Supplementary Materials

Our supplementary material is organized as follows:

1. We detail our evaluation metrics, including the proposed Qwen Semantic Score (Sec. A).
2. We present additional implementation details for our experiments (Sec. B).
3. We show the prompts used during our dataset construction and their corresponding results (Sec. C).
4. We showcase example comparisons between our dataset and other existing datasets (Sec. D).
5. We provide expanded comparisons of our method against additional competing methods and on more unseen styles (Sec. F).
6. We show more generation results from our method (Sec. G).

A. Detailed Evaluation Metrics

The Qwen Semantic Score is proposed to evaluate whether a stylization process utilizes semantic information from the style reference image. To assess this, we feed the style reference image and the stylized output image into a VLM, tasking it with determining if semantic information from the style image was incorporated. The specific prompt used for this VLM-based evaluation is detailed in Listing 1.

As shown in the prompt, `style_path` corresponds to the style reference image and `output_path` corresponds to the stylized output image. We use Qwen3-VL-8B-Instruct [3] as the VLM. If the model's response begins with 'YES', the score is 1; otherwise, it is 0.

In practice, for the *Stylized Dataset Curation* section, we evaluate the semantic focus of each style by randomly sampling five content-style-stylized triplets. We compute the Qwen Semantic Score for each triplet and sum them to produce a final semantic focus score for that style, ranging from 0 to 5. Conversely, in the *Quantitative Comparisons* section, we evaluate a model's ability to transfer semantic information. We compute the Qwen Semantic Score for every stylized triplet in the test set and then average these scores. The resulting metric is a value between 0 and 1.

Additionally, the reported CLIP score averages the similarities to both the content and style references, preventing models from merely copying the content image.

B. Experiments Details

B.1. IoU Experiment Details

In the IoU experiment presented in the *Discussion* section, we calculate the IoU of the expert indices selected for similar and dissimilar styles. Specifically, we ensure all reference images share the same content. First, we randomly select an anchor image I from our test set. We then define the candidate pool D as the set of all other test images that share the same content as I. We compute the CLIP similarity between I and all $I_j \in D$, then designate the similar style I_s and dissimilar style I_d as:

$$I_s = \arg\max_{I_j \in D} \text{CLIPSim}(I, I_j), \tag{11}$$

$$I_d = \arg\min_{I_j \in D} \text{CLIPSim}(I, I_j). \tag{12}$$

For each MoE-enabled layer l, let $E_l(I)$ be the set of expert indices selected by the gating network G_l for image I. We then compute the IoU for the similar pair (I, I_s) and the dissimilar pair (I, I_d) at this layer:

$$\text{IoU}_{\text{sim},l} = \frac{|E_l(I) \cap E_l(I_s)|}{|E_l(I) \cup E_l(I_s)|}, \tag{13}$$

$$\text{IoU}_{\text{dissim},l} = \frac{|E_l(I) \cap E_l(I_d)|}{|E_l(I) \cup E_l(I_d)|}. \tag{14}$$

We repeat this process for 100 sampled triplets (I, I_s, I_d) and compute the average IoU for each layer. In the table, "Early Stage," "Mid Stage," and "Late Stage" correspond to the first 1/3, middle 1/3, and final 1/3 of the MoE-injected layers, respectively.

B.2. Evaluation Prompts

Some stylization methods require text prompts to function. Methods like OmniStyle [44] and CSGO [52] require a *content prompt* to describe the subject. Others, such as DreamO [30], OmniGen2 [47], and Qwen-Image-Edit [46], treat stylization as a sub-task and require a *task prompt* to activate it. For the first category, we supply a clean content prompt during evaluation to ensure optimal performance. For the second, we identified a robust, general-purpose task prompt through testing: `"Make the whole first image have the same image style as the second image."` We use this prompt in all evaluations for these methods.

