et al. 2023]. Recent work further allows users to express their intent with natural language instructions via cross-modal alignment [Brooks et al. 2023; Fu et al. 2023; Ge et al. 2024; Google 2025; Wu et al. 2025].Recent diffusion transformer based methods further unify diverse spatial conditions through multimodal attention and lightweight conditioning modules [Bhat et al. 2024; Luo et al. 2024; Tan et al. 2025]. While these approaches provide effective spatial control for natural images, their control representations are less suitable for thematic diagram creation. Dense conditions such as edge maps are cumbersome to create and edit, whereas sparse controls such as sketches often lead to structural drift during generation. More importantly, these methods treat condition signals uniformly, ignoring the distinct structural functions of diagram components. We introduce a role-aware skeleton representation to explicitly encode diagram elements and their functions, and incorporate this information into the generative model through role-aware positional encoding.

Editing by Visual Analogy . Visual analogy aims to infer a transformation from a reference pair and apply it to a new input, commonly expressed as $A : A' :: B : B'$. Early image analogy and style transfer methods focus on low-level appearance transfer, such as filters, textures, and artistic styles [Diamanti et al. 2015; Fišer et al. 2016; Gatys et al. 2016; Hertzmann et al. 2023; Huang and Belongie 2017]. Later work extends visual analogy to higher-level transformations by modeling consistent operations in learned embedding spaces [Reed et al. 2015], establishing dense feature correspondences across semantically different objects [Benaim and Wolf 2021; Kong et al. 2019; Liao et al. 2017], and incorporating scene-level reasoning or diffusion priors [Bitton et al. 2023; Ganeshan et al. 2025; Šubrtová et al. 2023; Sun et al. 2023]. More recent in-context approaches exploit pretrained generative models by composing reference and target examples into a single input, enabling example-based style or subject

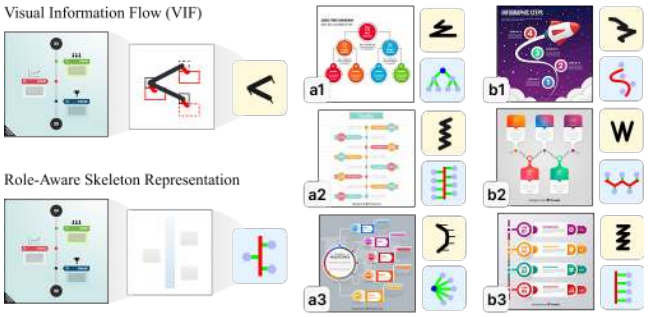
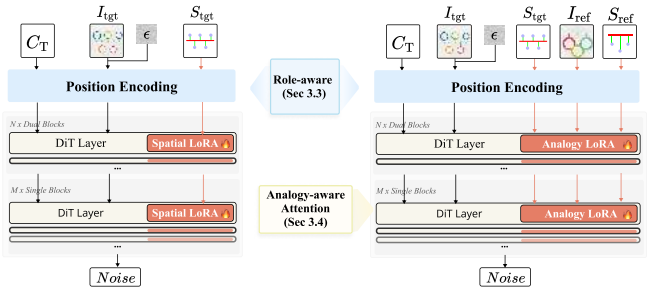


Fig. 2. Illustration of how VIF [Lu et al. 2020] and skeleton-based representations abstract infographics into concise structures, along with examples of our proposed skeleton representation for real-world infographics.



(a) Skeleton-to-diagram generation (b) Reference-based diagram editing

Fig. 3. Training Pipeline of two tasks: skeleton-to-diagram generation and reference-based diagram editing. For the second task, we incorporate semantic-aware position encoding and analogy-aware attention to perform structure variation with style preservation.

transfer without task-specific optimization [Gu et al. 2024; Huang et al. 2024; Shin et al. 2024]. Despite their progress, these methods are primarily designed for spatially aligned inputs or globally defined transformations. In contrast, diagram creation often requires transferring visual style across structurally different layouts, we therefore propose a analogy-aware attention to regulate the token interaction.

3 Method

3.1 Role-Aware Skeleton Representation

We show representative examples of real-world diagram designs in Fig. 2. Across these examples, information-bearing components are arranged through explicit spatial structures and connected by visual links. These examples highlight three requirements for a control representation of thematic diagrams. First, diagram layouts vary widely, including linear, radial, and hierarchical structures, while their information-bearing elements can take diverse visual forms. The representation should therefore be abstract enough to generalize across different diagram families, while preserving the spatial relationships that define each layout. Second, diagram components serve different structural functions, such as central hubs or connectors, and components with the same role often share visual appearance. The representation should therefore distinguish roles instead of treating all marks uniformly. Third, to support users in specifying intended structures, the representation should remain simple to construct and revise at the structural level.

A closely related abstraction is Visual Information Flow (VIF) [Lu et al. 2020], which summarizes a diagrammatic design as a perceptual flow path. We compare the abstraction process of VIF and our proposed skeleton representation in Fig. 2 left. VIF is useful for describing reading order, but it is not designed as a generative control representation. In particular, it may omit important structural components such as backbones (a2, b2, b3), and it can flatten hierarchical organization into a single sequence. For thematic diagram generation (a1, a3), we instead need a representation that preserves component identity, spatial relations, and structural roles. Specifically, our role-aware skeleton is composed of:

- **Nodes**, which represent content-bearing components such as milestones, stages, or content anchors. We further distinguish node roles by color: darker nodes • indicate central components, while lighter nodes • denote regular or peripheral ones.
- **Edges**, which encode structural relations among components. We distinguish two types of edges: backbone edges, which capture the primary structural flow of a diagram, and branch edges, which connects backbone and nodes.

3.2 Overall Pipeline

Our goal is to support two tasks in thematic diagram creation: generation and reference-based structure editing. In generation, the model creates a thematic diagram from a target skeleton and a style prompt. The key challenge is to augment spatial control with structural-role information. In editing, the model takes a reference skeleton–diagram pair and a new target skeleton, then generates a diagram that follows the target structure while preserving the reference style. This is challenging because the reference and target structures may differ substantially, yet style must be transferred between components with corresponding roles.

3.2.1 Two Tasks, One Framework. This work builds upon `FLUX.1`, a latent rectified flow model with a multimodal diffusion transformer for denoising image tokens. Prior conditional generation methods often introduce spatial conditions through separate trainable branches, which duplicate the backbone and increase computational cost [Zhang and Agrawala 2023]. Recent DiT-based conditioning methods instead concatenate image tokens and condition tokens into a unified sequence, and inject task-specific control through Low-Rank Adaptation (LoRA) in the denoising blocks [Tan et al. 2025; Wang et al. 2025]. This design allows image and condition tokens to interact through multimodal attention, while keeping the adaptation lightweight.

As shown in Fig. 3, we adopt this condition-LoRA framework for both thematic diagram generation and reference-based structure editing. For generation, we train a *Spatial LoRA* conditioned on a target skeleton S_{tgt} and a style prompt C_t, and generate the target diagram I_{tgt}. For editing, we train an *Analogy LoRA* conditioned on a reference skeleton–diagram pair $(S_{\text{ref}}, I_{\text{ref}})$, a target skeleton S_{tgt}, and a style prompt C_t, and generate I_{tgt}. Both pipelines share

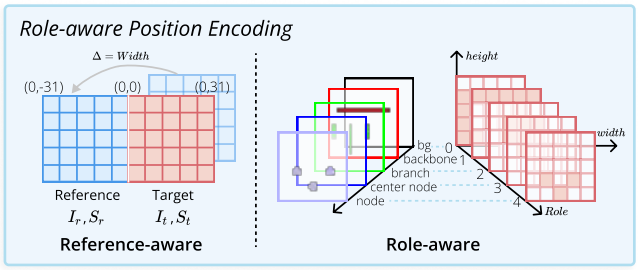


Fig. 4. Illustration of positional encoding method. (left) Reference indices are shifted to enable reference-aware alignment. (Right) Role axis encodes structural roles, including node, center node, branch, and backbone.

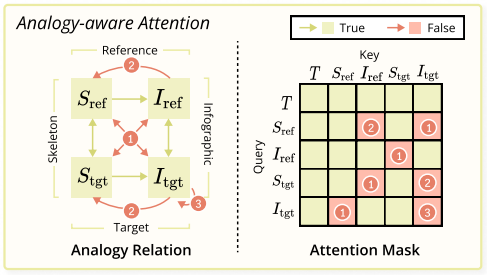


Fig. 5. Illustration of analogy-aware attention. We map the analogy relation into token interaction via an attention mask. Permissive and restrictive interactions are represented by green and red respectively.

the same pretrained diffusion backbone with task-specific LoRA modules. We further introduce two lightweight mechanisms on top of this framework. *Role-aware positional encoding* is used in both tasks to incorporate structural-role information into spatial control. *Analogy-aware attention* is used for reference-based editing to regulate token interactions between reference and target conditions, enabling style transfer across non-aligned structures.

3.3 Role-aware Position Encoding

Positional encoding is designed to inject spatial inductive bias into attention-based models. Given an input image of resolution $H \times W$, the VAE encoder produces a latent feature map of size $\frac{H}{f} \times \frac{W}{f}$, where f denotes the VAE downsampling factor. Each latent token is associated with a 2D spatial index. Rotary Position Embedding [Su et al. 2024] is incorporated into the attention layers to encode these spatial indices by applying rotation matrices to the query and key vectors. Recent work has demonstrated its capability beyond spatial layout, including depth ordering [Bai et al. 2025], layer semantics [Pu et al. 2025], and subject-driven generation [Tan et al. 2025].

We introduce two modifications to adapt positional encoding for structure-aware diagram generation and editing. First, following OmniControl [Tan et al. 2025], we apply a constant positional offset $\Delta = (\frac{H}{f}, \frac{W}{f})$ to the reference diagram I_{ref}, placing the reference and target diagrams side by side in latent space. This facilitates global style consistency between the two diagrams. Second, diagrams are inherently compositional: components such as nodes and backbones serve different structural roles and often require distinct visual treatments. To support role-aware generation and role-wise style correspondence during editing, we extend the standard 2D positional encoding with an additional *role axis*. Each latent token is assigned a role index according to its corresponding skeleton component. The positional index is therefore extended from a 2D coordinate to a triplet, with the role dimension encoded by an independent rotary embedding. This encourages tokens with the same structural role to share visual features, while separating appearance modeling across different roles.

3.4 Analogy-aware Attention

The unified sequence of noise image tokens and conditional tokens enables flexible interactions, but also risks attention entanglement or direct copying from reference (Fig. 6a). To instantiate the analogy relation explicitly, we introduce *analogy-aware attention*, which regulates token interactions using a Boolean attention mask. As shown in Fig. 5, interactions consistent with the analogy relation are marked as allowed, while interactions that may cause incorrect correspondence or copying are blocked.

Permissive interactions. The green arrows indicate valid analogy relations. First, tokens within the same modality are mutually accessible, enabling structural correspondence between skeletons ($S_{\text{ref}} \leftrightarrow S_{\text{tgt}}$) and style correspondence between diagrams ($I_{\text{ref}} \leftrightarrow I_{\text{tgt}}$). Second, to provide spatial guidance, diagram tokens are allowed to attend to their corresponding skeleton tokens ($I_{\text{ref}} \rightarrow S_{\text{ref}}$ and $I_{\text{tgt}} \rightarrow S_{\text{tgt}}$).

Restrictive interactions. Conversely, we impose three constraints, illustrated by the numbered red arrows. (1) Skeleton tokens from one pair are prohibited from attending to diagram tokens from the other pair (i.e., $S_{\text{ref}} \nleftrightarrow I_{\text{tgt}}$ and $S_{\text{tgt}} \nleftrightarrow I_{\text{ref}}$), preventing incorrect structural–visual associations. (2) Skeleton tokens are restricted from attending to diagram tokens within the same pair. Crucially, (3) the target diagram I_{tgt} is prevented from attending to itself, forcing it to rely on the reference condition rather than self-copying.

In practice, enforcing Constraint (3) across all layers degrades visual quality, producing blurred and patchy artifacts (Fig. 6b). We therefore apply this constraint only in the first 19 single-stream blocks, where correspondence routing is most critical. In later layers, we restore the default attention pattern to refine visual coherence. This mixed masking strategy produces the result shown in Fig. 6c, which faithfully follows the target structure and preserves the style. We provide ablations on the block type and layer range in the diffusion transformer for analogy-aware attention in Sec. 5.4.2.

4 Diagrams10K Dataset

We introduce Diagrams10K, a dataset of 10,000 paired diagrams and skeletons. Existing diagram datasets are limited in scale and diversity, and often contain text placeholder or decorative elements. Moreover, constructing skeletons from diagram structures remains challenging. To construct Diagrams10K at scale, we adopt a dual data-engine strategy:

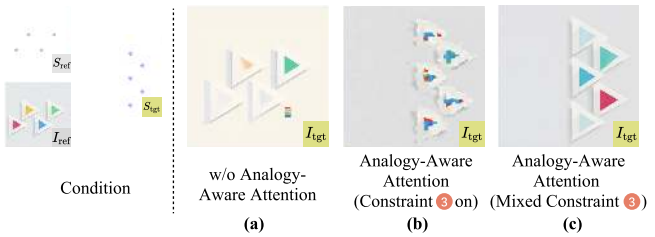


Fig. 6. Effect of analogy-aware attention. (a) Without analogy-aware attention, the model directly copies from I_{ref}. (b) Enforcing Constraint 3 across all layers significantly degrades generation quality. (c) Applying Constraint 3 selectively yields the best results.

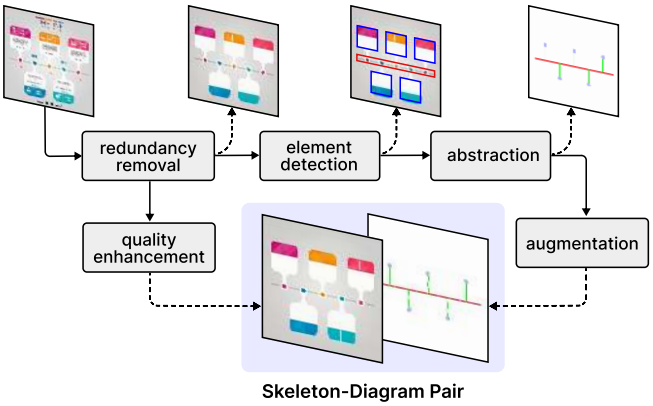


Fig. 7. Illustration of using data engine 1 to produce skeleton and diagram.

