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:

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

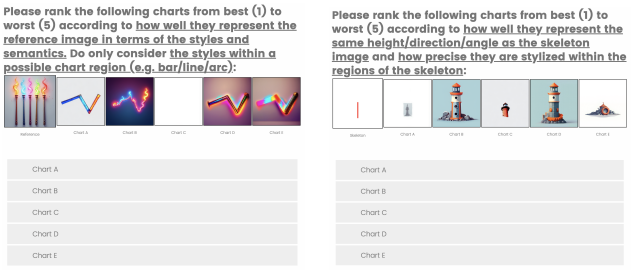
Weighted Region Scheme. Since points proximity to the skeleton indicates varying importance, we introduce multi-level weighted regions based on distance. for instance, for each region i, we sample a fixed number of points and compute weighted precision as:

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

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.

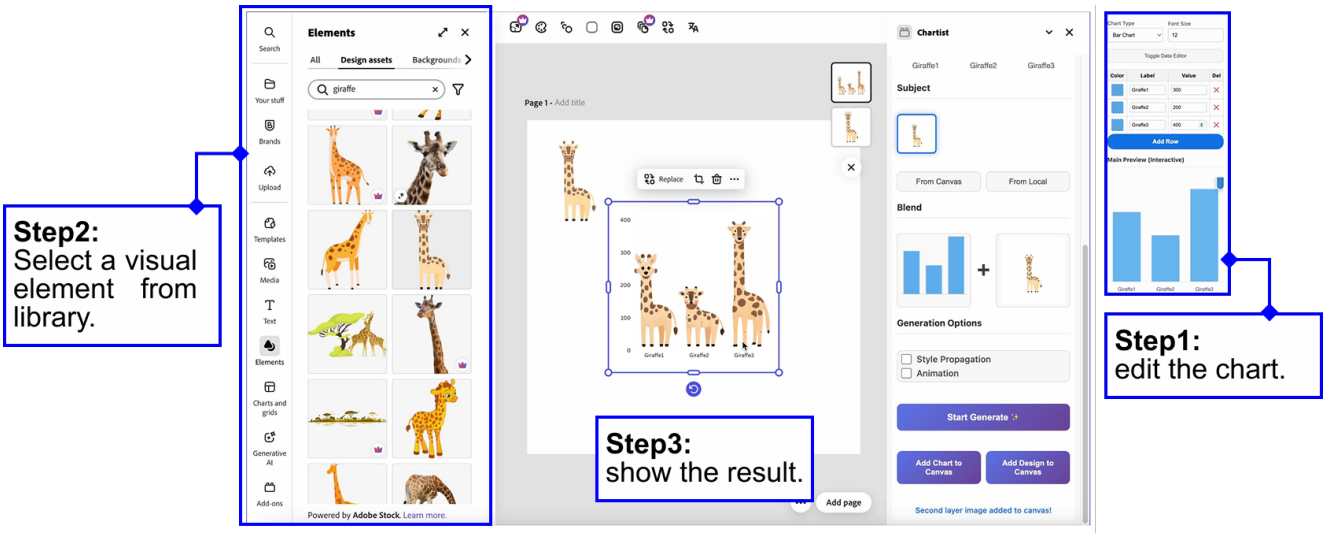


Figure S2. Interface of ChArtist build in existing design platform. The left side is the visual library, the right side is chart and blending configuration. Canvas is in the central for user manipulation.

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.

