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

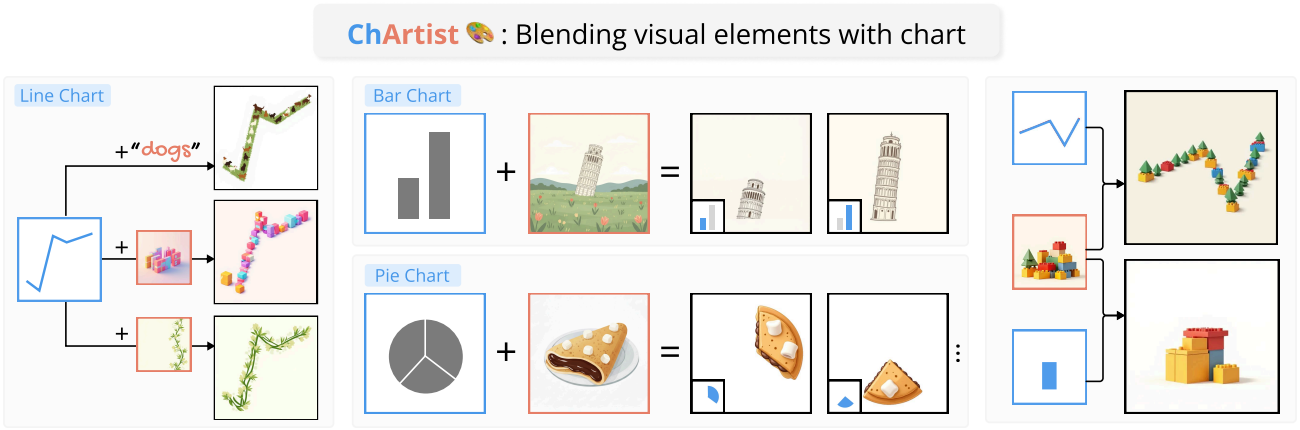


Figure 1. Illustrations of pictorial charts generated by ChArtist. We convert chart primitives such as bars, lines, and segments into vivid visual elements. With user-provided text or reference images, ChArtist combines spatial and subject control to maintain data fidelity while achieving visual consistency.

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).

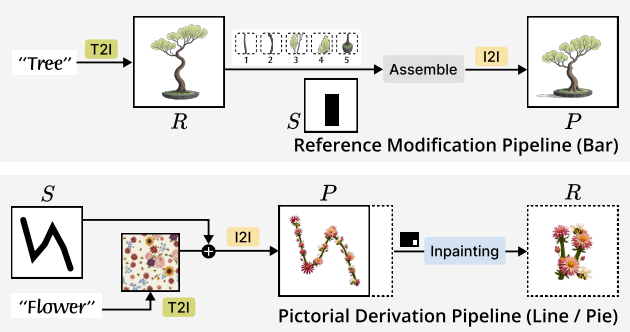


Figure 2. Pipelines for CHARTIST-30K Dataset Construction.

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.

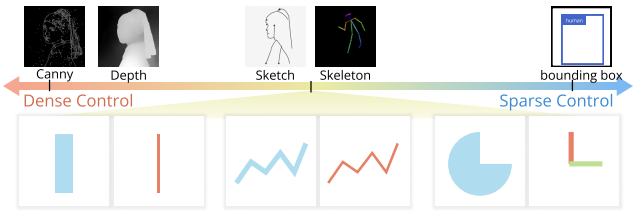


Figure 3. The spectrum of control representation based on their complexity. Top: Existing control representations for natural image; Bottom: Our proposed skeleton-based representations for pictorial chart generation, where charts (blue) are represented as skeletons (red and green lines).

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.