Data Engine 1: Real-world diagrams → Skeleton. We start from the 13K dataset used in VIF [Lu et al. 2020]. Since the raw collection contains many low-quality samples, such as purely illustrative images, duplicated designs, and isolated cropped elements, we first filter out noisy examples, resulting in 3,275 curated diagrams. The remaining diagrams still contain redundant text placeholders and decorative icons, which may introduce artifacts or semantically irrelevant content during training. We detect these elements using GroundingDINO [Liu et al. 2024] and remove them with LaMa [Suvorov et al. 2022], followed by an image refinement step [Trom et al. 2023] to obtain clean diagram images. To extract skeletons, we mamually annotate 600 diagrams with structural labels and train a YOLO detector [Khanam and Hussain 2024] to recognize central nodes, regular nodes, and backbones. The detected bounding boxes are then abstracted into skeleton nodes and backbone edges according to our representation. Branches are inferred based on the Gestalt principle of proximity. Finally, we augment the extracted skeletons with small geometric perturbations to simulate the natural variation of user-drawn structures. The processing pipeline is shown in Fig. 7.

Data Engine 2: Synthesized Skeleton → Diagrams. Our skeleton representation is modular and easy to parameterize, allowing us to specify structural properties such as node count and layout type. Based on these parameters, we use a large language model to generate SVG skeletons with standardized visual attributes, such as node color, edge type, and stroke width. This allows us to synthesize diverse skeleton structures and supplement underrepresented patterns in the real-world data. Given the synthesized skeletons, we generate corresponding thematic diagrams using the latest model in a self-consuming training loop, and add high-quality outputs back into the dataset to further expand Diagrams10K.

Table 1. Percentage ranked best in forced ranking results across different input control methods. Skeleton performed best in efficiency and usability, and with overall rank as 1.75.

Method	Aesthetics	Faithfulness	Efficiency	Usability	Rank
Text-only	34.6%	3.6%	2.6%	1.8%	4.00
Scribble	26.4%	32.1%	21.3%	14.5%	2.00
Painting	9.6%	17.7%	17.7%	11.1%	3.75
Canny	7.5%	19.9%	10.0%	25.9%	3.50
Skeleton	21.9%	26.7%	48.4%	46.7%	**1.75**

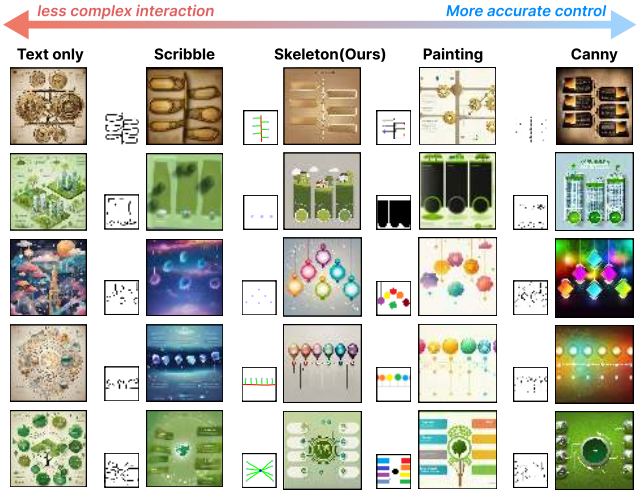


Fig. 8. Qualitative results of task 1. Comparison of different control methods.

5 Experiments

5.1 Setup

5.1.1 Evaluation Goals. We evaluate the following two tasks:

- **Task 1: Skeleton-to-diagram synthesis** evaluates the effectiveness of using skeletons as a structural control representation and measures structure faithfulness.
- **Task 2: Reference-based structure editing** also emphasizes structure faithfulness, while additionally evaluating style consistency with respect to the reference diagram.

5.1.2 Evaluation Dataset. Our evaluation dataset contains 50 samples. Target skeletons are evenly split between LLM-generated and manually drawn ones. For the generation task, we prepare style prompts generated by GPT-4o with different theme. For the editing task, reference diagrams are evenly divided between outputs generated by our model and real-world templates created by designers.

Table 2. Comparisons with baselines on structure faithfulness, style consistency, and image quality. For all criteria, a higher score indicates better performance. Highest scores are shown in bold. Second-highest scores are underlined.

Method	Structure Faithfulness			Style Consistency		Image Quality		
	Node F1 ↑	Backbone F1 ↑	Overall F1 ↑	CLIP ↑	DINO ↑	CLIP-IQA ↑	MUSIQ ↑	MAN-IQA ↑
IC-LoRA [FLUX.1]	0.467	0.519	0.491	0.423	0.341	0.543	61.838	0.466
IC-LoRA [FLUX. 1 Kontext]	0.512	0.588	0.540	**0.777**	**0.718**	<u>0.601</u>	<u>67.075</u>	0.488
StyleAligned + ControlNet	0.487	0.530	0.508	0.586	0.579	0.581	64.906	0.457
Qwen-Image-Edit	0.503	<u>0.653</u>	0.551	0.559	0.527	0.585	66.610	<u>0.493</u>
Analogist	<u>0.704</u>	0.500	<u>0.638</u>	0.584	0.513	0.526	47.407	0.383
Ours	**0.914**	**0.812**	**0.907**	<u>0.702</u>	<u>0.662</u>	**0.615**	**68.210**	**0.501**

5.2 Task 1: Skeleton-to-diagram Generation.

5.2.1 Baseline Methods. We evaluate common spatial control methods for thematic diagram generation, ranging from coarse to fine-grained control. *Text-only* generation [Podell et al. 2023] uses only the prompt with spatial description. *Scribble* control [Zhang and Agrawala 2023] provides a lightweight sketch of layout and component shapes. *Painting* control [Meng et al. 2021] specifies coarse layout, shape, and color through a visual draft. *Canny-edge* control [Zhang and Agrawala 2023] provides dense contour-level guidance. We additionally train a diagram-specific LoRA on Diagrams10K and use it across all baseline methods during evaluation.

5.2.2 User Study. We conducted a user study with 37 participants from diverse academic backgrounds, all of whom had prior experience with image generation. Using our interactive interface, participants created or extracted control inputs and inspected the corresponding generated diagrams. Participants compared four baseline control methods and ours across 20 examples. The method order was randomized in each trial. For each example, participants selected the best method under four criteria: *visual aesthetics*, *structure faithfulness*, *efficiency*, and *practical usability*. The first two criteria evaluate the generated results: visual aesthetics measures perceived visual quality, and structure faithfulness measures preservation of the intended layout and component relationships. The last two criteria evaluate the control representation: efficiency measures the ease of creating the control input, and practical usability measures its usefulness in real authoring workflows. We report the percentage of trials in which each method was selected as best, compute the final rank by averaging per-criterion ranks, and report Cramér's V to measure agreement strength among participants. After the ranking task, each participant completed a 15-minute semi-structured interview.

5.2.3 Results. We report the user study results in Tab. 1, Fig. 8 and Fig. 9. As shown in Tab. 1, Skeleton-based control achieves the highest rankings in *Efficiency* ($V = 0.387$) and *Usability* ($V = 0.381$), and ranks second and third in *Structural Faithfulness* ($V = 0.239$) and visual aesthetics ($V = 0.255$), indicating that skeletons provide an effective authoring interface aligned with user intended structure. As evidenced by the visual comparisons in Fig. 8, different control modalities reveal a trade-off between interaction effort and structural controllability. Text-only method requires minimal effort but provides weak control over layout and component organization. Canny-edge control offers precise spatial control, but requires users to prepare highly detailed control images. Scribble and painting controls are closer to common design practices, yet their generated results may deviate from the intended structure or introduce visual artifacts. In contrast, our skeleton representation captures high-level layout and component relationships without requiring users to specify low-level contours, colors, or textures.

5.3 Task 2: Reference-based Structure Editing.

5.3.1 Baseline Methods. We consider four representative baselines spanning style transfer, in-context image editing, and image analogy, including StyleAligned [Hertz et al. 2024], IC-LoRA [Huang et al. 2024], Qwen-Image-Edit [Wu et al. 2025], and Analogist [Gu et al. 2024]. StyleAligned applies DDIM inversion to the reference diagram to recover its diffusion trajectory for style transfer. Since it does not provide spatial control, we pair it with a ControlNet trained on Diagrams10K. IC-LoRA serves as an in-context baseline built on pretrained text-to-image diffusion transformers. To adapt it to our task, we pack reference and target pairs into 2×2 grid inputs and train IC-LoRA on both `FLUX.1-dev` [Labs 2024] and `FLUX.1 Kontext-dev` [Labs et al. 2025]. For `FLUX.1-dev`, inpainting is used at inference time to generate the target diagram. Qwen-Image-Edit and Analogist are included as training-free baselines and are applied directly using their official implementations.

5.3.2 Evaluation Metrics. We evaluate our method from four aspects. *Structural faithfulness* is measured using the F1 score by comparing structural elements extracted via SAM-based segmentation [Carion et al. 2025] with the ground-truth skeleton; node and backbone components are evaluated separately, and the final score is computed as a weighted average based on their counts. Style consistency is measured following Analogist [Gu et al. 2024] by computing the cosine similarity between the directional changes from S_{ref} to I_{ref} and from S_{tgt} to I_{tgt}, using image features extracted from CLIP [Radford et al. 2021] and DINO [Caron et al. 2021]. Image generation quality is assessed using CLIP-IQA [Wang et al. 2022], MUSIQ [Ke et al. 2021], and MAN-IQA [Yang et al. 2022].

5.3.3 Results. We report qualitative and quantitative comparisons in Fig. 10 and Tab. 2. As shown in Fig. 10, IC-LoRA-FLUX, IC-LoRA-Kontext, and Qwen-Image-Edit suffer from information leakage,

Table 3. Ablation study on effect of semantic-aware position encoding.

Method	Style ↑		Structure ↑		
	CLIP	DINO	Node F1	Backbone F1	Overall F1
Ours w/o SPE	0.651	0.578	**0.915**	0.748	0.892
Ours w/ SPE	**0.702**	**0.662**	0.914	**0.812**	**0.907**

where the generated results largely copy from I_{ref} or overlay patterns from S_{tgt}. StyleAligned produces outputs with noticeable variation from I_{ref}; however, their structures often deviate substantially from the target skeleton S_{tgt}. Analogist occasionally generates layouts that roughly follow S_{tgt}, but frequently misses node elements and exhibits clear style drift from I_{ref}. While both methods preserve global style to some extent, they often introduce visual artifacts and structural inconsistencies. In contrast, our method stably preserves both the target structure and reference style. Quantitatively, it outperforms all baselines in structural faithfulness, achieving a Node F1 of 0.914, a Backbone F1 of 0.812, and an overall F1 of 0.907, seeing a 26.9% increase over the second-best method. In terms of style consistency, our method ranks second only to IC-LoRA-Kontext. However, many IC-LoRA-Kontext results are nearly identical to I_{ref}, indicating trivial copying that inflates similarity-based metrics. Finally, our method also delivers superior image quality, outperforming all baselines on CLIP-IQA, MUSIQ, and MAN-IQA, with scores of 0.615, 68.210, and 0.501, respectively.

5.4 Ablation Study

5.4.1 Effectiveness of Semantic-aware Position Encoding (SPE). As shown in Tab. 3, introducing semantic-aware position encoding yields approximately 7.8% and 14.5% relative improvements in style consistency across both CLIP and DINO metrics. Although SPE is not explicitly designed to optimize structural accuracy, we also observe noticeable gains in backbone-level structural faithfulness. These results indicate that incorporating semantic information into positional encoding stabilizes both style preservation and structural faithfulness.

5.4.2 Variants of Constraint 3 in Analogy-aware Attention. We study different implementations of Constraint 3 in analogy-aware attention by varying both the block type (single vs. dual block) and the number of layers where the constraint is applied. As shown in Tab. 4, enforcing Constraint 3 within a single-stream block consistently outperforms the dual-block variant across all style consistency and image quality metrics. It shows that applying Constraint 3 only in early layers (e.g., Layer 0 or 9) yields limited style consistency, while enforcing it throughout deeper layers (e.g., Layer 29 or 38) significantly degrades image quality. The best performance is achieved when Constraint 3 is applied to the first 19 layers, striking a balance between preventing trivial copying from I_{ref} and allowing sufficient flexibility for visual refinement in later layers.

Table 4. Ablation study on the block type and layer location of self-attention.

Setting	Style Consistency ↑		Image Quality ↑		
	CLIP	DINO	CLIP-IQA	MUSIQ	MAN-IQA
Single Block	**0.702**	**0.662**	**0.615**	**68.210**	**0.501**
Dual Block	0.505	0.435	0.440	52.252	0.380
Layer 0	0.260	0.227	0.574	58.048	0.629
Layer 9	0.334	0.301	0.555	65.717	0.585
Layer 19	**0.702**	**0.662**	**0.615**	**68.210**	**0.636**
Layer 29	0.334	0.172	0.315	42.824	0.373
Layer 38	0.334	0.097	0.380	51.813	0.501

6 Applications

6.1 Procedural Editing

diagram creation is inherently iterative. As shown in Fig. 11, our method supports procedural editing, enabling both small incremental changes (e.g., adding or removing components) and larger structural modifications such as redrawing the skeleton. Throughout the iterative process, the generated diagrams consistently preserve visual style.

6.2 Illustrative diagrams

Fig. 12 demonstrates that our method is capable of generating content beyond diagrams, extending its applicability to the domain of illustrative images with well-defined structural layouts. Compared to structured diagrams, illustrative graphics offer richer aesthetic expressiveness. We hope our work can offer more opportunities to create designs with a clear structure and rich visuals.