C. Dataset Curation Details

To generate a dataset with high style consistency, we must obtain clean prompts that are free of any style descriptors, as these descriptors could adversely affect the stylization process. To this end, we first generate an initial caption by feeding the content image and the prompt from Listing 2 into Qwen3-VL [3]. Although Listing 2 explicitly requests a content-only description, the resulting captions often retain style information. For example, as shown in Fig. 9, an initial caption might be: "A young child wearing a white

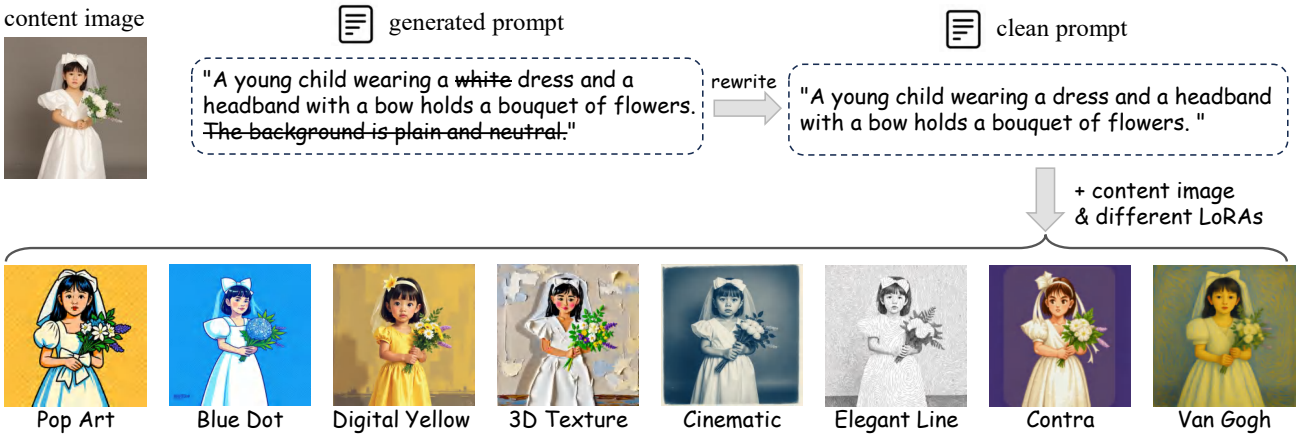

Figure 9. Illustration of our prompt filtering process. Initial VL-generated captions often retain descriptors (*e.g.*, colors, atmosphere) that conflict with styles defined by their own intrinsic properties. This is detrimental to styles with unique color palettes (*e.g.*, blue dot, Digital Yellow, Van Gogh, Elegant Line) or distinct atmospheres (*e.g.*, Pop Art, 3D Texture, Cinematic, Contra). Our rewriting step removes these conflicting descriptors to produce clean prompts, ensuring consistent stylization.

dress... The background is plain and neutral." These descriptors are detrimental because many artistic styles define their own intrinsic color palettes or background compositions. A prompt specifying "white" or a "plain" background would conflict with such styles and degrade the stylization quality. Therefore, we perform an additional rewriting step. We use Qwen3 [3] and the prompt from Listing 3 to remove this specific style information, yielding a final clean caption such as: "A young child wearing a dress and a headband with a bow holds a bouquet of flowers."

To filter out triplets where the layout has changed significantly, we primarily check for variations in composition and specific attributes, such as the number or gender of persons. The prompts used for this filtering process are provided in Listing 4 and Listing 5.

D. Dataset Comparison

To further illustrate the differences between our dataset and previous datasets, Fig. 10 compares our StyleExpert-500K dataset with the previous OmniStyle-150K dataset. Our dataset exhibits significant stylistic diversity, with styles categorized by their depth of semantic focus, spanning Color, Line, Texture, and Semantic levels. In contrast, the majority of examples in OmniStyle-150K degenerate to simple color transfer, failing to utilize deeper semantic information from the style reference. For instance, they fail to capture the "illustration style" (Row 1, Col 1) or the "symbolic style" (Row 1, Col 2). Furthermore, because the reference colors are not always suitable for the content image, these stylizations often exhibit poor aesthetic quality.

Beyond qualitative observations, Tab. 4 provides a quantitative comparison between the two datasets. Unlike OmniStyle-150K, our StyleExpert-500K successfully maintains color-semantic balance. Furthermore, it outperforms the baseline across most metrics, particularly achieving a significantly higher CSD score and a lower DreamSim distance. While performing comparably on DINO, our dataset yields better CLIP similarity and Aesthetic scores, quantitatively demonstrating its superior stylistic diversity and overall visual quality.

Table 4. Quantitative comparison between OmniStyle-150K and our StyleExpert-500K.