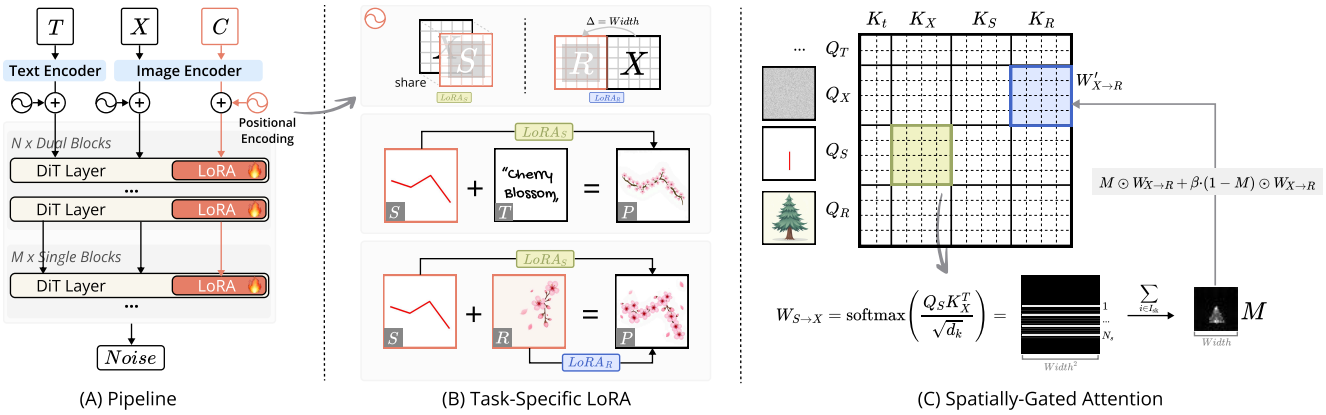


Figure 4. The whole architecture of ChArtist consists of (A) a pretrained DiT-based diffusion model with two conditional-LoRAs with different positional encoding on the image inputs, where (B) one is for spatial control ($LoRA_S$) and another for subject control ($LoRA_R$). (C) To avoid the interference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

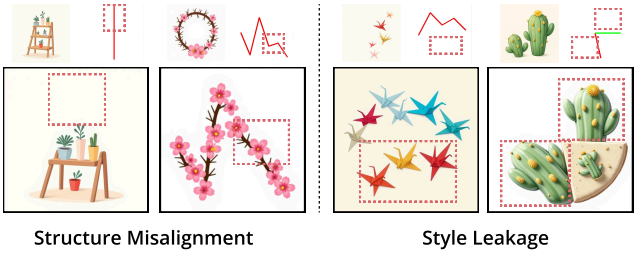


Figure 5. Artifacts observed when merging multiple LoRAs in parallel. $LoRA_R$ can distort the spatial control and disobey the chart structure (structure misalignment), and can also introduce extra visual elements beyond the spatial constraints (style leakage).

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

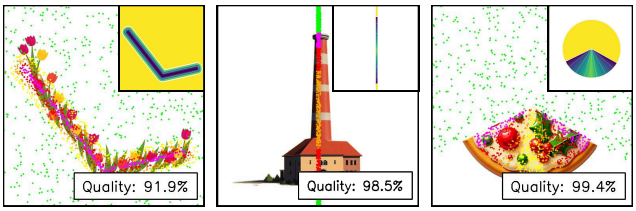


Figure 6. Illustrations of the data accuracy metric. We construct a distance field based on the chart skeleton. Then we randomly sample the points based on the ranges on the distance field, and calculate the accuracy based on a weighted F1 Score.

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

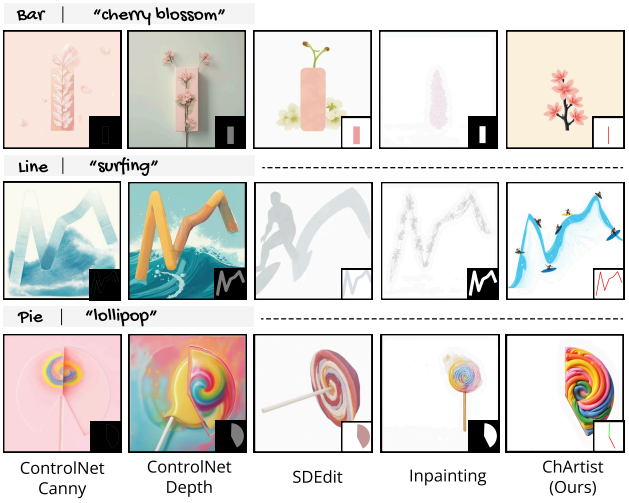


Figure 7. Result of spatially aligned evaluation with different control representations. (Task 1)