7 Limitations

Quality Reduction in Sequential Editing. As shown in Fig. **??**, our method effectively preserves the styles of the reference infographic I_{ref} during the initial skeleton-to-infographic Synthesis. However, sequential editing may result in detail loss, color shift, or noise accumulation. Using object-level descriptions alongside the default prompt may alleviate this issue.

Scope of Applicability. Our work focuses on the structured subset of infographics with explicit geometric layouts. However, real-world infographics can be far more structurally complex. Geometric structures can be either implicitly embedded or entirely absent, relying instead on artistic expression and advanced visual techniques beyond the scope of our current framework. Further exploration into generating implicit structures that accommodate the full spectrum of design possibilities will contribute to this field.

8 Conclusion

We introduce a skeleton-based control representation for explicitly describing the structural organization of infographics. Building on this representation, we present a unified framework that supports both infographic generation and structure editing. We formulate structure editing as an image analogy problem and propose two minimal yet effective mechanisms: semantic-aware positional encoding and analogy-aware attention. They jointly enable explicit and robust instantiation of analogy within a diffusion transformer framework. We further contribute Infographics10K, a large-scale

dataset of paired infographics and skeletons that spans a diverse range of structures and visual styles, including both real-world designs and synthesized data. Extensive experiments demonstrate that our method achieves strong structural faithfulness, style consistency, and practical usability for iterative infographic creation.

References

Yunpeng Bai, Haoxiang Li, and Qixing Huang. 2025. Positional Encoding Field. *arXiv preprint arXiv:2510.20385* (2025).

Sagie Benaim and Lior Wolf. 2021. Structural Analogy from a Single Image Pair. *Comp. Graph. Forum* 40, 2 (2021), 487–499.

Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. 2024. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Yonatan Bitton et al. 2023. Visual Analogies of Situation Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. 2025. SAM 3: Segment Anything with Concepts. arXiv:2511.16719 [cs.CV] https://arxiv.org/abs/2511.16719

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Zhutian Chen, Yun Wang, Qianwen Wang, Yong Wang, and Huamin Qu. 2019. Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. *IEEE Trans. Vis. & Comp. Graphics* 26, 1 (2019), 917–926. doi:10.1109/TVCG.2019.2934810

Darius Coelho and Klaus Mueller. 2020. Infomages: Embedding data into thematic images. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 593–606.

Olga Diamanti, Connelly Barnes, Sylvain Paris, Eli Shechtman, and Olga Sorkine-Hornung. 2015. Synthesis of complex image appearance from limited exemplars. *ACM Transactions on Graphics (TOG)* 34, 2 (2015), 1–14.

Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sỳkora. 2016. Stylit: illumination-guided example-based stylization of 3d renderings. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. 2023. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102* (2023).

Aditya Ganeshan, Thibault Groueix, Paul Guerrero, Radomír Měch, Matthew Fisher, and Daniel Ritchie. 2025. Pattern Analogies: Learning to Perform Programmatic Image Edits by Analogy. In *Proc. IEEE Conf. on Comp. Vis. and Pat. Rec.*

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. 2025. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)* 44, 4 (2025), 1–11.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396* (2024).

Google. 2025. Google Nano Banana. https://gemini.google/overview/image-generation/. Online demo, sometimes referred to as "Nano Banana".

Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. 2024. Analogist: Out-of-the-box Visual In-Context Learning with Image Diffusion Model. *arXiv preprint arXiv:2405.10316* (2024).

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4775–4785.

Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. 2023. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 557–570.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. 2024. In-Context LoRA for Diffusion Transformers. (2024).

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. arXiv:2108.05997 [cs.CV] https://arxiv.org/abs/2108.05997

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* (2024).

Nam Wook Kim, Eston Schweickart, Zhicheng Liu, Mira Dontcheva, Wilmot Li, Jovan Popovic, and Hanspeter Pfister. 2016. Data-driven guides: Supporting expressive design for information graphics. *IEEE Trans. Vis. & Comp. Graphics* 23, 1 (2016), 491–500. doi:10.1109/TVCG.2016.2598620

Bailey Kong, James Supancic III, Deva Ramanan, and Charless C Fowlkes. 2019. Cross-domain image matching with deep feature maps. *International Journal of Computer Vision* 127, 11 (2019), 1738–1750.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1931–1941.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742* (2025).

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088* (2017).

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 38–55.

Min Lu, Chufeng Wang, Joel Lanir, Nanxuan Zhao, Hanspeter Pfister, Daniel Cohen-Or, and Hui Huang. 2020. Exploring visual information flows in infographics. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. 2024. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8217–8227.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021).

Anna Offenwanger, Matthew Brehmer, Fanny Chevalier, and Theophanis Tsandilas. 2023. Timesplines: Sketch-based authoring of flexible and idiosyncratic timelines. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 34–44.

Anna Offenwanger, Theophanis Tsandilas, and Fanny Chevalier. 2024. Datagarden: Formalizing personal sketches into structured visualization templates. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2024), 1268–1278.

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

Yifan Pu, Yiming Zhao, Zhicong Tang, Ruihong Yin, Haoxing Ye, Yuhui Yuan, Dong Chen, Jianmin Bao, Sirui Zhang, Yanbin Wang, et al. 2025. Art: Anonymous region transformer for variable multi-layer transparent image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7952–7962.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. 2015. Deep Visual Analogy-Making. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1252–1260.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. 2024. Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator. (2024).

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983* (2023).

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. 2023. Diffusion Image Analogies. In *Proceedings of SIGGRAPH 2023 (Conference Papers)*. 79. doi:10.1145/3588432.3591558

Yasheng Sun, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. 2023. ImageBrush: Learning Visual In-Context Instructions for Exemplar-Based Image Manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2025. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14940–14950.

Benjamin Trom, Pierre Chapuis, and Cédric Deltheil. 2023. Refiners: The simplest way to train and run adapters on top of foundation models. https://github.com/finegrain-ai/refiners.

Theophanis Tsandilas. 2020. StructGraphics: Flexible visualization design through data-agnostic and reusable graphical structures. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 315–325.

Anjul Tyagi, Jian Zhao, Pushkar Patel, Swasti Khurana, and Klaus Mueller. 2022. Infographics wizard: Flexible infographics authoring and design exploration. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 121–132.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 conference proceedings*. 1–11.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. 2025. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277* (2025).

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. 2022. Exploring CLIP for Assessing the Look and Feel of Images. arXiv:2207.12396 [cs.CV] https://arxiv.org/abs/2207.12396

Yun Wang, Haidong Zhang, He Huang, Xi Chen, Qiufeng Yin, Zhitao Hou, Dongmei Zhang, Qiong Luo, and Huamin Qu. 2018. InfoNice: Easy creation of information graphics. In *Proc. ACM CHI*. 1–12. doi:10.1145/3173574.3173909

Zhenyu Wang and Min Lu. 2024. A Versatile Collage Visualization Technique. *arXiv preprint arXiv:2406.04008* (2024).

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15943–15953.

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

Jiaqi Wu, John Joon Young Chung, and Eytan Adar. 2023. viz2viz: Prompt-driven stylized visualization generation using a diffusion model. *arXiv preprint arXiv:2304.01919* (2023).

Haijun Xia, Nathalie Henry Riche, Fanny Chevalier, Bruno De Araujo, and Daniel Wigdor. 2018. DataInk: Direct and creative data-oriented drawing. In *Proc. ACM CHI*. 1–13. doi:10.1145/3173574.3173797

Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. 2023. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics* (2023).

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment. arXiv:2204.08958 [cs.CV] https://arxiv.org/abs/2204.08958

Abhay Zala, Han Lin, Jaemin Cho, and Mohit Bansal. 2024. DiagrammerGPT: Generating Open-Domain, Open-Platform Diagrams via LLM Planning. In *COLM*.

Jiayi Eris Zhang, Nicole Sultanum, Anastasia Bezerianos, and Fanny Chevalier. 2020. DataQuilt: Extracting visual elements from images to craft pictorial visualizations. In *Proc. ACM CHI*. 1–13. doi:10.1145/3313831.3376172

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2016. Generative Visual Manipulation on the Natural Image Manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*.

ChArtist: Generating Pictorial Charts with Unified Spatial and Subject Control

Shishi Xiao*
Brown University
shishi_xiao@brown.edu

Tongyu Zhou
Adobe Research
tongyuz@adobe.com

David H. Laidlaw
Brown University
dhl@cs.brown.edu

Gromit Yeuk-Yin Chan
Adobe Research
ychan@adobe.com

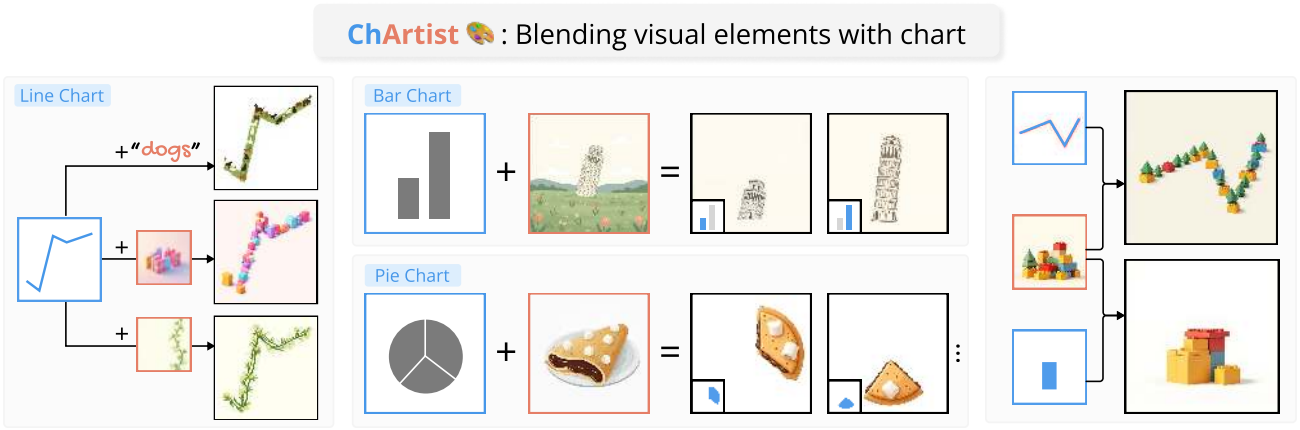


Figure 1. Illustrations of pictorial charts generated by ChArtist. We convert chart primitives such as bars, lines, and segments into vivid visual elements. With user-provided text or reference images, ChArtist combines spatial and subject control to maintain data fidelity while achieving visual consistency.

Abstract

A pictorial chart is an effective medium for visual storytelling, seamlessly integrating visual elements with data charts. However, creating such images is challenging because the flexibility of visual elements often conflicts with the rigidity of chart structures. This process thus requires a creative deformation that maintains both data faithfulness and visual aesthetics. Current methods that extract dense structural cues from natural images (e.g., edge or depth maps) are ill-suited as conditioning signals for pictorial chart generation. We present ChArtist, a domain-specific diffusion model for generating pictorial charts automatically, offering two distinct types of control: 1) spatial control that aligns well with the chart structure, and 2) subject-driven control that respects the visual characteristics of a reference image. To achieve this, we introduce a skeleton-based spatial control representation. This representation encodes only the data-encoding information of the chart, allowing for the easy incorporation of reference visuals without a rigid outline constraint. We implement our method based on the Diffusion Transformer (DiT) and leverage an adaptive position encoding mechanism to manage these two controls. We further introduce Spatially Gated Attention to modulate the interaction between spatial control and subject control. To support the fine-tuning of pre-trained models for this task, we created a large-scale dataset of 30,000 triplets (skeleton, reference image, pictorial chart). We also propose a unified data accuracy metric to evaluate the data faithfulness of the generated charts. We believe this work demonstrates that current generative models can achieve data-driven visual storytelling by moving beyond general-purpose conditions to task-specific representations. Project page: https://chartist-ai.github.io/.

1. Introduction

Pictorial charts that embed semantic data directly into chart structures provide a rich and engaging way to tell visual stories [8, 34, 48, 49]. These visualizations captivate viewers by blending representative imagery with data, converting information into experiences that enhance both understanding and long-term recall [4, 14].

However, creating such a pictorial chart from scratch is highly challenging, involving both creativity and design skills. The integration of malleable visual imagery with the rigid shape of a standard chart requires a careful balance

*Work done during internship at Adobe Research.

of data faithfulness and visual aesthetics. Some examples of this are shown in Fig. 1. For example, to create a line chart about dogs, the designer may first query for and identify an image that aligns with the overall trend of the line chart, then overlay the chart on top [8, 34]. Alternatively, they could also use collage tools to fill in data points on the line chart with custom glyphs [9, 37]; this is operationally simpler at the cost of the pictorial chart's expressiveness. While recent authoring tools have been developed to assist in pictorial chart creation, these workflows still require substantial manual refinement and lack effective quantitative evaluation of data faithfulness [48, 49].

In this paper, we propose an end-to-end pipeline for pictorial chart generation that supports two types of control: 1) *spatial* and 2) *subject-driven*. These controls were inspired by observations of real design workflows: users either adopted a "data-first" approach—finalizing the chart first and searching for compatible images afterwards—or a "visual-first" approach—selecting an appealing image and deforming it to fit the chart structure.

Existing spatial-control methods typically rely on image-based conditions such as Canny edges [18, 28, 32, 41, 58], which are designed for natural images that contain far greater structural complexity than charts. To tackle this problem, we propose a new control representation for data charts. By focusing solely on the data-encoding dimension, each chart is condensed to a skeleton representation that enables semantic and stylistic adaptability during generation. As shown in Fig. 3, this skeleton encodes semantically meaningful visual properties such as bar heights, line trends, and pie-slice angles.

Our approach builds upon FLUX [24], a pretrained Diffusion Transformer (DiT) model, and trains two task-specific LoRA modules: $LoRA_S$, which enforces spatial control from the chart skeleton, and $LoRA_R$, which injects subject control from reference images. These two controls can be applied independently or jointly to support diverse design workflows, as illustrated in Fig. 4 B.