Dataset	Color-Semantic Balance	CLIP (↑)	DINO (↑)	CSD (↑)	Aesthetic (↑)	DreamSim (↓)
OmniStyle-150K	✗	67.68	**67.71**	38.47	6.08	59.49
StyleExpert-500K	✓	**68.41**	66.75	**74.49**	**6.56**	**36.19**

E. User Study

To subjectively assess our method, we conducted a user study involving 30 participants and 1,200 total votes. Users were asked to select the best stylized output from randomized baseline results, evaluating three main aspects: style consistency, content preservation, and overall aesthetics. The results, presented in Tab. 5, demonstrate that our StyleExpert dominates human preference, achieving a Top-1 selection rate of 74.5%, outperforming all other methods.

Table 5. User study results. Values indicate the Top-1 preference rate (%) across all methods.

	CSGO	DreamO	OmniGen2	OmniStyle	QwenEdit	USO	**StyleExpert**
Top1	3.0	1.1	8.5	4.8	3.6	4.5	**74.5**

F. Additional Comparisons

Figures 11 and 12 present expanded qualitative comparisons against a broader range of stylization methods. To provide a more comprehensive analysis, we extend

our comparison to include Qwen-Image-Edit [46], Nano-Banana [12], and ChatGPT [1], in addition to the baselines discussed in the main text (OmniStyle [44], CSGO [52], USO [48], OmniGen2 [47], and DreamO [30]). As observed, our method consistently achieves the best overall stylization performance. In contrast, Qwen-Image-Edit frequently fails to perform stylization, often merely replicating the content or style reference image, or generating irrelevant content (*e.g.*, Fig. 11, top-right and Fig. 12, middle-right). Similarly, Nano-Banana often fails to capture deep semantic cues from the style reference, resulting in a mere replication of the content image. While ChatGPT achieves relatively better stylization, it still suffers from insufficient stylization intensity, yielding suboptimal results (*e.g.*, Fig. 11, top-left and middle-right).

G. Additional Visual Results

Figures 13, 14, 15, and 16 present additional generation results from our method across a diverse range of content and style inputs. As observed, our method excels not only at traditional color transfer (*e.g.*, Fig. 13, 1st style column) but also at capturing the overall atmosphere (*e.g.*, Fig. 13, 4th style column; Fig. 14, 3rd style column) and transferring intricate textures and lines (*e.g.*, Fig. 13, 3rd style column; Fig. 14, 1st style column). These results fully demonstrate the robustness and reliability of our approach in effectively transferring style attributes across multiple semantic levels.