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

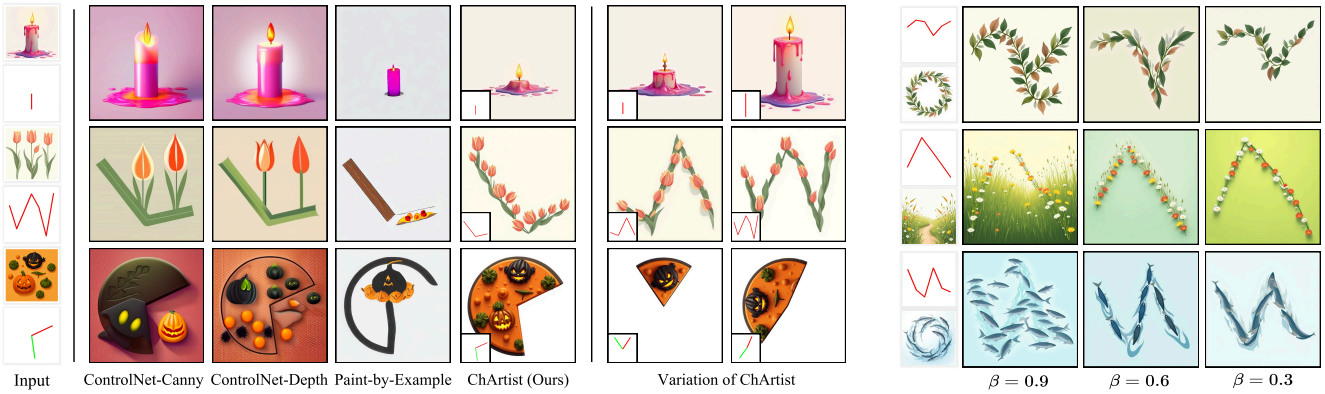


Figure 8. (a) Dual-control generation results conditioned on both spatial structure and subject reference. (b) Results of ChArtsit with different β showcases the importance of Spatially Gated Attention of dual control in ensuring the data accuracy.

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

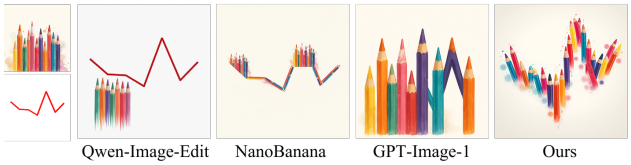


Figure 9. Comparison with current SoTA image editing models.

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:

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

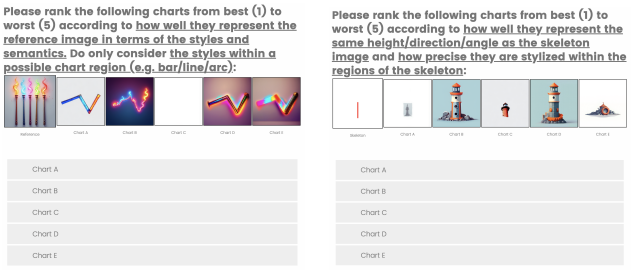
Weighted Region Scheme. Since points proximity to the skeleton indicates varying importance, we introduce multi-level weighted regions based on distance. for instance, for each region i, we sample a fixed number of points and compute weighted precision as:

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

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.