However, composing multiple LoRAs in parallel introduces substantial cross-condition interference, a common issue in multi-control setups, yet particularly harmful for charts, where any subject-induced distortion may directly break data faithfulness. To address this, we propose Spatially-Gated Attention, which sequentially modulates the interaction between spatial and subject control, establishing an explicit dependency between them during inference and substantially improving spatial fidelity and visual consistency under harmonious dual control.

To validate the effectiveness of our approach, we construct CHARTIST-30K, a high-quality synthetic dataset of 30,000 skeleton–reference–pictorial chart triplets, generated through two specialized pipelines tailored to distinct chart topologies. In addition, we introduce a unified data accuracy metric to quantitatively evaluate different forms of pictorial charts. We hope that this dataset and evaluation framework will foster future research on data-driven storytelling with generative models. Our evaluations show that ChArtist achieves robust balance between data faithfulness and visual expressiveness. Our key contributions are summarized as follows:

- A chart-specific control representation that is minimal, precise, and flexible.
- An end-to-end framework for pictorial chart generation with both spatial and subject control.
- A Spatially-Gated Attention mechanism to mitigate cross-condition interference.
- A CHARTIST-30K dataset and a unified data-accuracy metric for pictorial charts.

2. Related Works

2.1. Control Representations for Generative Models

Current generative models have established a series of control representations that can be broadly categorized by their levels of granularity. On one hand, dense representations such as Canny edges, depth maps, and semantic segmentation maps [18, 28, 32, 41, 58] can provide strict guidance and are effective for tasks such as photorealistic generation. While powerful, these dense representations assume pixel-perfect correspondence between the conditions and the generated image. This rigidity restricts the range of visual deformations needed for expressive tasks. Recent work, such as LooseControl [3], has identified this limitation and proposed a looser adherence to depth conditions. More relevant to our work are abstract, sparse representations, including sketches [22, 43, 65], bounding boxes [26, 60, 62], and human pose keypoints [7, 19, 30]. These representations allow for structural control while preserving stylistic creativity. However, directly applying them to pictorial chart generation remains challenging: sketches focus on overall shape instead of data-encoding dimension, and pose keypoints rely on a fixed set of joints, making them unsuitable for generalizing across chart types with varying structures. In response, we introduce a control representation specifically for charts, enabling precise data information while generalizing consistently across different chart topologies.

2.2. Controllable generation

Spatially Aligned Generation It typically employs image-conditioned controls to enforce structural constraints. Some approaches inject structural features into frozen diffusion backbones through task-specific branches or adapters [27, 32, 58]. Recent studies move toward unified and multi-modal conditioning frameworks [15, 21, 25, 61], which encode diverse spatial priors into shared representations, enabling cross-condition generalization.

Subject Driven Generation Subject-driven generation aims to preserve the identity of a reference subject while generating it across diverse visual contexts. Pioneering works rely on subject-specific optimization for each new concept, whether by optimizing model weights [23, 36], text embeddings [39, 46], or modulation signals in diffusion transformers [11, 64]. Some methods achieve high-fidelity subject preservation by leveraging the in-context generation capability of pretrained diffusion models, either via lightweight LoRA tuning or by direct inpainting with side-by-side concatenated images [13, 16, 38].

Unified Generation Recent works aim to unify spatial and subject control within a single framework. Luo et al. [29] propose parameter-efficient readout heads that extract intermediate features and project them into the target domain, which are used to guide latent optimization during sampling. UniCombine [44] and OmniControl [41] leverage unified sequence processing for DiT, allowing flexible token interactions. Building on this line of research, we extend unified controllable generation to jointly model spatial constraints and reference-driven visual consistency in pictorial chart synthesis.

2.3. Visual Blending

Visual blending, the process of integrating multiple visual concepts into a coherent and interpretable composition, is an effective strategy for visual communication. It has been widely explored in various forms, such as creating illusion art [5, 12], hiding QR codes in images [51, 59], and semantic text stylization [17, 42, 50, 53, 54]. A key challenge across these tasks is preserving both structural fidelity and visual expressiveness. To tackle this, earlier work often relied on retrieval-based composition [42] to match shape constraints or on style transfer techniques to apply texture to a given contour [2, 10]. More recent approaches, particularly in semantic typography, optimize the outline itself to incorporate new concepts by leveraging the rich natural image priors from pretrained diffusion models [17]. Our work extends visual blending to data charts, a domain that demands even stricter structural guidance due to the need to encode precise numerical data. We aim to achieve this through an end-to-end generative pipeline, rather than manual authoring tools [8, 49, 57], and we introduce a unified metric for evaluating data accuracy.

3. CHARTIST-30K Dataset

We construct a large-scale synthetic dataset including 30,000 triplets for pictorial chart generation. Each sample is a (S, R, P) triplet where S is the chart skeleton image, R is the reference image, and P is the resulting pictorial chart that blends R and S. We generated 10,000 examples of bar, line, and pie charts each to form the whole dataset.

Different chart types require distinct generation pipelines, as the deformation needed to adapt the reference image R with the geometric properties in S varies substantially across charts. For example, bar charts require the height of an object to be adjusted, while line and pie charts require more dramatic deformation to match the target topologies like curvature and angles. Thus, we propose two specialized generation pipelines, which are illustrated in Fig. 2. They generalize by covering two encoding rules: Pipeline (1) applies to charts with single linear or rectilinear encodings (e.g., bar, histogram, box, treemap, heatmap), while Pipeline (2) applies to charts with multi-dimensional or angular encodings (e.g., line, pie/donut, scatter, radar).

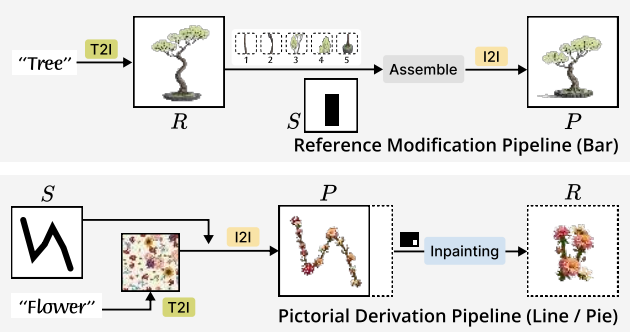


Figure 2. Pipelines for CHARTIST-30K Dataset Construction.

3.1. Reference-Modification-based Pipeline

For bar charts, this pipeline follows an $(R, S) \rightarrow P$ process: we first generate a reference image R, and then adjust it according to the height constraints specified by the chart skeleton S to form the pictorial chart P.

Reference Object Generation. We first generate the reference object R using text-to-image (T2I) model with a prompt emphasizing a single object on a clean background. We then remove the background with BiRefNet [63] to obtain its precise height. Such a result is expected to be equivalent to a pictorial bar chart as well.

Pictorial Chart Assembly. We vertically divide generated R into $K = 5$ equal-sized grids $g_1, ..., g_K$. Our goal is to assemble these grids to align with the target height defined in S. To identify structurally editable grids, we compute the Structural Similarity Index (SSIM) between all grid pairs (g_i, g_j) where $i \neq j$. Grids with higher mean SSIM represent repetitive textures that can be safely extended or trimmed without disrupting the object's visual coherence. We rank all grids by their mean SSIM scores to obtain an editability priority queue. Guided by this ranking, we either replicate the highest-ranked grid or remove lower-ranked grids to match the height specified in S. Finally, the assembled image is injected into an image-to-image (I2I) pipeline for refinement, producing the final seamless pictorial chart P. This design is robust even for non-vertical objects.

3.2. Pictorial-Derivation-based Pipeline

Directly warping a canonical object R to fit the curved skeleton S is ill-posed (e.g., fitting a flower into a line chart). Thus, we adopt a reverse pipeline $S \rightarrow P \rightarrow R$: we first generate a plausible pictorial chart P that matches the shape of S, and then derive the reference object R.

Pictorial Chart Generation. We begin by generating a textured background that contains repetitive patterns of a reference scene using a T2I model. The chart skeleton S is then used as a binary mask to crop the chart-shaped region, forming an initial P. This initial P often contains incompleted objects in the chart region. We refine it using an I2I model to form the final P such that the objects obtain the missing details.

Reference Object Derivation. To derive R, we employ a diptych prompting strategy [38] with an inpainting model: we append a blank panel to the right of P and prompt the model to generate an object that matches the appearance of P but with a natural, standalone shape. The filled region is cropped as the reference object R. The result (i.e., the filled region) is cropped to serve as our reference object R. Noted that sometimes R might still carries chart-like characteristics. To tackle this, we repeat the inpainting process using the latest R as input to obtain a more natural object. Manual verification and filtering are applied at both stages to ensure visual consistency and data faithfulness.

4. Method

Our method builds on a pretrained Diffusion Transformer (DiT) model to blend visual imagery with chart data. Given an input chart, we support both text-driven and reference-image-driven generation to allow different levels of customization. The output image should align with the structure defined by the chart while conforming to the visual semantics specified by the text or reference image. To achieve this, we first introduce a skeleton-based representation that explicitly encodes the chart structure to guide spatial control. We then train two LoRA modules to achieve spatial and subject control respectively. However, we find that naively combining LoRAs in parallel leads to cross-condition interference, where conflicts between the two controls lead to degraded structure misalignment and noticeable style leakage; see Fig. 5. To address this, we propose a cascaded conditioning inference paradigm that establishes an explicit dependency between conditions, dynamically gating the subject's influence by the spatial signal.

4.1. Data-Driven Control Representation

To unify spatial and subject control, the control representation for pictorial chart generation must faithfully encode data while maintaining visual flexibility for stylistic adaptation. Existing control representations often fail to meet this requirement, tending either to be overly dense or overly sparse. As we conceptualize on a spectrum (Fig. 3), dense pixel-level conditions (e.g., depth, Canny) are overly complex, leaving little room for stylistic infusion, whereas sparse conditions (e.g., bounding boxes) are too weak to guide internal structures. We follow the concise design of sketch and skeleton, and propose a skeleton-based representation that occupies a sweet spot in the spectrum, where it 1) faithfully encodes the primary data-encoding dimension, and 2) is structurally minimal to enable flexible semantic or stylistic injection.

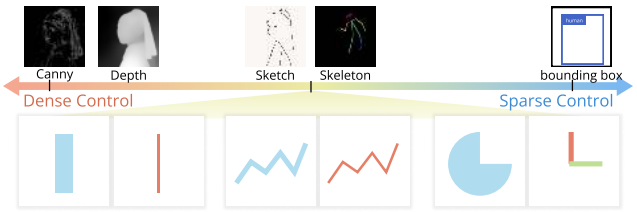


Figure 3. The spectrum of control representation based on their complexity. Top: Existing control representations for natural image; Bottom: Our proposed skeleton-based representations for pictorial chart generation, where charts (blue) are represented as skeletons (red and green lines).

Specifically, we define three types of chart-type-specific skeletons: a single vertical line represents each bar in bar charts, a polyline traces the trend in line charts, and two colored radial lines indicate the clockwise start and end angles of each slice in pie charts.

4.2. Learning Spatial and Subject Control

Training Pipeline. We adapt a DiT-based diffusion model following a conditional-LoRA architecture [41, 44]. The image condition token C (*i.e.*, either chart skeleton S for spatial or reference image R for subject) is concatenated with the text tokens T and noisy image tokens X to form a unified input sequence $[T, X, C]$, as illustrated in Fig. 4 (A). During forward process, T and X are handled by the frozen pretrained backbone, whereas C is processed by trainable LoRA adapters. The unified sequence allows flexible interactions among textual, latent, and conditional tokens through multimodal attention.

Task-Specific LoRA. We train two task-specific LoRA adapters, $LoRA_S$ for spatial control and $LoRA_R$ for subject control. They can be used independently or jointly to support different sources of semantic information, such as concept semantics from text prompts or appearance features from reference images (Fig. 4 B). Their training differs in two aspects. First, $LoRA_S$ is trained on skeleton–pictorial pairs (S, P), while $LoRA_R$ is trained on reference–pictorial pairs (R, P). Second, spatial control requires spatial alignment with the latent X, whereas subject control does not.

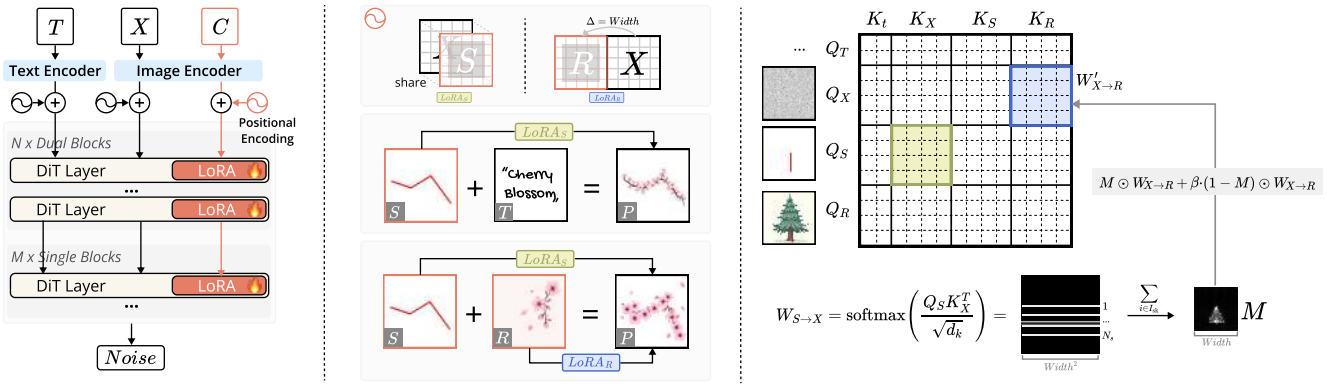


Figure 4. The whole architecture of ChArtist consists of (A) a pretrained DiT-based diffusion model with two conditional-LoRAs with different positional encoding on the image inputs, where (B) one is for spatial control ($LoRA_S$) and another for subject control ($LoRA_R$). (C) To avoid the interference from $LoRA_R$ to $LoRA_S$, we employ a Spatially-Gated Attention mechanism at inference time.