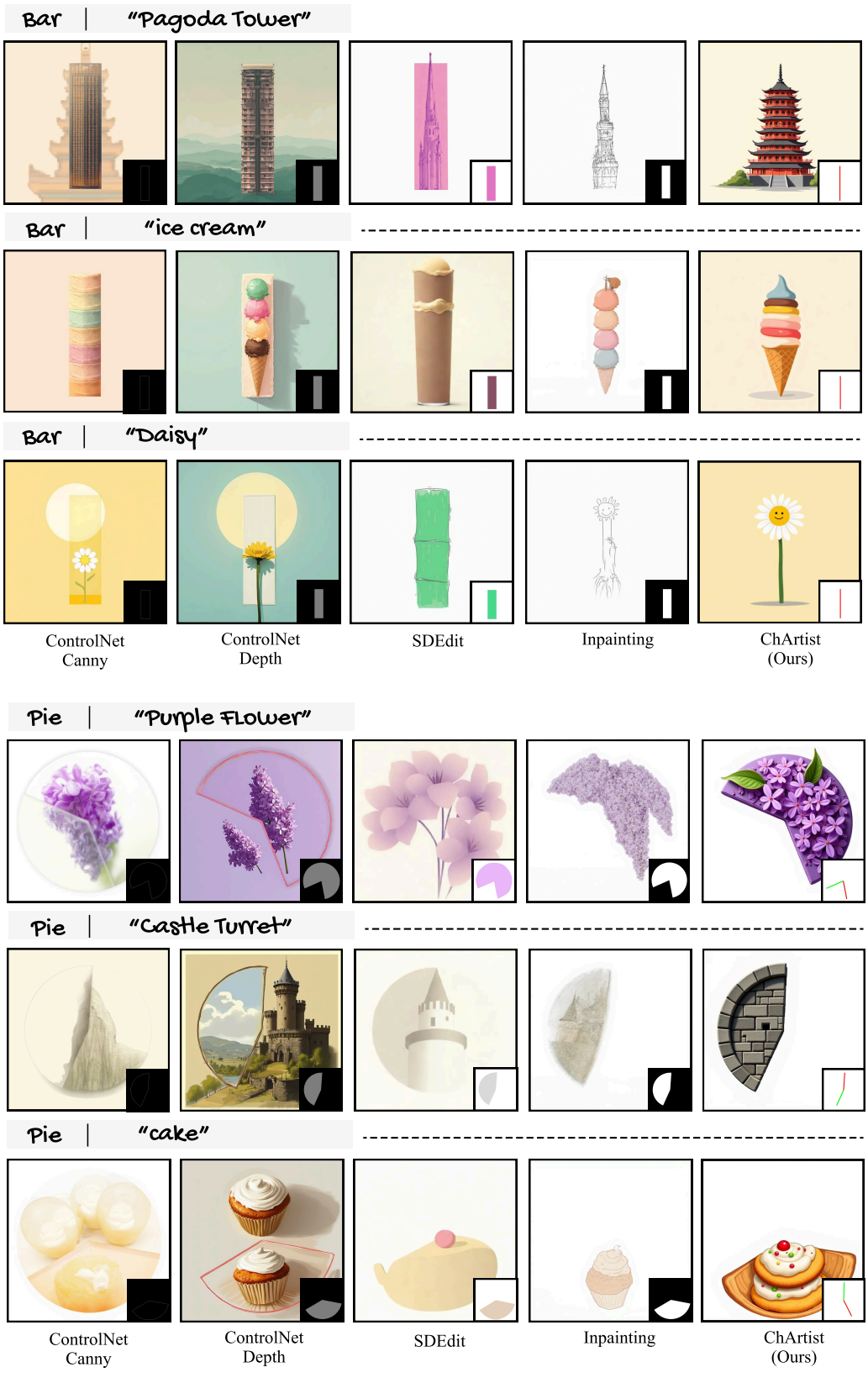


Figure S3. Result of spatially aligned evaluation with different control representations. (Task 1).