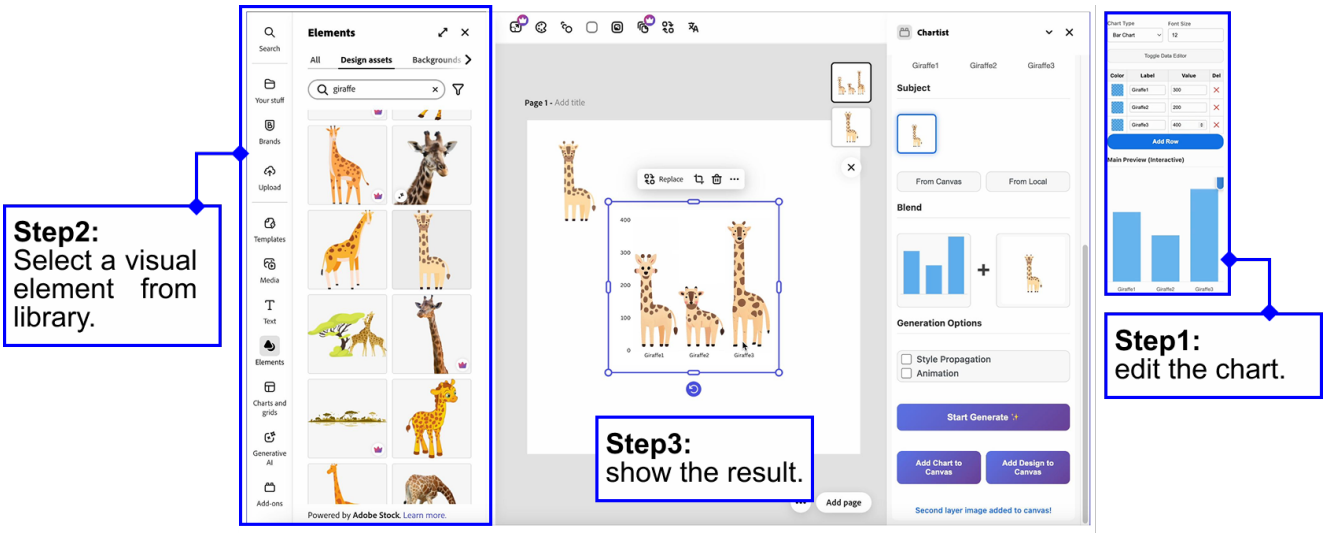


Figure S2. Interface of ChArtist build in existing design platform. The left side is the visual library, the right side is chart and blending configuration. Canvas is in the central for user manipulation.

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.