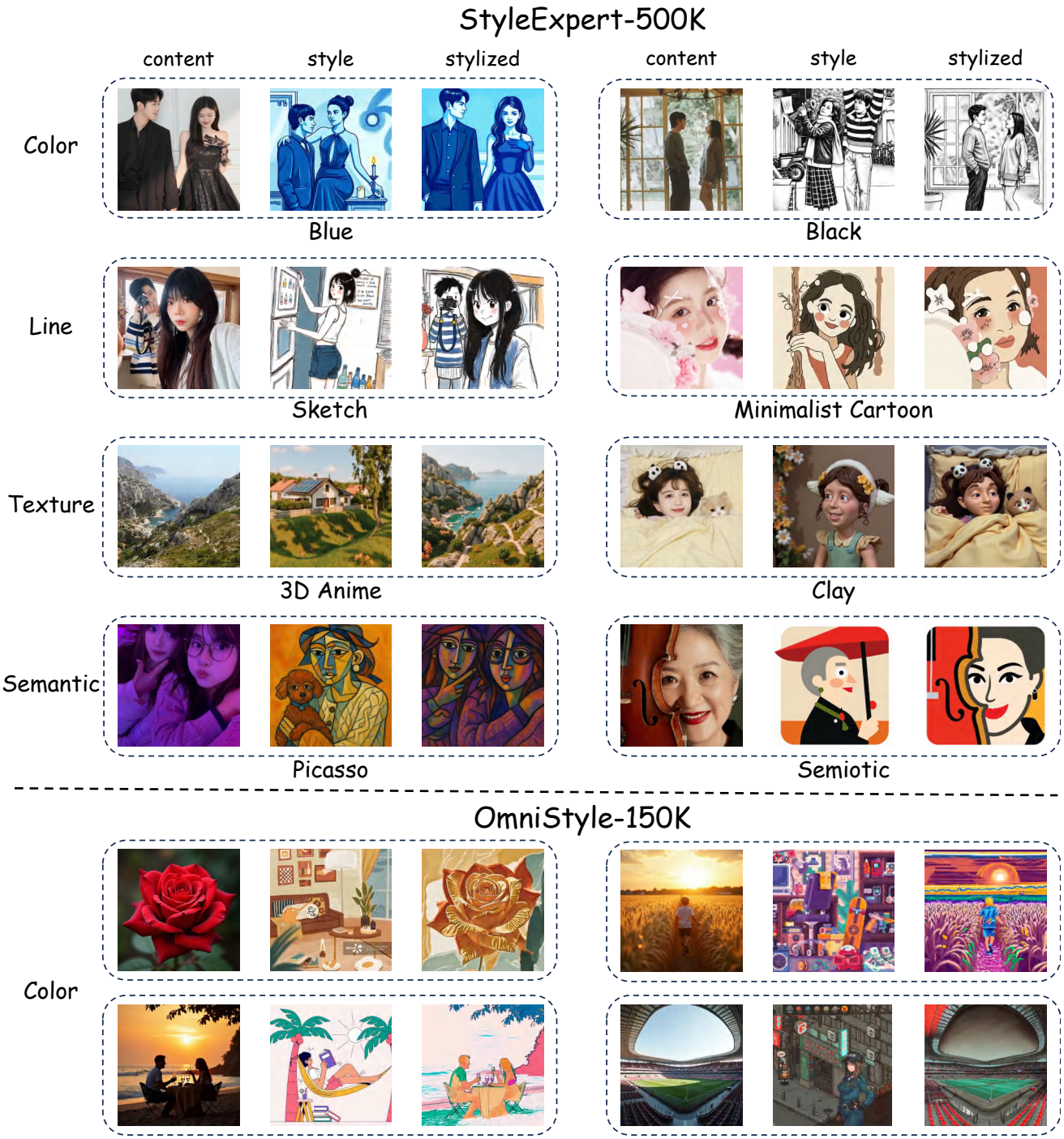

Figure 10. Comparison of data samples from our StyleExpert-500K dataset and the OmniStyle-150K dataset.