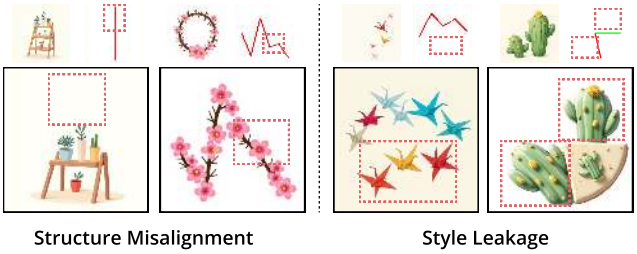


Figure 5. Artifacts observed when merging multiple LoRAs in parallel. $LoRA_R$ can distort the spatial control and disobey the chart structure (structure misalignment), and can also introduce extra visual elements beyond the spatial constraints (style leakage).

To unify them within the same framework, we adopt a RoPE-based position-aware strategy [40, 41]. The skeleton S shares positional indices with the latent tokens X, while the reference R is shifted by an offset Δ to the side of X in latent space.

4.3. Dual-Control Inference

Challenge of Merging Multiple LoRAs. Conventional multi-condition control typically composes LoRAs in parallel, treating each signal as an independent input. However, such parallel composition introduces severe cross-condition interference, which is especially harmful for pictorial charts where spatial and subject constraints must be tightly coordinated. This interference causes the generated chart to misrepresent the underlying data, manifesting in two primary failure modes, as shown in Fig. 5: the result may fail to adhere to the skeleton structure (*structure misalignment*), or it may fill the chart region correctly but still leak subject content into the background, making the chart visually ambiguous (*style leakage*).

Spatially-Gated Attention. To resolve this interference, we move from a parallel-competition paradigm to a sequential-conditioning paradigm at inference time. Pictorial charts inherently require an explicit dependency between spatial and subject control: the subject must be coordinated by the spatial constraint. To enforce this dependency, we propose a training-free inference mechanism called *Spatially-Gated Attention*. The core idea is to derive a spatial mask from the spatial condition and use it to dynamically gate the influence of the subject signal.

We first construct a spatial mask M. Since the skeleton S is a sparse and abstract representation that does not directly encode the concrete object shape, we compute an attention map between the skeleton queries Q_S and latent keys K_X:

$$W_{S\to X} = \text{softmax}\left(\frac{Q_S K_X^T}{\sqrt{d_k}}\right),$$

where each element $(W_{S\to X})_{i,j}$ represents the probability that latent token j aligns with skeleton token i. To identify the set of data-encoding skeleton tokens, denoting as $I_S \subseteq \{1, \ldots, N_S\}$, we map foreground pixel coordinates (e.g., colored pixels) from the skeleton image to their corresponding 1D token indices in the latent sequence via coordinate downsampling and flattening. The spatial mask M is then computed by aggregating the attention over this set I_S:

$$M = \sum_{i\in I_S} (W_{S\to X})_i.$$

This mask is then applied to gate the subject attention. The original subject attention weights $W_{X\to R}$ (from Q_X to K_R) are replaced by gated weights $W'_{X\to R}$:

$$W'_{X\to R} = M \odot W_{X\to R} + \beta \cdot (1 - M) \odot W_{X\to R},$$

where $\odot$ denotes element-wise multiplication, and β controls the degree of subject expressiveness in background re-

gions. Finally, a normalization step is performed to restore the probability distribution.

5. Experiments

5.1. Settings

Training. We follow the default settings of OMNICONTROL [41] and fine-tune FLUX.1-DEV on our CHARTIST-30K using LoRA. For each chart type, we train two LoRA adapters for spatial control and subject control with rank 16 at 512×512 resolution. All experiments are conducted on $2\times$ NVIDIA A100 (80GB) GPUs. Each model is trained for 25,000 iterations per task.

Evaluation. We evaluate the model on two tasks: spatially aligned control and subject-guided control. A method is considered robust if it can handle a wide range of visual concepts while adapting to different chart structures. To assess this, we generate 500 evaluation images per task for each chart type, using prompts produced by ChatGPT that span 30 major categories (e.g., plants, animals, buildings, sports, etc.). For the subject-guided task, the reference image is produced by FLUX.1-SCHNELL. β is set to 0.6. More Evaluation details can be found in Appendix A.3.

Task 1: Spatially Aligned Only Task. We compare our method against baselines with various spatial control representations: ControlNet (Canny/Depth) [58], SDEdit [31], and Inpainting [35]. All baselines use FLUX.1 as the backbone. For SDEdit relying on the colored strokes, we prompt ChatGPT to return a theme-related color based on the text prompt and apply it to the strokes.

Task 2: Subject-Guided Task. We compare against the combination of ControlNet (Canny/Depth) and IP-Adapter [56], Paint-by-Example [52], and current advanced image editing models including Qwen-Image-Edit [47], Nano Banana [1], and GPT-Image-1 [33] . We only evaluate the similarity between the semantics of the references and generations *within* the chart boundaries.

5.2. Evaluation Metrics

Data Accuracy Metric. We design a *structure-aware* F1 Score to evaluate structural alignment between generated pictorial charts and their skeletons. As chart skeletons are sparse, conventional spatial metrics such as IoU fail to capture structural alignment. To address this, we construct a distance field along the chart's data-encoding dimension, where larger distances indicate a stronger geometric bias (see line chart example in Fig. 6). We divide the chart area into multiple regions defined by distance and sample points within each region to approximate precision and recall. Each region is assigned a weight based on its spatial importance, forming a weighted score that reflects both geometric accuracy and structural completeness. The implementation details can be found in A.1. Importantly, our

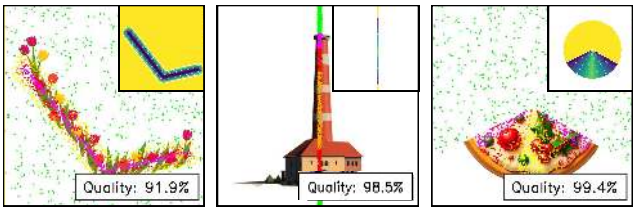


Figure 6. Illustrations of the data accuracy metric. We construct a distance field based on the chart skeleton. Then we randomly sample the points based on the ranges on the distance field, and calculate the accuracy based on a weighted F1 Score.

weighting scheme is guided by the data-encoding type:

- **Line chart:** Regions closer to the data trajectory receives higher weights to emphasize positional accuracy.
- **Bar chart:** weights peak near the bar endpoints that represent height information.
- **Pie chart:** we assign higher weights to the radial division lines that determine the angle of the pie.

Semantic Alignment and Image Quality. Text alignment is measured using CLIP Text score. For evaluating visual consistency between reference and generated image, we report CLIP-Image and DINO [6]. To reduce background bias, visual consistency scores are computed on the cropped data-encoding region rather than on the full image. Image quality is measured using CLIP-IQA [45], MANIQA [55], and MUSIQ [20].

User Perceptual Study. To assess human preferences regarding data faithfulness and semantic quality, we conducted an online comparison study with 300 participants. Participants were evenly assigned to bar, line, or pie question sets, each containing 50 data-accuracy ranking questions and 50 semantic-alignment ranking questions. Each question asked participants to rank four methods, with the method order randomized to reduce bias; in total, we obtained 100 responses for each of 300 unique questions. Each participant was compensated $15.

5.3. Results

Spatially Aligned Evaluation with different Control Representations. We show the qualitative results in Fig. 7 and quantitative results in Tab. 1, with more results can be found in A.5. As shown in the figure, baseline methods exhibit a clear trade-off between preserving chart structure and aligning with textual semantics. For instance, ControlNet rigidly follows its Canny or depth conditions, often pushing semantic content into the background instead of the chart regions, compromising data fidelity. Similarly, SDEdit, despite having good color initialization, adapts the global appearance to match the prompt but frequently

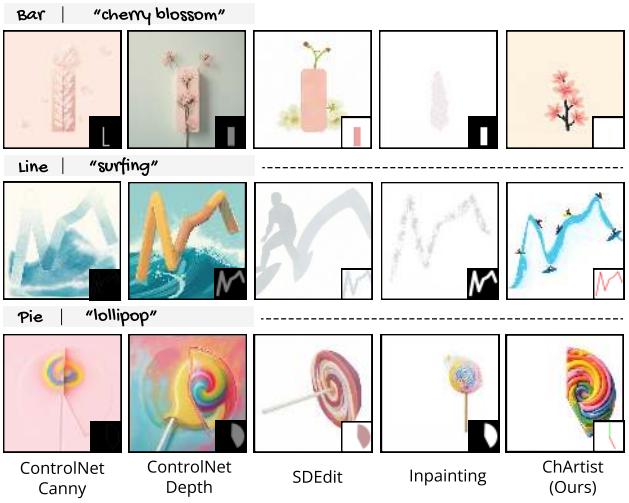


Figure 7. Result of spatially aligned evaluation with different control representations. (Task 1)

distorts underlying data information. Inpainting performs well on simple chart types such as bar charts, but struggles when deformation is required, often failing to fully populate the chart region, as illustrated by the incomplete lollipop-shaped pie chart. Furthermore, because inpainting is operated on a blank background without visual context, its results tend to be less expressive and often degrade into sketch-like appearances. In contrast, ChArtist achieves a far more harmonious balance, integrating semantic visual elements into chart structures without sacrificing expressiveness. It does not rely on rigid contour adherence as ControlNet does, preventing content from being confined to hard structural boundaries. Meanwhile, its spatial guidance is substantially stronger than weak image-conditioned approaches like SDEdit or Inpainting, which often drift away from the intended chart shape or become unrecognizable.

This observation is further supported by the quantitative results in Tab. 1. ChArtist consistently delivers stable overall performance across chart types by evaluating both data accuracy and text alignment. Notably, it outperforms other methods on CLIP-T on both types of chart, demonstrating it can incorporate the visuals while respecting structural constraints. Note that while some baselines excel in isolated cases (e.g., Inpainting achieves highest bar chart data accuracy), they do so at the cost of yielding a low CLIP-T score.

Human Preference from Controlled Online Study. From Tab. 2, our method demonstrates consistent performance (2nd place for both skeleton and reference questions), which corroborates our qualitative results that it offers the most balanced performance across both data accuracy and semantic alignment. More analysis and significance tests can be found in Appendix A.2.

Table 1. Quantitative comparison of spatially aligned evaluation.

		Method				
		ControlNet-C	ControlNet-D	SDEdit	InPainting	ChArtist (Ours)
Bar	Data Acc ↑	0.741	0.686	0.774	0.923	0.894
	CLIP-T ↑	0.249	0.243	0.233	0.231	0.304
	Avg ↑	0.495	0.465	0.504	0.577	0.599
Line	Data Acc ↑	0.819	0.858	0.792	0.754	0.920
	CLIP-T ↑	0.227	0.243	0.190	0.179	0.247
	Avg ↑	0.523	0.550	0.491	0.467	0.584
Pie	Data Acc ↑	0.725	0.626	0.836	0.794	0.778
	CLIP-T ↑	0.136	0.158	0.190	0.217	0.252
	Avg ↑	0.430	0.392	0.513	0.506	0.515

Table 2. Human evaluation results from controlled online study (300 participants). Average rank ranges from 1 (best) to N (worst).

Method	**Spatial**	**Subject**	**Method**
ControlNet-Canny	3.008	2.914 (3rd)	ControlNet-Canny
ControlNet-Depth	2.969 (3rd)	2.822 (1st)	ControlNet-Depth
SDEdit	2.897 (1st)	3.085	Paint-by-Example
Inpainting	3.168	-	-
ChArtist (Ours)	2.957 (2nd)	2.845 (2nd)	ChArtist (Ours)

Dual Control Evaluation. We present a quantitative comparisons in Tab. 3 and the qualitative analysis in Fig. 8a, Fig. 9. As shown in Tab. 3, with more results can be found in A.5. Our method consistently outperforms all baselines across all chart types, in terms of data accuracy, visual consistency with reference image, and image quality. This observation of high data accuracy is the most evident in line charts, where ChArtist achieves a data accuracy of 0.905, surpassing the second-best method (0.728) significantly. For visual consistency with the reference image (DINO and CLIP-I), ChArtist again achieves the best performance across all chart types. Regarding overall image quality, we use the average score across three chart types; ChArtist still delivers high image quality, except for the CLIP-IQA score. It is worth noting that although Paint-by-Example achieves a high data accuracy (0.912) for bar charts, this comes at the cost of visual consistency. Its DINO and CLIP-I scores are relatively low, indicating a failure to preserve the reference input's visual style. The qualitative results in Fig. 8a corroborate these quantitative findings. Baseline methods, despite their strong performances on natural images, generally struggle to blend the reference appearance into the correct chart regions. The bottom of Fig. 8a and Fig. 1 presents variations of ChArtist generated from the same reference image under different structural inputs. The consistent preservation of appearance demonstrates that our dual-control framework enables both strong visual consistency and robust adaptability to structural variations. The results in Fig. 9 further demonstrate the potential of image editing models in preserving visual consistency. However, they still exhibit limited ability to faithfully follow structural constraints.

Table 3. Quantitative comparison result of dual control of spatial and subject (Task 1 + Task 2).