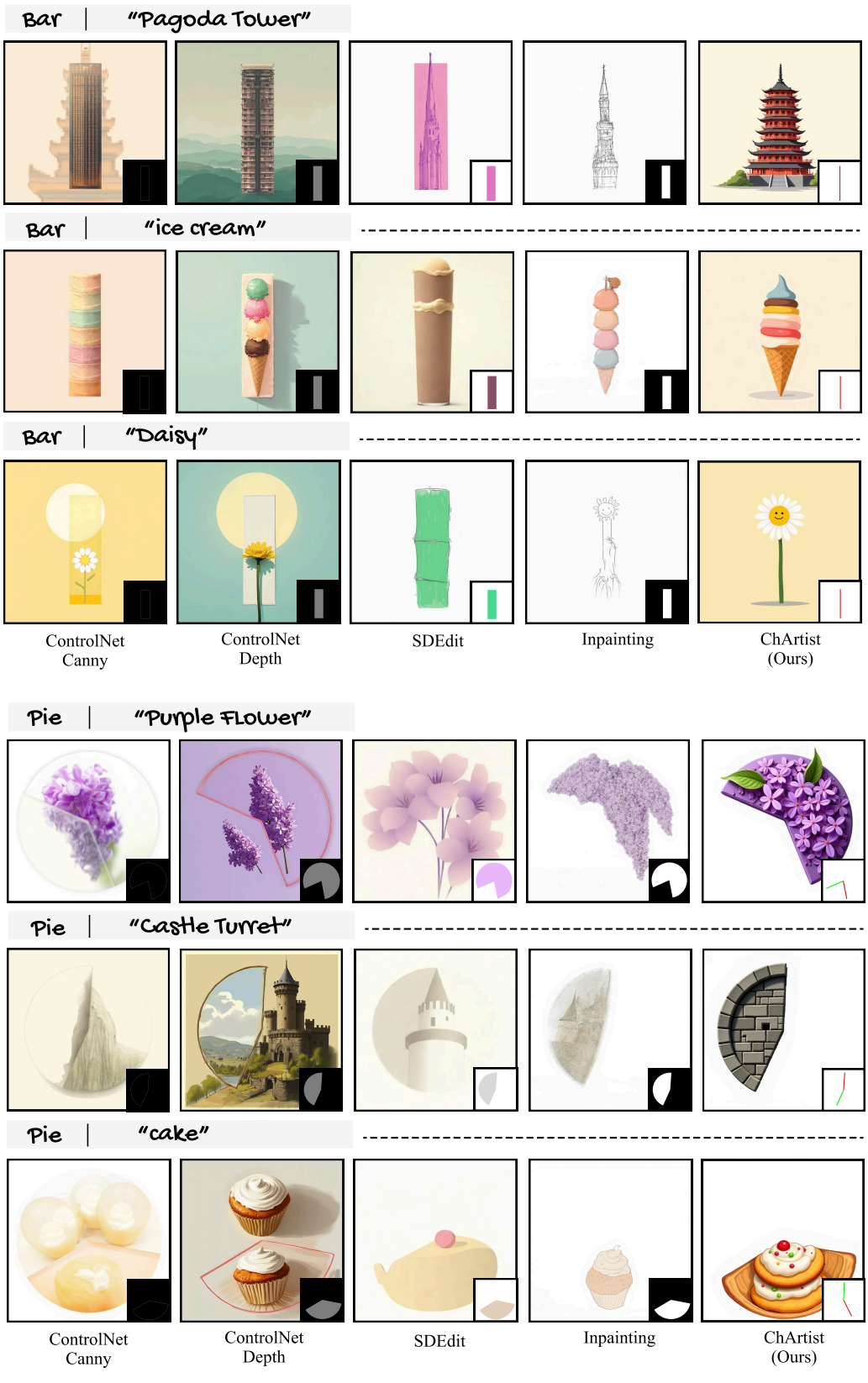


Figure S3. Result of spatially aligned evaluation with different control representations. (Task 1).

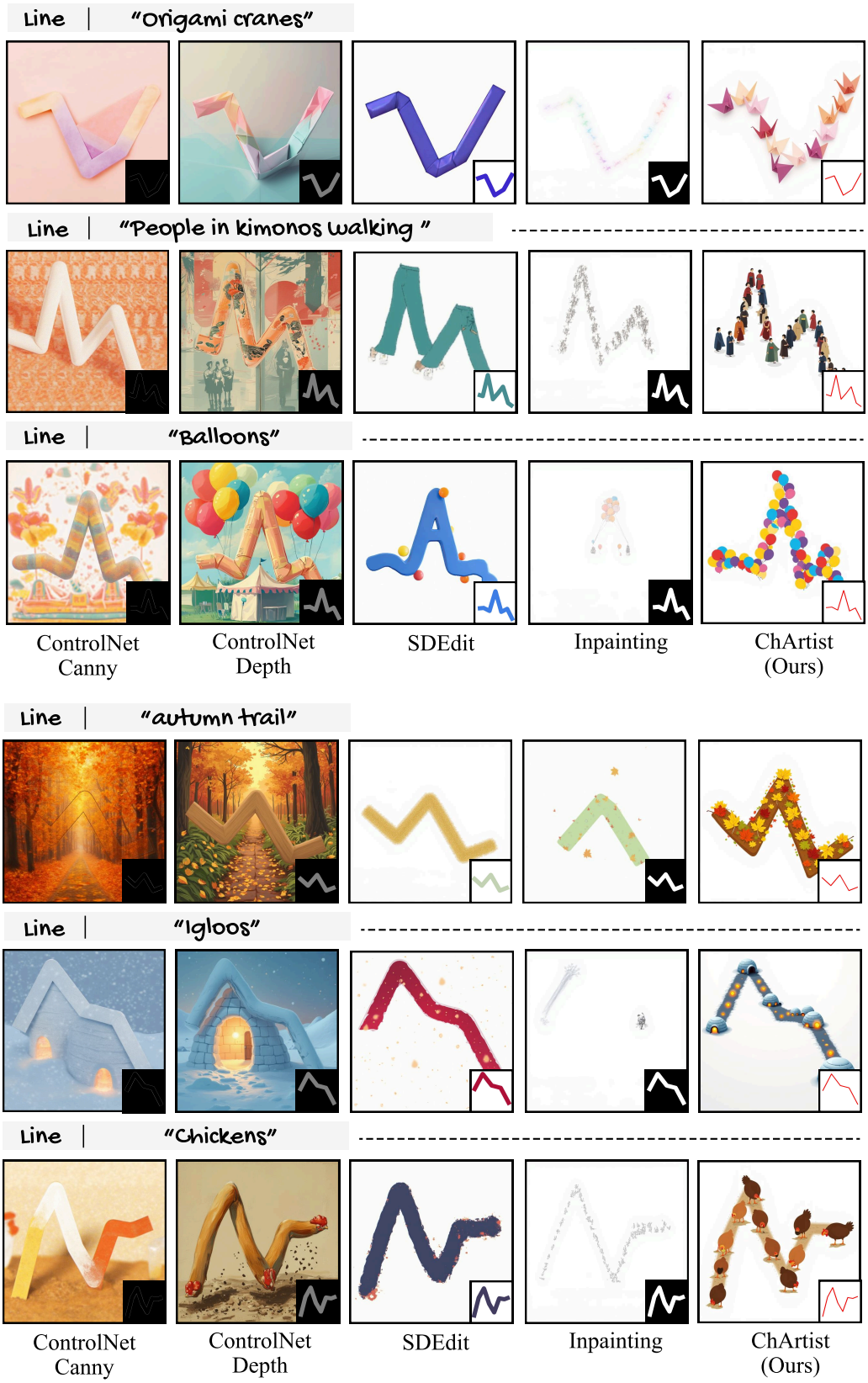


Figure S4. Result of spatially aligned evaluation with different control representations. (Task 1).