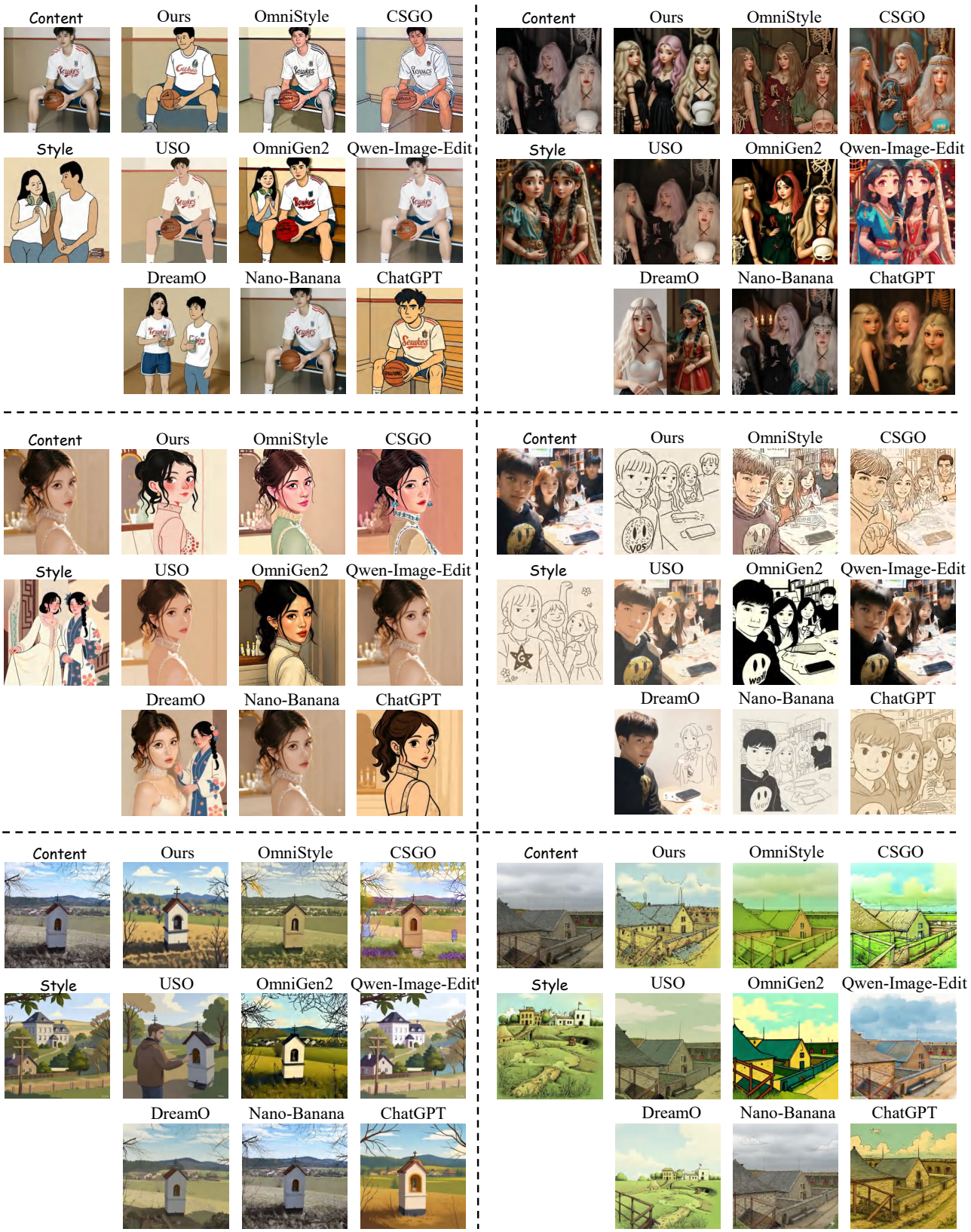

Figure 11. Qualitative comparison of our method against competing approaches under various content and unseen style inputs.

Figure 12. Qualitative comparison of our method against competing approaches under various content and unseen style inputs.