Method	Bar			Line			Pie			Image Quality		
	Data Acc ↑	DINO ↑	CLIP-I ↑	Data Acc ↑	DINO ↑	CLIP-I ↑	Data Acc ↑	DINO ↑	CLIP-I ↑	CLIP-IQA ↑	MUSIQ ↑	MAN-IQA ↑
ControlNet-Canny + IP-Adater	0.634	0.652	0.824	0.728	0.613	0.758	0.652	0.651	0.701	0.674	67.38	0.487
ControlNet-Depth + IP-Adater	0.593	0.635	0.805	0.683	0.615	0.752	0.579	0.668	0.723	0.620	54.66	0.372
Paint-by-Example	0.912	0.586	0.708	0.513	0.429	0.615	0.420	0.495	0.634	0.683	65.37	0.574
Qwen-Image-Edit	0.733	0.697	0.856	0.574	0.621	0.732	0.765	0.578	0.621	0.660	63.18	0.567
Nano Banana	0.727	0.731	0.812	0.716	0.606	0.685	0.546	0.692	0.727	0.679	65.32	0.565
GPT-Img-1	0.758	0.745	0.830	0.628	0.679	0.756	0.422	0.657	0.718	0.712	67.98	0.581
ChArtist (Ours)	0.931	0.837	0.892	0.905	0.728	0.815	0.753	0.689	0.747	0.657	69.35	0.589

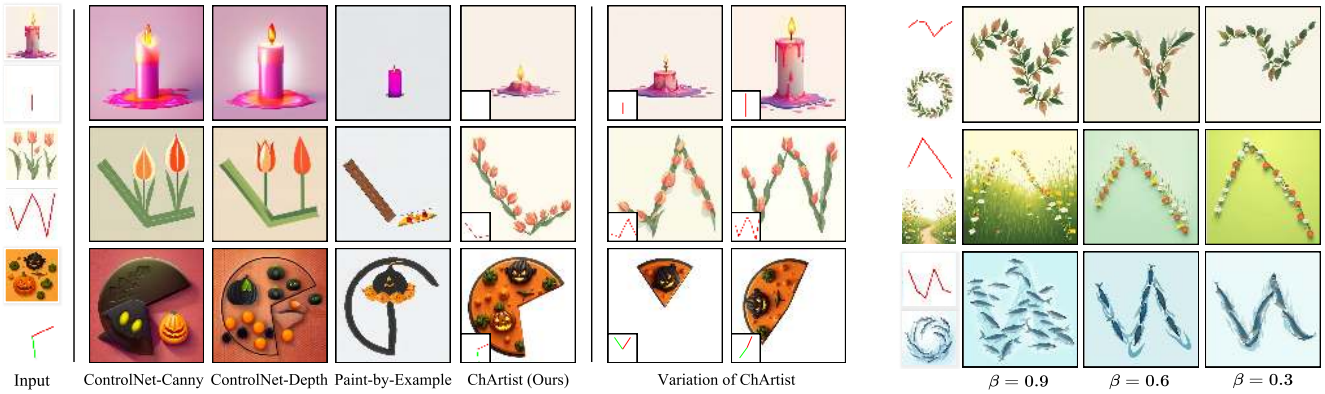


Figure 8. (a) Dual-control generation results conditioned on both spatial structure and subject reference. (b) Results of ChArtsit with different β showcases the importance of Spatially Gated Attention of dual control in ensuring the data accuracy.

Table 4. Ablation on control factor β on line pictorial chart.

β	Data Acc ↑	DINO ↑	CLIP-T ↑
0.3	**0.927**	0.732	0.324
0.6	0.876	0.748	**0.349**
0.9	0.729	**0.775**	0.337

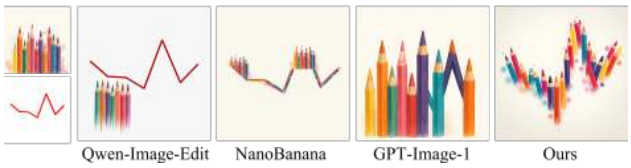


Figure 9. Comparison with current SoTA image editing models.

Subject Control Factor β Ablation. We conduct an ablation study to evaluate the effectiveness of our spatially-gated attention mechanism (Sect. 4.3). Fig. 8(b) presents qualitative results as the suppression factor β varies across different settings. As β decreases, the gating more strongly suppresses attention to non-chart regions. This helps prevent reference-image content from leaking into the background; for example, with $\beta = 0.9$, the model tends to copy the the background of reference image, whereas lowering β effectively mitigates this issue (second row of Fig. 8 (b)). Meanwhile, stronger suppression encourages the model to focus more on the designated chart region, improving data accuracy, as shown by the result with β set to 0.3. The quantitative results in Tab. 4 corroborate these observations, showing how different β values affect data accuracy. Besides, they also show a trade-off between data accuracy and visual consistency, where the best data accuracy is not accompanied by the highest visual consistency score.

6. Conclusion

We introduce ChArtist, a pipeline for generating pictorial charts with text and image references. We address the core challenge of balancing data faithfulness and visual expressiveness by providing both spatial and subject-level control, which can be used independently or composed jointly. We propose a skeleton-based representation, a minimal abstraction that encodes only the core data dimensions, preserving freedom for visual integration. We train spatial and subject LoRA separately and employ Spatially-Gated Attention during inference to mitigate their mutual interferences. We support our method with a large-scale dataset, ChArtist-30K, and a data accuracy metric for quantifying the faithfulness of generated charts. We demonstrate its usability and compatibility in the current design workflow, see A.4. We hope that it will unlock the door for further research of image generation to empower data-driven visual storytelling.

References

[1] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, and et al. Gemini: A family of highly capable multimodal models, 2025. 6

[2] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018. 3

[3] Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[4] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W Grant, Irene Reppa, Luciano Floridi, and Min Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012. 1

[5] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 6

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2

[8] Darius Coelho and Klaus Mueller. Infomages: Embedding data into thematic images. In *Computer Graphics Forum*, pages 593–606. Wiley Online Library, 2020. 1, 2, 3

[9] Weiwei Cui, Jinpeng Wang, He Huang, Yun Wang, Chin-Yew Lin, Haidong Zhang, and Dongmei Zhang. A mixed-initiative approach to reusing infographic charts. *IEEE Transactions on Visualization and Computer Graphics*, 28 (1):173–183, 2021. 2

[10] Noa Fish, Lilach Perry, Amit Bermano, and Daniel Cohen-Or. Sketchpatch: Sketch stylization via seamless patch-level synthesis. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 3

[11] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025. 3

[12] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024. 3

[13] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 3

[14] Steve Haroz, Robert Kosara, and Steven L Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200, 2015. 1

[15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[16] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 3

[17] Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. Word-as-image for semantic typography. *ACM Transactions on Graphics (TOG)*, 42(4): 1–11, 2023. 3

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[19] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 2

[20] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 6

[21] Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. Diffblender: Scalable and composable multimodal text-to-image diffusion models. *arXiv preprint arXiv:2305.15194*, 2023. 2

[22] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6850–6861, 2023. 2

[23] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 3

[24] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[25] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and

Jiasen Lu. One diffusion to generate them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2671–2682, 2025. 2

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023. 2

[27] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. *arXiv preprint arXiv:2404.06451*, 2024. 2

[28] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8217–8227, 2024. 2

[29] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *CVPR*, 2024. 3

[30] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 2

[31] 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. In *International Conference on Learning Representations*, 2022. 6

[32] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2

[33] OpenAI. Gpt-image-1: Openai image generation model. https://developers.openai.com/api/docs/models/gpt-image-1, 2025. Accessed: March 2026. 6

[34] Ji Hwan Park, Arie Kaufman, and Klaus Mueller. Graphoto: Aesthetically pleasing charts for casual information visualization. *IEEE computer graphics and applications*, 38(6): 67–82, 2019. 1, 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 6

[36] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[37] Yang Shi, Pei Liu, Siji Chen, Mengdi Sun, and Nan Cao. Supporting expressive and faithful pictorial visualization design with visual style transfer. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):236–246, 2022. 2

[38] Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7986–7996, 2025. 3, 4

[39] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 3

[40] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[41] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 2, 3, 4, 5, 6

[42] Purva Tendulkar, Kalpesh Krishna, Ramprasaath R Selvaraju, and Devi Parikh. Trick or treat: Thematic reinforcement for artistic typography. *arXiv preprint arXiv:1903.07820*, 2019. 3

[43] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[44] Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, et al. Unicombine: Unified multi-conditional combination with diffusion transformer. *arXiv preprint arXiv:2503.09277*, 2025. 3, 4

[45] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[46] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023. 3

[47] 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. 6

[48] Jiaqi Wu, John Joon Young Chung, and Eytan Adar. viz2viz: Prompt-driven stylized visualization generation using a diffusion model. *arXiv preprint arXiv:2304.01919*, 2023. 1, 2

[49] Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):284–294, 2023. 1, 2, 3

[50] Shishi Xiao, Liangwei Wang, Xiaojuan Ma, and Wei Zeng. Typedance: Creating semantic typographic logos from image through personalized generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024. 3

[51] Mingliang Xu, Qingfeng Li, Jianwei Niu, Hao Su, Xiting Liu, Weiwei Xu, Pei Lv, Bing Zhou, and Yi Yang. Art-up: A novel method for generating scanning-robust aesthetic qr codes. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 17(1):1–23, 2021. 3

[52] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 6

[53] Shuai Yang, Jiaying Liu, Wenhan Yang, and Zongming Guo. Context-aware unsupervised text stylization. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1688–1696, 2018. 3

[54] Shuai Yang, Jiaying Liu, Wenjing Wang, and Zongming Guo. Tet-gan: Text effects transfer via stylization and destylization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1238–1245, 2019. 3

[55] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 6

[56] 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. 6

[57] Jiayi Eris Zhang, Nicole Sultanum, Anastasia Bezerianos, and Fanny Chevalier. Dataquilt: Extracting visual elements from images to craft pictorial visualizations. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020. 3

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 6

[59] Yongtai Zhang, Shihong Deng, Zhihong Liu, and Yongtao Wang. Aesthetic qr codes based on two-stage image blending. In *International Conference on Multimedia Modeling*, pages 183–194. Springer, 2015. 3

[60] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[61] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023. 2

[62] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2

[63] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. 3

[64] Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. *arXiv preprint arXiv:2505.18612*, 2025. 3

[65] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2

ChArtist: Generating Pictorial Charts with Unified Spatial and Subject Control

Supplementary Material

A. Evaluation

A.1. Data Accuracy Evaluation Details

We aim to measure how faithfully the generated image represents data using a structure-aware F1 score.

Preprocessing. Before scoring, we first remove the background of each pictorial chart and apply a slight blur to its alpha channel to handle stylistic variance, such as hollow or soft strokes. For most chart types, we then collect all pixels with $\alpha > 0$ as the foreground region of the generated chart. For bar charts, however, we use the bounding box of the non-transparent region as the effective area, since slanted objects (e.g., Pisa tower) may not fully overlap with the skeleton despite being visually aligned.

Sampling-based F1 score. We randomly sample points from both the skeleton and the generated chart to approximate **precision** and **recall**:

$$\text{Precision} = \frac{|P \cap S|}{|P|}, \quad \text{Recall} = \frac{|P \cap S|}{|S|}. \tag{S1}$$

Here, P and S denote the sampled point sets from the generated chart and the skeleton, respectively. The overall score is computed as the harmonic mean of both terms:

$$F1score = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall}) + \epsilon}. \tag{S2}$$

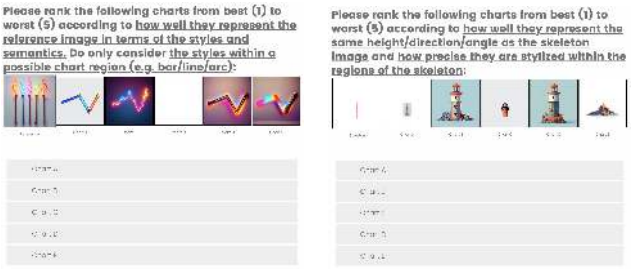
Weighted Region Scheme. Since points proximity to the skeleton indicates varying importance, we introduce multi-level weighted regions based on distance. for instance, for each region i, we sample a fixed number of points and compute weighted precision as:

$$\text{Precision}_\text{w} = \frac{\sum_i w_i \times |P_i \cap S|}{\sum_i w_i \times |P_i|}, \tag{S3}$$

Structure-aware Adaptation. We further adapt the weighting scheme to each chart type guided by their data encoding, ensuring that regions most relevant to data semantics receive higher importance.

A.2. Significance Tests

For the controlled online study results, as shown in Fig. S1, a Friedman test revealed no significant difference among methods for Skeleton-condition questions ($\chi^2(4) = 4.02, p = 0.403$), indicating comparable perceived quality compared against the structure of the chart skeleton. For Reference-condition questions, where participants had target reference image, the difference was significant ($\chi^2(4) = 24.52, p < 0.001$). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction ($\alpha = 0.005$) identified significant pairwise differences. Specifically, Chartist significantly outperformed InContext ($p < 0.001$), on par with the other top-ranked methods (Depth and Canny), and was among the top two methods in average ranking ($avg = 2.85$).



(a) Reference question sample (b) Skeleton question sample

Figure S1. Controlled online study questions for evaluating (a) style and semantic representation quality, and (b) skeleton alignment accuracy and precision.

Table S1. Human evaluation results from controlled online study (300 participants). Average rank ranges from 1 (best) to N (worst).

Method	Spatial (150 Questions)		Subject (150 Questions)		Method
	Avg Rank	Kendall τ	Avg Rank	Kendall τ	
ChArtist (Ours)	2.957 (2nd)	0.052	2.845 (2nd)	0.080	ChArtist (Ours)
SDEdit	2.897 (1st)	0.070	3.085	0.086	Paint-by-Example
ControlNet-Depth	2.969 (3rd)	0.058	2.822 (1st)	0.057	ControlNet-Depth
ControlNet-Canny	3.008	0.051	2.914 (3rd)	0.069	ControlNet-Canny
Inpainting	3.168	0.074	–	–	–

While structural control (Skeleton questions) produced minimal perceptual differentiation, Chartist maintained strong relative preference, demonstrated by their average ranks, under conditions requiring *both* faithful spatial correspondence and visual detail—suggesting that its control representation satisfies a niche that prefers a balance of semantic alignment and pictorial fidelity.

A.3. Evaluation Details

Implementation of Baseline. For the Spatially Aligned Only task (Task 1), all baseline methods use FLUX as the backbone model. For the subject-guided task, we use SDXL combined with ControlNet and IP-Adapter. Since the default line width is too thin to effectively present visual elements, we increase it to 30 pixels when creating the mask condition. Several baseline methods are sensitive to hyperparameters: we set the strength factor of SDEdit to 0.75 and the conditioning factor of ControlNet to 0.9.