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.

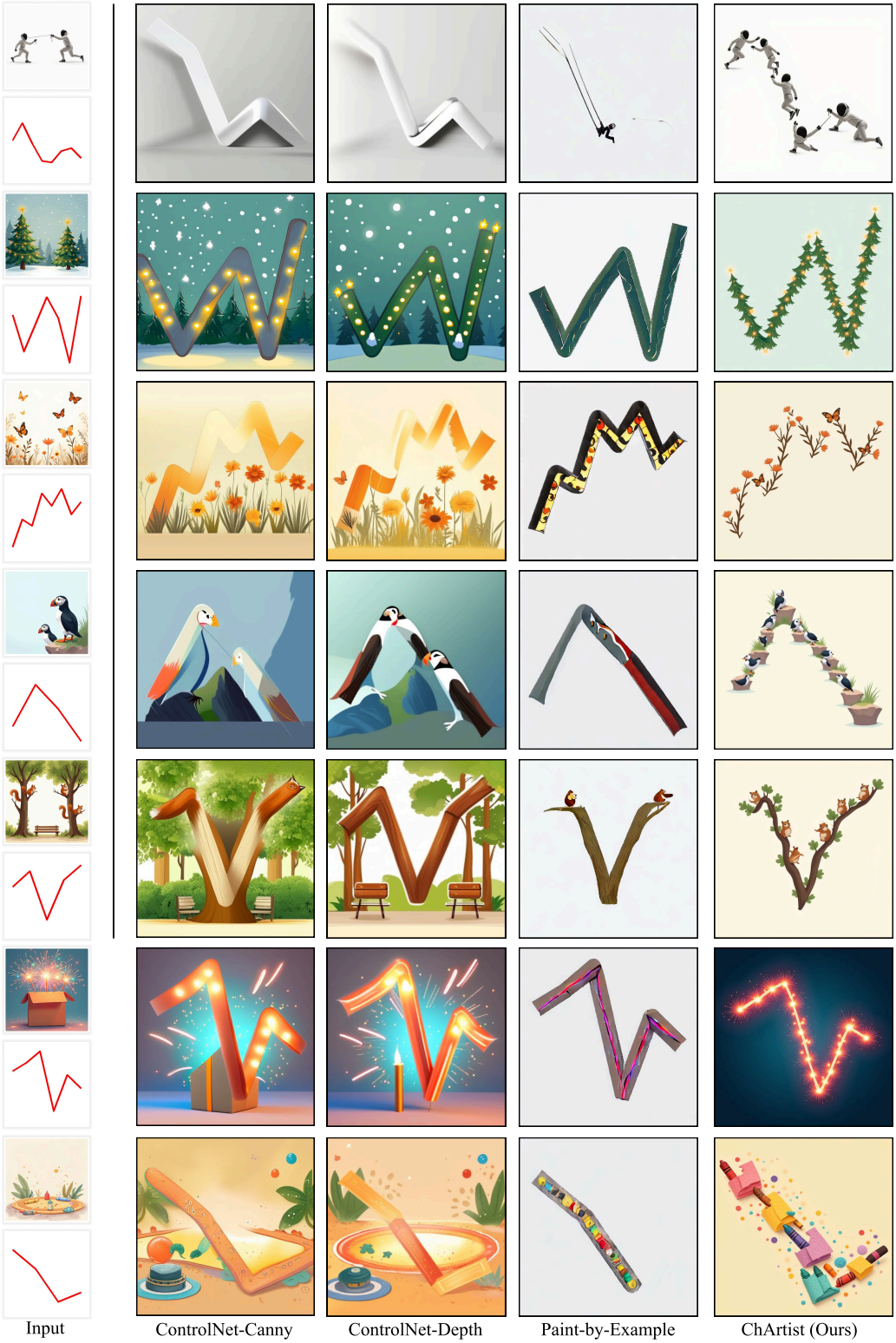


Figure S7. Dual-control generation line pictorial results conditioned on both spatial structure and subject reference.