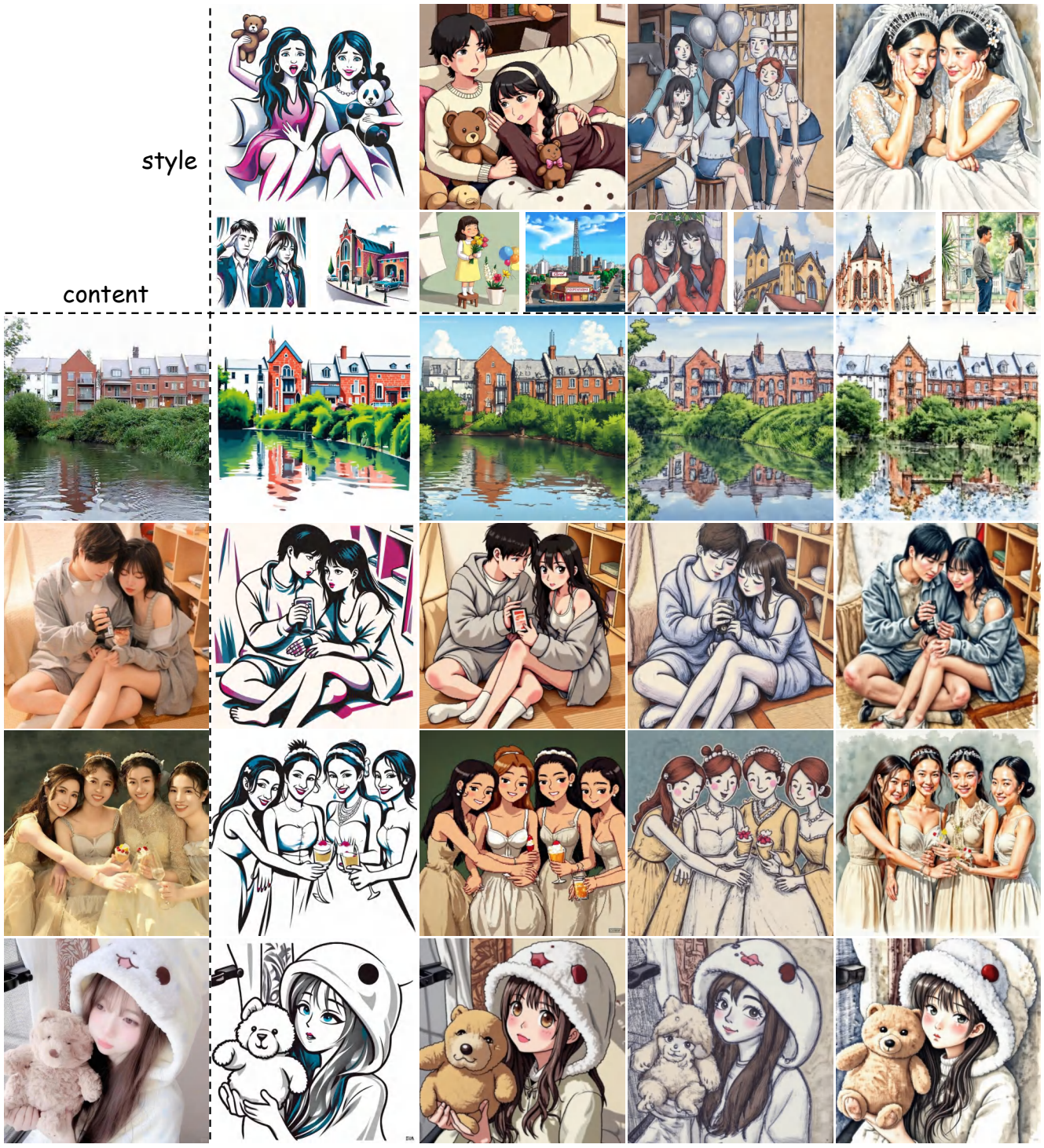

Figure 13. Additional visualization results generated by our method across a diverse range of styles and content subjects.

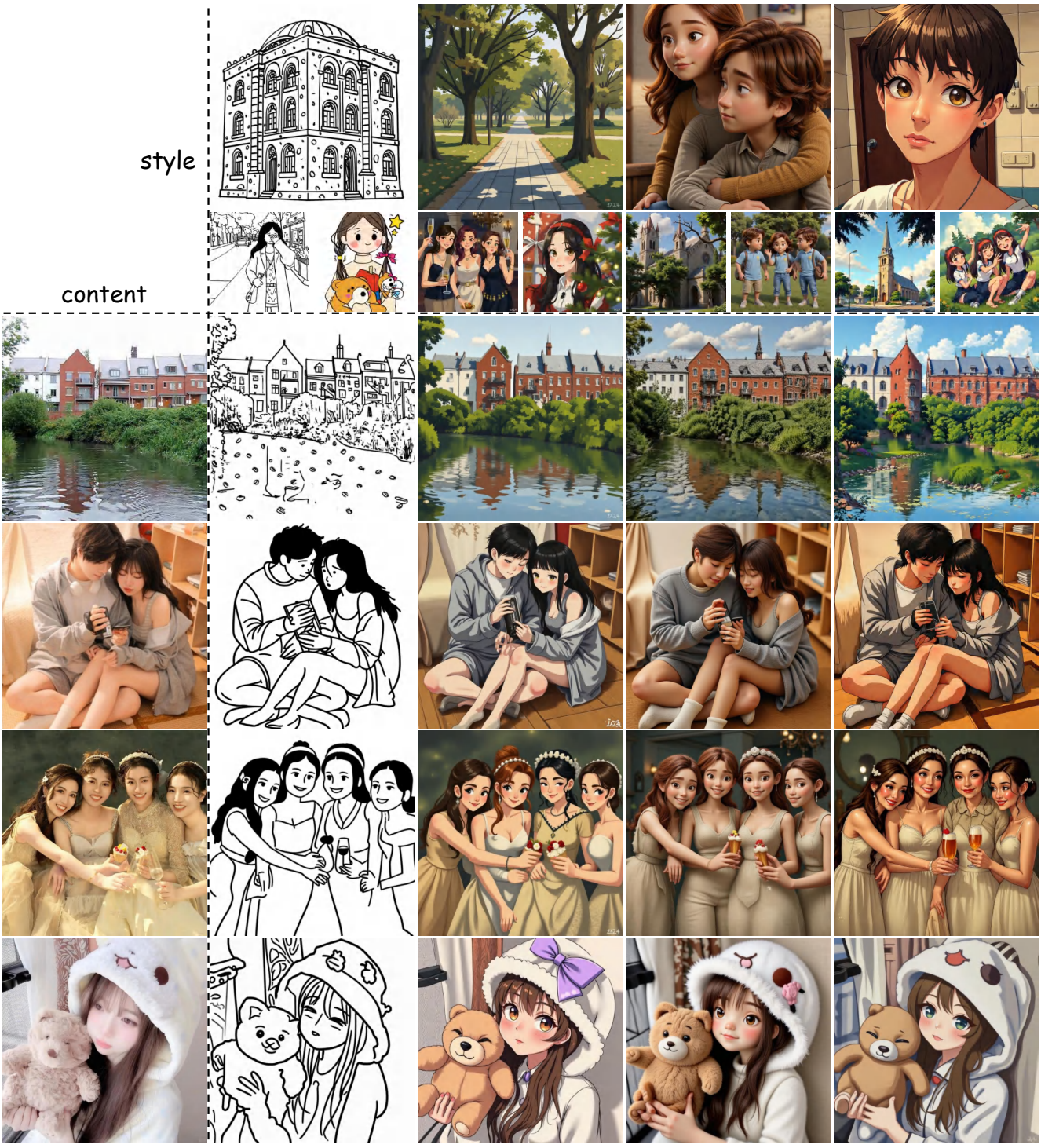

Figure 14. Additional visualization results generated by our method across a diverse range of styles and content subjects.