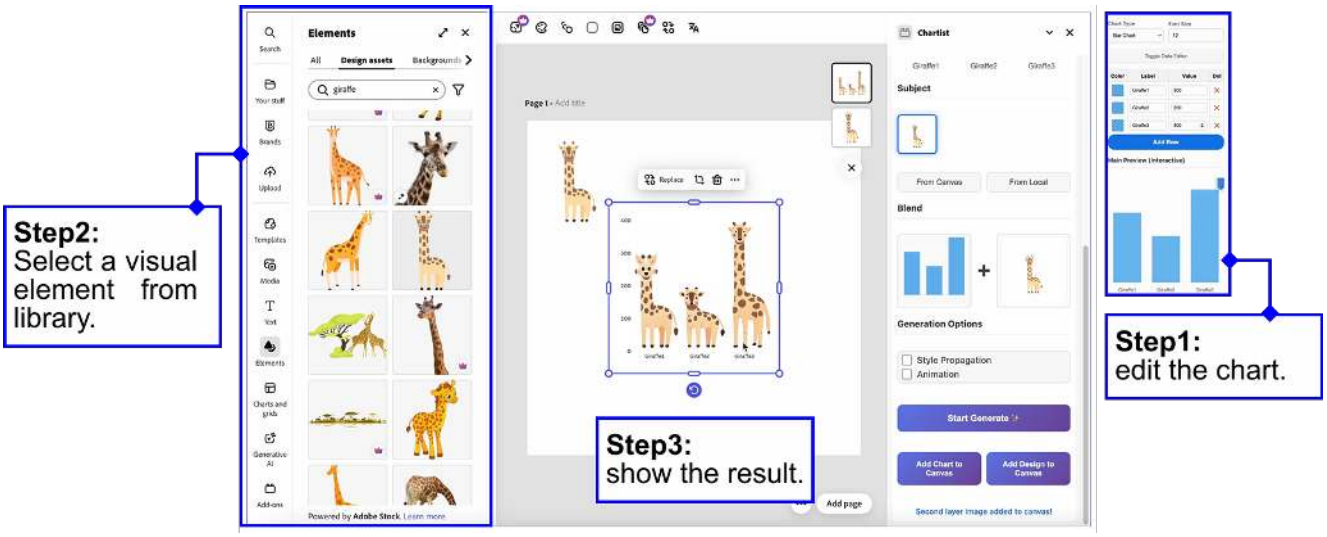


Figure S2. Interface of ChArtist build in existing design platform. The left side is the visual library, the right side is chart and blending configuration. Canvas is in the central for user manipulation.

Training Details For training both $LoRA_R$ and $LoRA_S$, we use the default prompt: "this item, in a white background." The training dataset is highly diversified through the use of various style LoRAs, including cartoon, 3D, illustration, photorealistic, and more.

A.4. Real-World Application

To illustrate the practical design value of ChArtist, we integrate it into a design software environment as a plugin (Fig. S2 and video in the supplementary). We observe that in most design platforms such as Adobe Express and Canva, visual design and chart creation are handled as two separate workflows. The visual-design workflow is rich, flexible, and supported by extensive asset libraries and recommendation systems, while chart creation is relatively limited, typically allowing customization only through simple color adjustments. Built on Adobe Express, our goal is to demonstrate ChArtist's ability to unify visual design and chart creation into a single workflow. First, user can edit chart data in the right panel, and then select a visual element from the asset library. The blended result will be shown on the central canvas, which can be further added with axis and annotation. We additionally incorporate external models to support a more complete design pipeline. For example, we employ StyleAligned to maintain stylistic consistency across visual elements, and we use an external image-to-video API to animate the final outputs.

A.5. More Experimental Results

We provide additional qualitative results:

- Spatially Aligned Only Task: Fig. S3, Fig. S4.
- Dual-control generation conditioned on both spatial structure and subject reference: Fig. S5, Fig. S6, and Fig. S7.

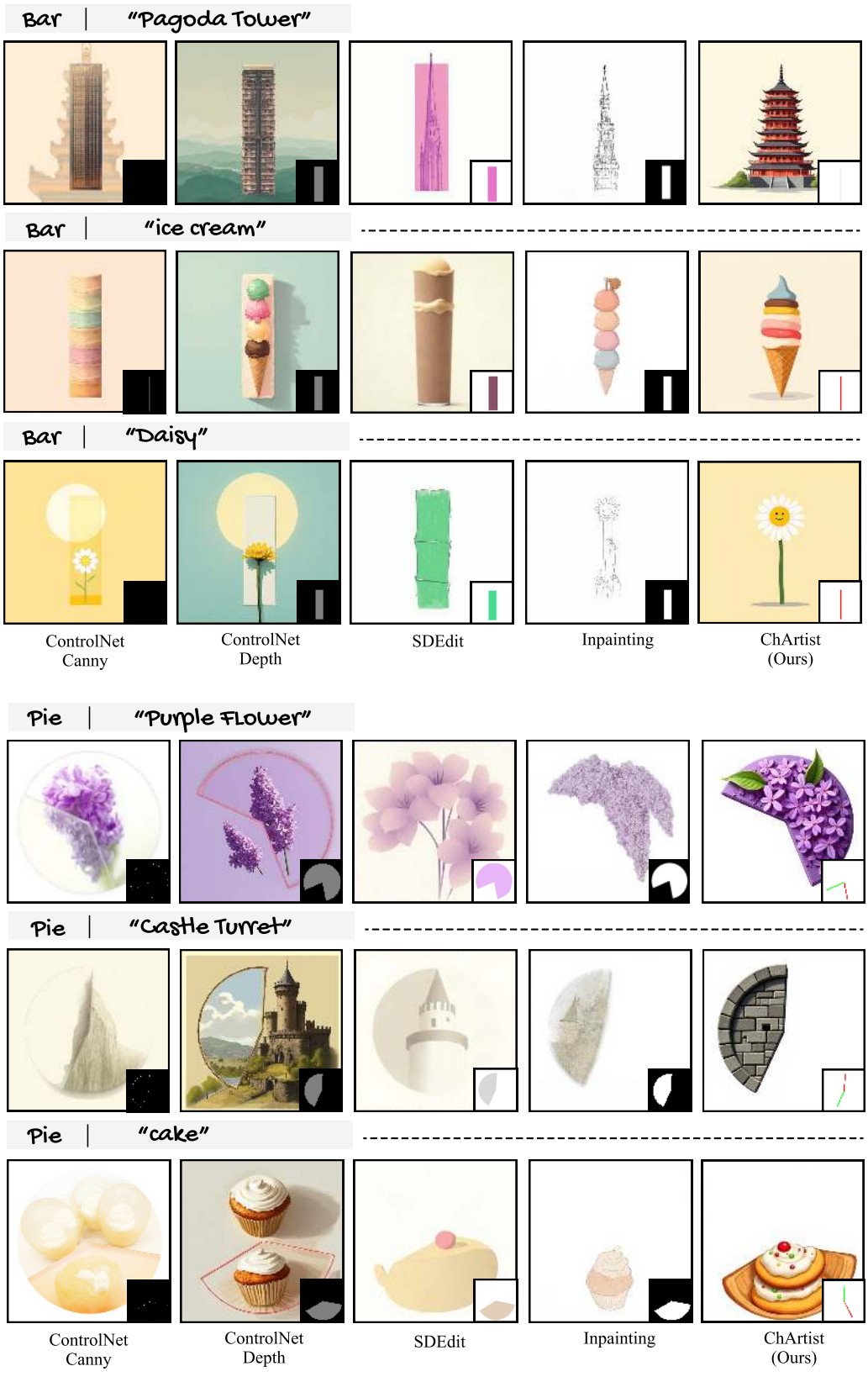


Figure S3. Result of spatially aligned evaluation with different control representations. (Task 1).

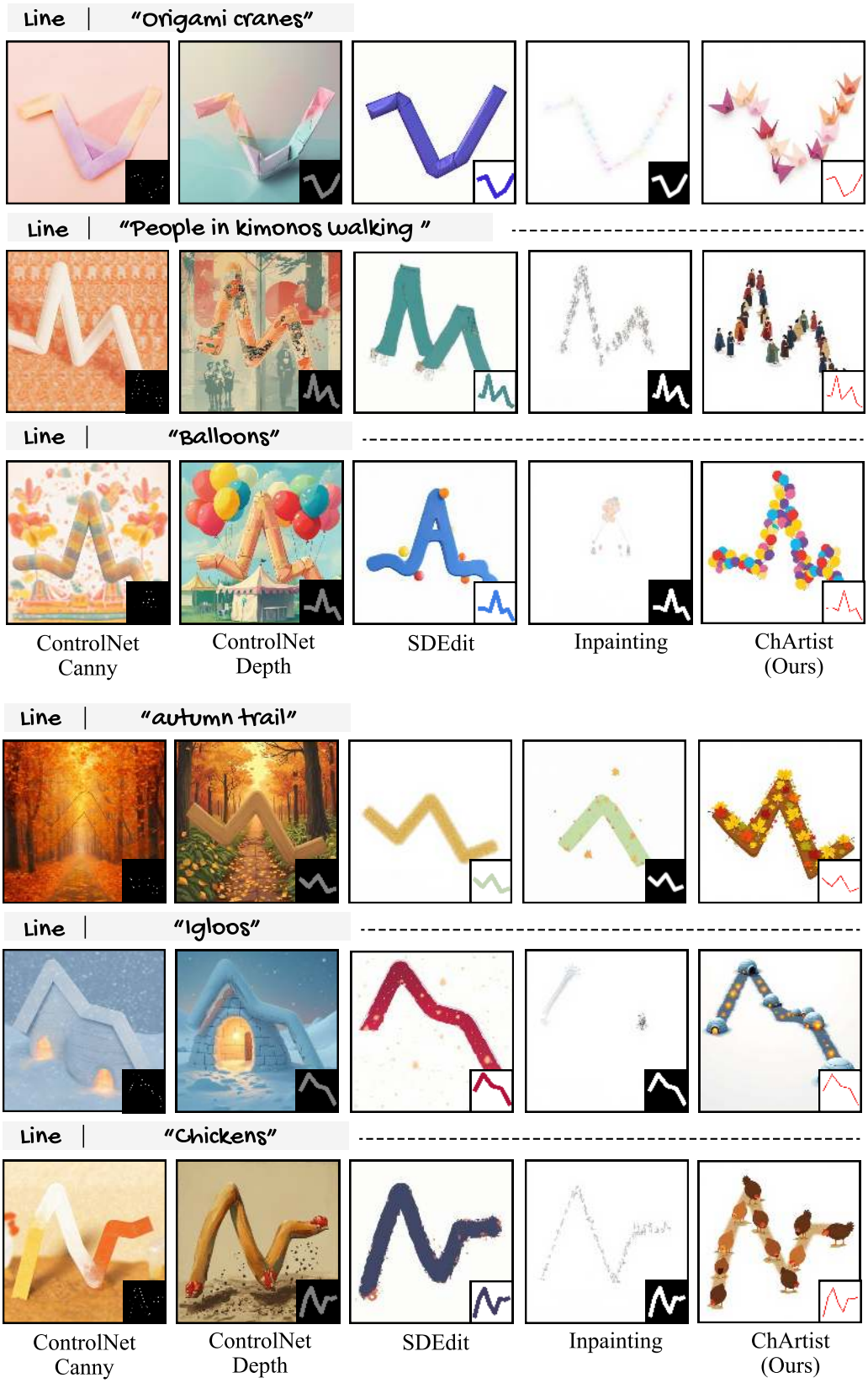


Figure S4. Result of spatially aligned evaluation with different control representations. (Task 1).

Figure S5. Dual-control generation bar pictorial results conditioned on both spatial structure and subject reference.

Figure S6. Dual-control generation pie pictorial results conditioned on both spatial structure and subject reference.

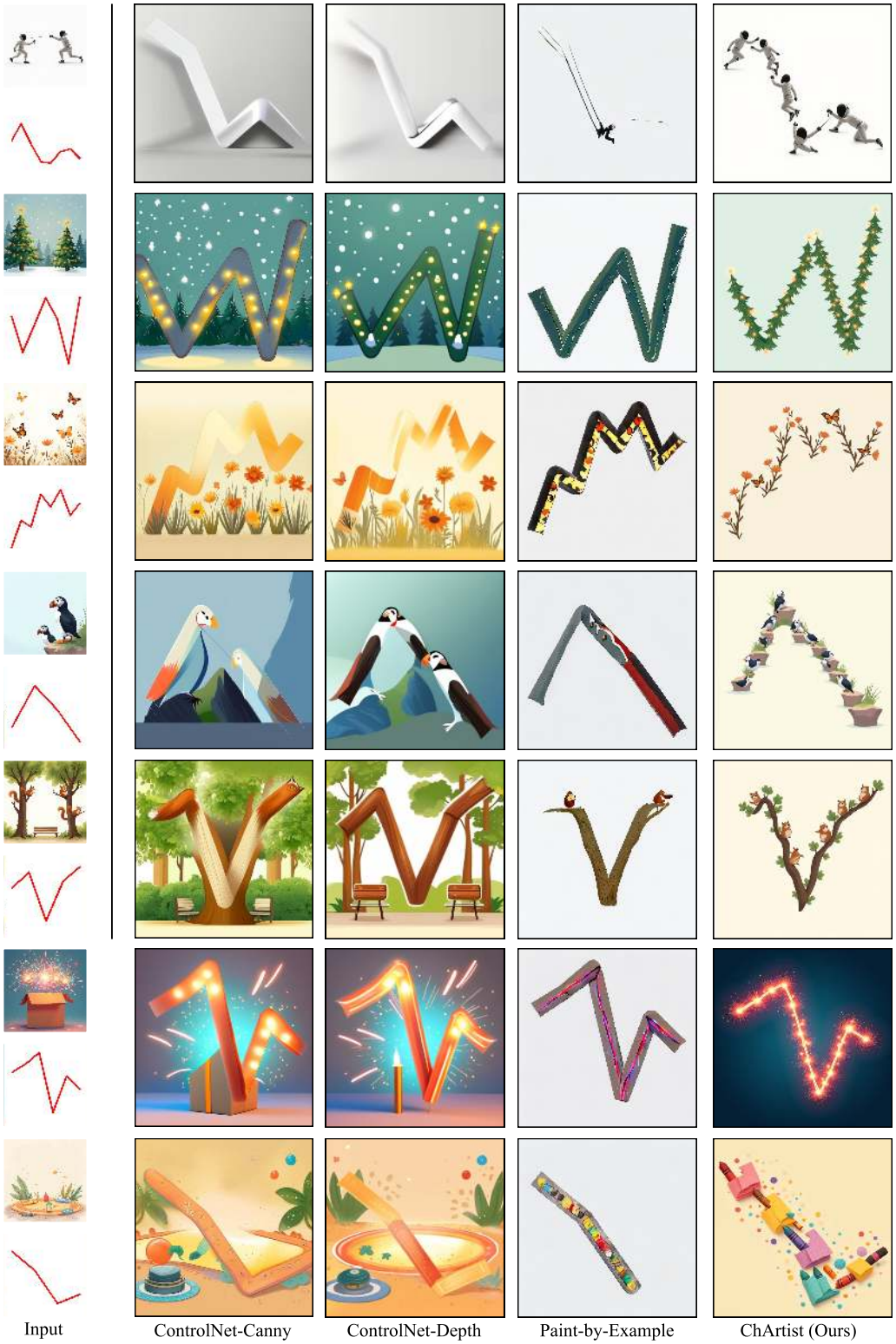


Figure S7. Dual-control generation line pictorial results conditioned on both spatial structure and subject reference.