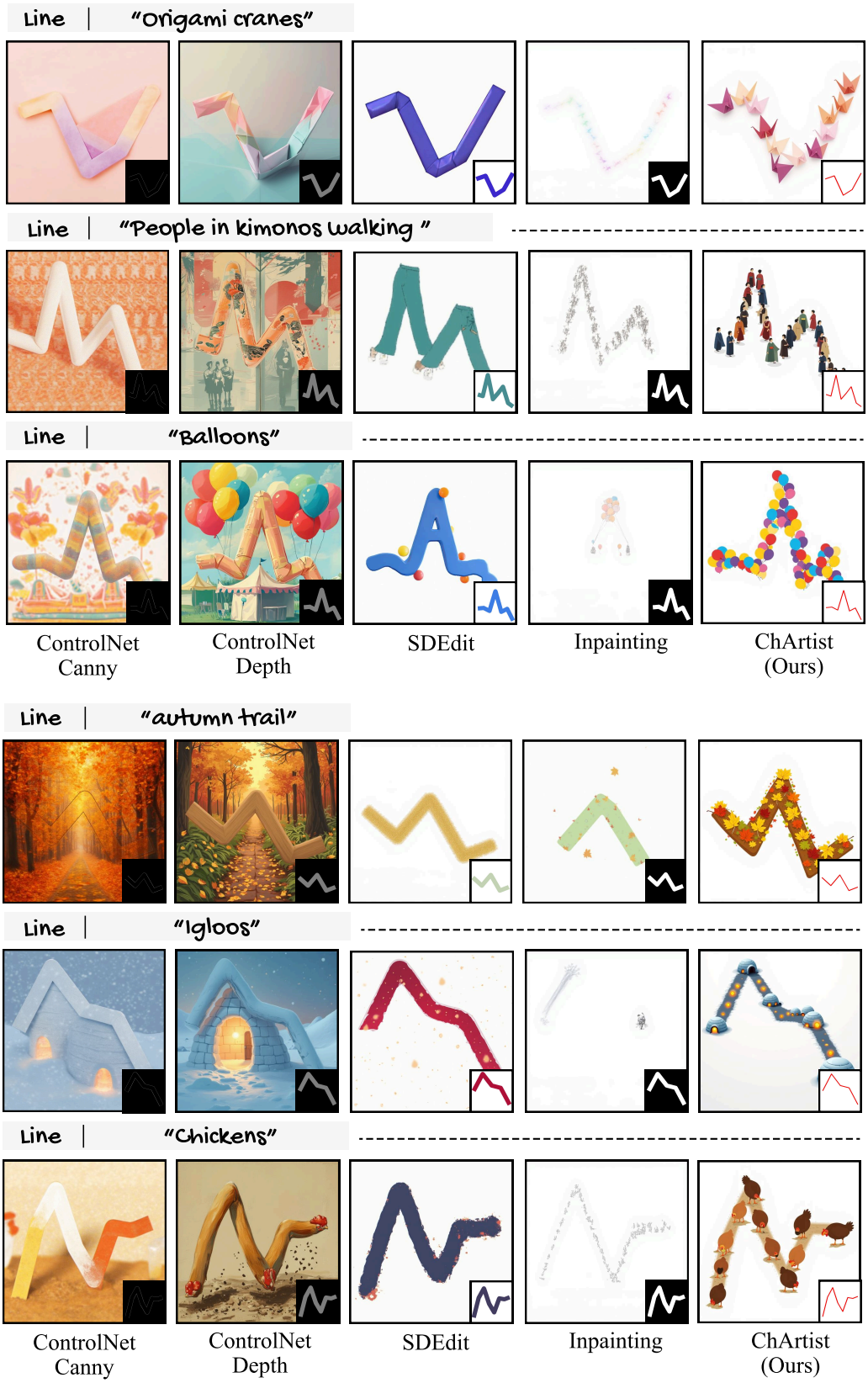


Figure S4. Result of spatially aligned evaluation with different control representations. (Task 1).

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.

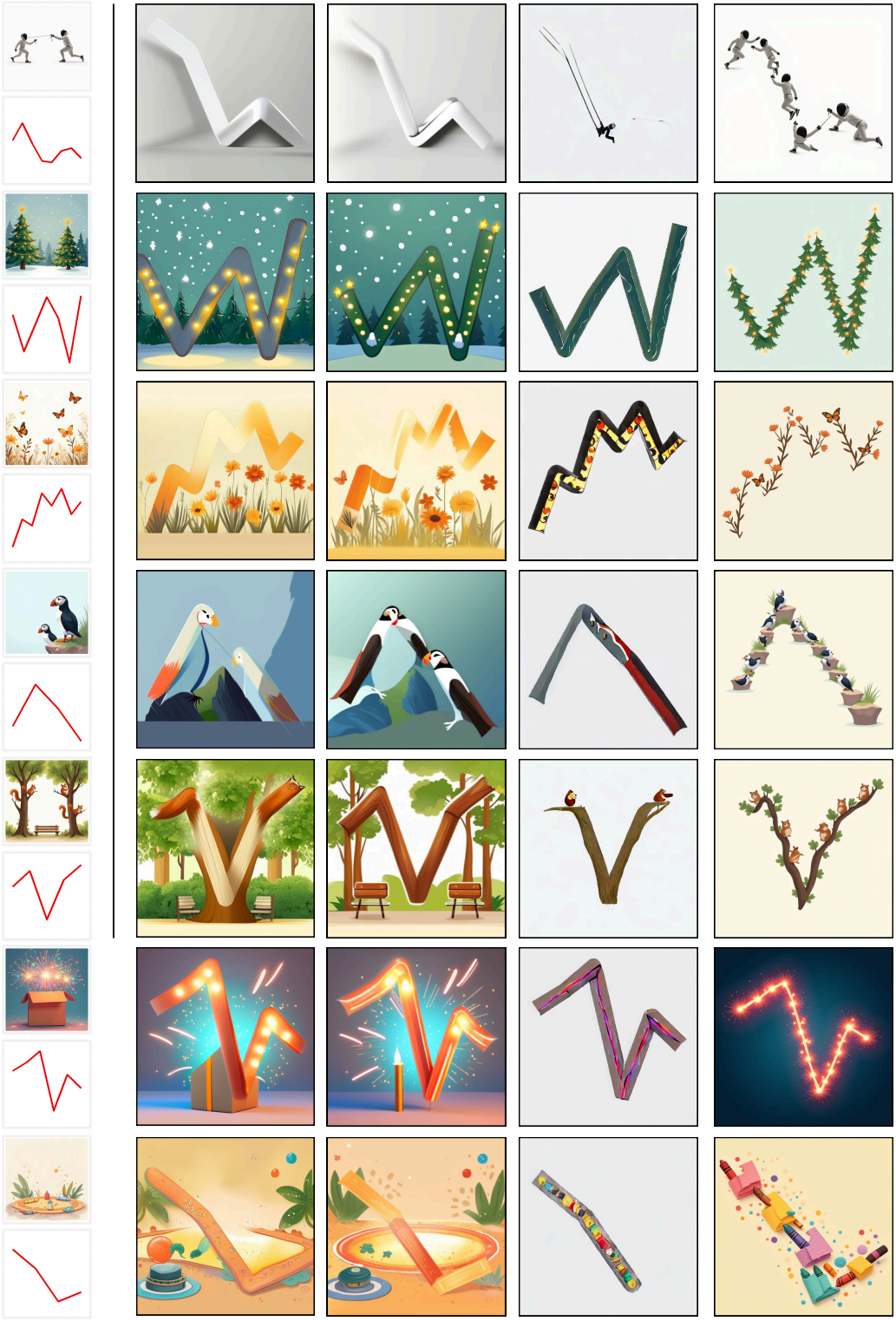


Figure S7. Dual-control generation line pictorial results conditioned on both spatial structure and subject reference.