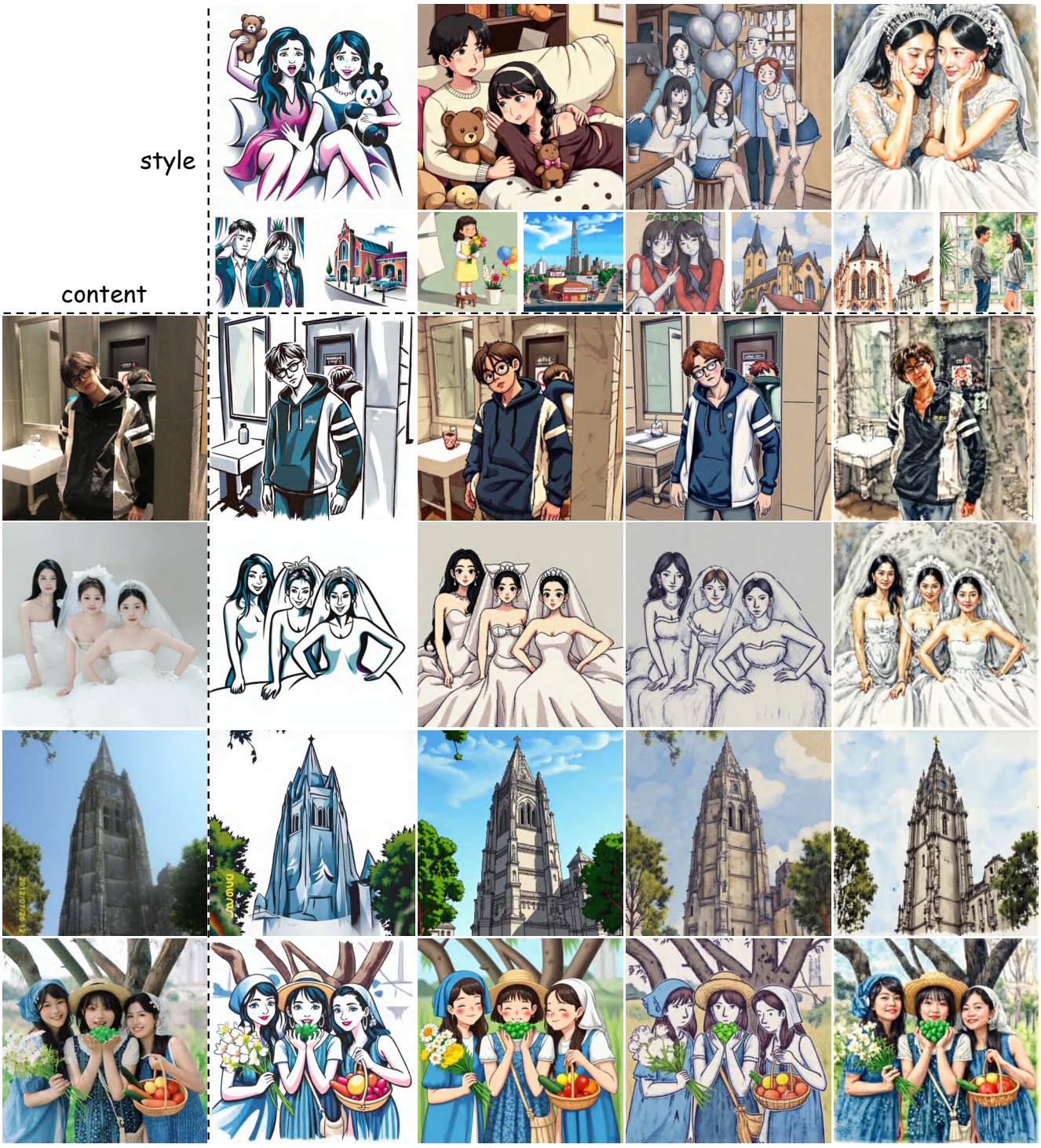

Figure 15. Additional visualization results generated by our method across a diverse range of styles and content subjects.