SpriteForge: Consistent Multi-View Pixel Art Generation via LoRA-Enhanced Image Conditioning

Jiayi Li
Brown University
jiayi_li@brown.edu

Abstract

In this paper, we propose a conditioning-based approach using LoRA fine-tuning on a pre-trained image-conditioned diffusion model to generate four canonical views (front, back, left, and right) of a pixel art character from a single input image. Unlike general multi-view consistency methods, our task focuses on structured viewpoint transformation rather than full 3D-consistent generation. Recent work such as SyncDreamer demonstrates the ability to generate multi-view consistent images from a single-view input. However, maintaining consistency in color palettes and discrete structural details remains challenging for pixel art due to its low resolution and highly constrained visual style. To address this issue, we propose a lightweight conditioning strategy that learns view-specific transformations directly from paired multi-view pixel art data, enabling improved structural alignment and palette consistency across generated views.

1. Introduction

Pixel art is widely used in games and digital media due to its efficiency and distinct visual style. Generating consistent multi-view representations (e.g., front, back, left, and right) from a single input image is important for character design and animation. Unlike natural images, which can often be interpreted as projections of underlying 3D geometry, pixel art characters are manually designed 2D representations. Different views are constructed independently to convey pose and orientation, making multi-view generation a problem of learning structured, view-dependent transformations rather than enforcing geometric consistency, as illustrated in Figure 1.

Existing approaches for novel view generation can be broadly categorized into three directions. The first relies on text-to-image generation, where models are prompted to produce a target view (e.g., "back view of a pixel art character"). While flexible, this approach lacks explicit conditioning on the input image, often resulting in inconsistencies in structure, pose, and color. The second direction uses structure-guided methods such as ControlNet [9], which incorporate signals such as edge maps or segmentation to guide generation. While effective for natural images, these signals are often unreliable in pixel art due to low resolution and discrete boundaries, leading to poor structural and color consistency. A third direction involves image-conditioned diffusion models (e.g., Flux-style models), which take both a reference image and a text prompt as input. Although these models can preserve high-level appearance, they often exhibit weak binding between the input image, textual condition, and generated output, resulting in loss of fine-grained details and incorrect viewpoint generation.

Recent work such as SyncDreamer [6] demonstrates the ability to generate multi-view consistent images from a single-view input by modeling underlying 3D structure. However, such approaches assume geometric consistency and continuous appearance variation, which do not hold for pixel art, where views are manually designed and exhibit discrete transformations.

To address these limitations, we propose a LoRA-based fine-tuning approach for pixel art multi-view generation that strengthens the mapping between input appearance and target views. By learning view-specific transformations from paired multi-view data, the model generates consistent back, left, and right views from a given front view. This improves orientation correctness and structural alignment while preserving discrete color palettes and pixel-level details. We further construct a curated dataset of pixel art assets with aligned multi-view annotations to support training and evaluation.

2. Related Work

2.1. Text-to-Image Generation

Text-to-image diffusion models such as Stable Diffusion [8] have demonstrated strong ability to generate diverse visual content from textual prompts. These methods provide flexible control over high-level semantics such as object category and viewpoint. However, since generation is condi-

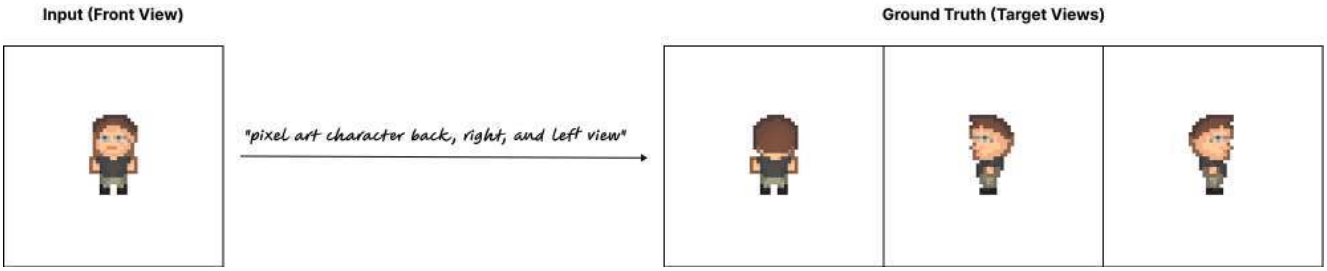


Figure 1. Overview of the multi-view pixel art generation task. Given a single front-view input, the model generates back, right, and left views conditioned on a text prompt.

tioned only on text, they lack explicit correspondence to a given input image, often leading to inconsistencies in structure, pose, and color when applied to view synthesis tasks.

2.2. Structure-Guided Generation

Structure-guided approaches improve controllability by conditioning on additional signals such as edge maps or segmentation masks. Frameworks such as ControlNet [9] incorporate these structural constraints into diffusion models to guide generation, while image-to-image translation methods such as pix2pix [4] learn mappings between paired inputs and targets. However, these methods rely on reliable structural signals or consistent training pairs. In pixel art, where images are low-resolution and composed of discrete pixels, such signals are often ambiguous or noisy. As a result, these approaches struggle to maintain structural alignment and consistent color reproduction across views.

2.3. Image-Conditioned and Multi-View Generation

Image-conditioned generative models take both a reference image and a text prompt to guide synthesis, aiming to preserve appearance while enabling controllable transformations. Methods based on diffusion models can maintain high-level visual consistency, but often exhibit weak binding between the input image and generated output.

Multi-view generation approaches such as SyncDreamer [6] model underlying 3D structure to enforce cross-view consistency. However, these methods assume geometric consistency and continuous appearance variation, which do not hold for pixel art. In this setting, views are manually designed and involve discrete transformations, making such assumptions ineffective.

3. Method

3.1. Base Model

We build upon Flux2Klein, a multi-view image generation model based on the FLUX architecture [5]. Flux2Klein extends the base transformer with a sequence-concatenation conditioning mechanism, allowing a reference image to be incorporated as a conditioning signal during generation without modifying the underlying architecture.

3.2. Sequence-Concatenation Conditioning

Given a conditioning image x_c and a target image x_t, both are encoded into latent space using a pretrained VAE encoder $\mathcal{E}$:

$$z_c = \mathcal{E}(x_c), \quad z_t = \mathcal{E}(x_t) \tag{1}$$

The latent representations are then patchified and flattened into token sequences. During training, the noisy target tokens are concatenated with the conditioning tokens along the sequence dimension:

$$p_{\text{input}} = [p_t \,\|\, p_c] \tag{2}$$

where p_t denotes the noisy target tokens and p_c the conditioning tokens.

To distinguish the two parts, positional indices of the conditioning tokens are offset relative to the target tokens. The transformer is trained to predict only the target tokens, while the conditioning tokens serve as read-only context. This design enables direct cross-attention between input and target representations, strengthening input–output binding without introducing additional architectural components.

3.3. LoRA Fine-Tuning

We adapt the model using Low-Rank Adaptation [3]. LoRA layers are applied to the attention and feed-forward modules of the transformer, while all base model parameters remain frozen. We use rank $r = 32$ and scaling factor $\alpha = 32$, enabling efficient fine-tuning with a small number of trainable parameters.

3.4. Training and Inference

We construct a curated dataset of pixel art assets with aligned multi-view annotations. Each sample consists of a front-view conditioning image and a target view (back or side), along with a corresponding text prompt. All images

are resized to 512×512 using nearest-neighbor interpolation to preserve discrete pixel structures. Left-view supervision is omitted during training and obtained via horizontal flipping at inference.

We adopt a flow matching formulation. For each sample, a timestep $t \sim \mathcal{U}(0, 1)$ is sampled and the latent is interpolated with Gaussian noise:

$$z_t = (1 - t)z_0 + t\epsilon \tag{3}$$

The model learns to predict a velocity field that transforms noisy latents toward clean representations, conditioned on both the input image and text prompt. Training is performed using AdamW with a learning rate of 5×10^{-5} for 10,000 steps with batch size 1 on a single NVIDIA RTX A6000 (48GB VRAM). At inference, the front-view image is used as conditioning input. The target latent is initialized from Gaussian noise and iteratively refined using classifier-free guidance ($s = 3.5$) for 28 steps.

4. Experiments

4.1. Baselines

We compare our method with representative approaches from three settings.

Multimodal generative systems. We evaluate publicly available multimodal systems, including GPT-4o [7], Gemini [2], and Claude [1], which take both an input image and a textual prompt to generate target views. These models provide strong general-purpose generation capabilities, but are not specifically designed for structured multi-view transformation.

Structure-guided rendering. We use ControlNet [9] with edge-based conditioning. This baseline evaluates whether structural guidance alone is sufficient for pixel art view generation.

Ablation baseline. We compare against Flux2Klein without LoRA fine-tuning to evaluate the effect of our proposed adaptation.

4.2. Evaluation Protocol

Given a front-view pixel art character, each method generates the corresponding back, left, and right views. Multimodal systems use both the input image and a textual prompt, while ControlNet uses edge-based conditioning.

We evaluate results using six criteria: *view rotation*, *color*, *pose*, *size*, *style*, and *structure*. Table 1 summarizes performance across these criteria based on visual assessment.

4.3. Qualitative Results

Figure 2 shows qualitative comparisons across different methods. Green boxes indicate correct alignment, while red boxes highlight structural or viewpoint errors. Multimodal systems generate plausible outputs but often fail to preserve consistent structure and scale. ControlNet maintains local structure but cannot perform reliable viewpoint transformation. Flux2Klein improves visual consistency but still shows weak alignment between input and generated views.

In contrast, our method produces consistent multi-view outputs with accurate viewpoint transformation and preserved pixel-level details.

4.4. Ablation Study

We evaluate the effect of LoRA fine-tuning by comparing our method with Flux2Klein without adaptation. Without LoRA, the model exhibits weaker alignment between input and generated views, leading to incorrect viewpoint transformation and structural inconsistencies. Fine-tuning with LoRA improves alignment, structural consistency, and color preservation.

5. Limitations

Despite the strong performance of our method, several limitations remain. First, our approach assumes approximate bilateral symmetry in pixel art characters and relies on horizontal flipping to generate left views. This assumption does not hold for asymmetric designs (e.g., characters with one-sided accessories), which may lead to incorrect or inconsistent results. Second, the model shows reduced performance on characters with very light or low-contrast boundary regions, where structural cues become ambiguous. In such cases, the generated views may exhibit weakened edge definition or structural inconsistencies. Third, our approach is designed for moderate-resolution pixel art and does not generalize well to extremely low-resolution inputs (e.g., 16×16 or 18×18), where limited pixel information makes consistent view transformation challenging. Finally, while our method improves structural consistency and viewpoint alignment, it is limited to static view generation and does not model temporal consistency. Extending the approach to animation or dynamic poses remains an open problem.

These limitations suggest directions for future work in handling asymmetric designs, improving robustness to low-contrast inputs, and extending the method to more complex and dynamic visual settings.

6. Conclusion

In this paper, we presented SpriteForge, a conditioning-based approach for multi-view pixel art generation that strengthens input–output binding through LoRA fine-tuning. Unlike prior methods that rely on text prompts or structural guidance, our approach directly learns view-specific transformations from paired multi-view data, enabling consistent generation of back, left, and right views from a single front-view input.

Figure 2. Qualitative comparison. Green boxes indicate correct alignment, red boxes indicate errors. Our method produces consistent multi-view outputs with correct structure and scale.

Method	View Rotation	Color	Pose	Size	Style	Structure	Open/Local
GPT-4o	✓	✓	✓	✓	✓	✓	✗
Claude	✓	✓	✓	✓	✓	✗	✗
Flux2Klein	✓	✓	✓	✗	✓	✗	✓
Gemini	✓	✓	✓	✗	✗	✗	✗
ControlNet [9]	✗	✗	✓	∼	∼	∼	✓
SpriteForge (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1. Qualitative comparison of multi-view pixel art sprite generation methods. ✓ indicates success, ✗ indicates failure, and ∼ indicates partial success.

Through qualitative comparisons, we demonstrate that existing multimodal and structure-guided methods struggle to preserve pixel-level structure, consistent scale, and accurate viewpoint alignment. In contrast, our method produces coherent multi-view outputs with stable structure, correct orientation, and faithful color reproduction.

These results highlight the importance of explicitly modeling view-dependent transformations for pixel art, where traditional geometric assumptions do not hold. Future work includes extending the approach to more complex animations, handling asymmetric designs, and incorporating higher-resolution or stylistically diverse pixel art datasets.

References

[1] Anthropic. Claude. https://www.anthropic.com/claude, 2024.

[2] Google. Gemini. https://deepmind.google/technologies/gemini/, 2024.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[5] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[6] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image, 2024.

[7] OpenAI. Gpt-4o. https://openai.com, 2024.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.