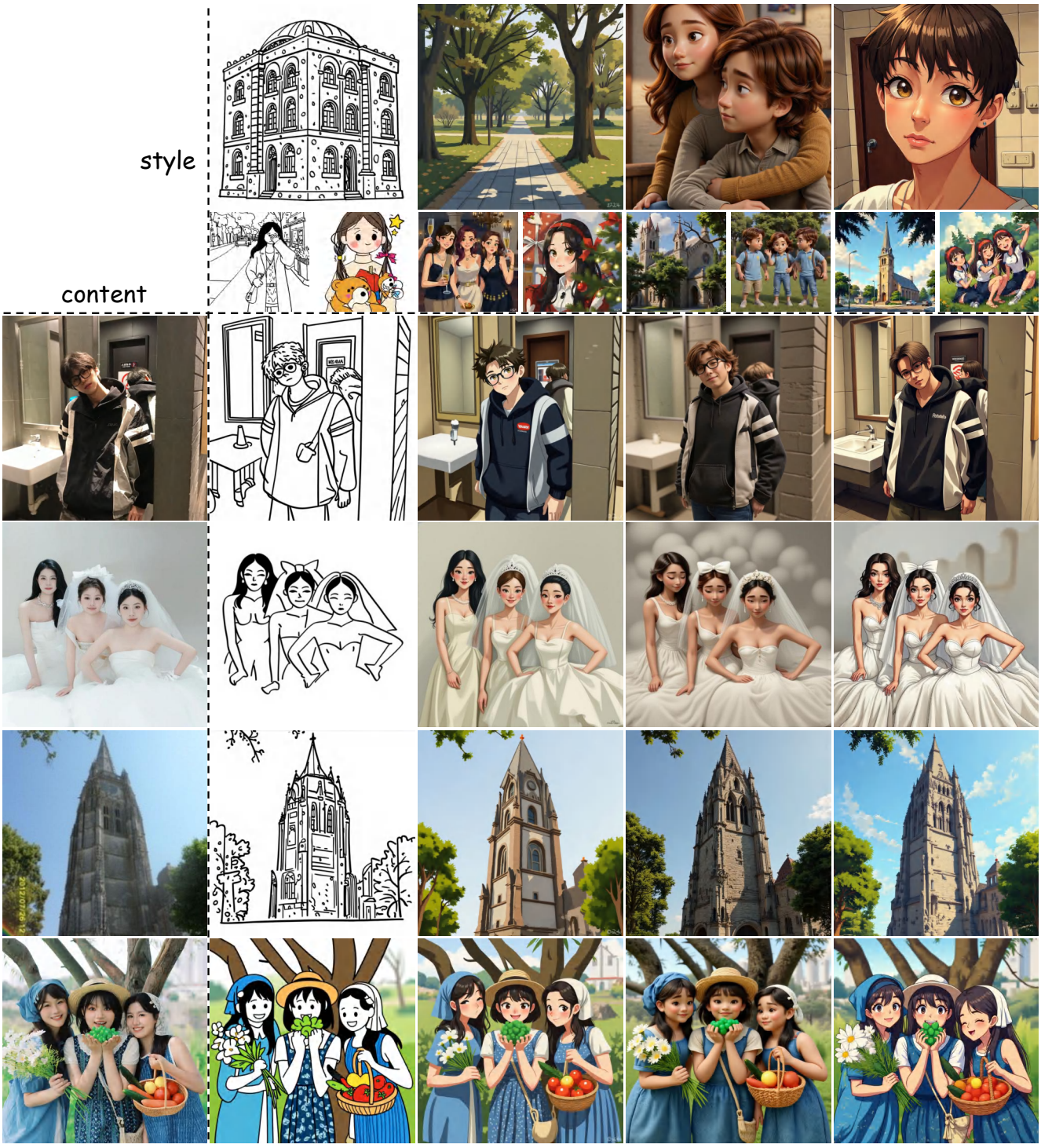

Figure 16. Additional visualization results generated by our method across a diverse range of styles and content subjects.

Listing 1. The prompt for the Qwen Semantic Score. This prompt instructs the VLM to evaluate whether the stylized output (Style image B, mapped to `output_path`) shares the same semantic style (*e.g.*, texture, line quality) as the style reference (Style image A, mapped to `style_path`), while ignoring simple color similarities.

```
{
   "role": "user",
   "content": [
      {"type": "text", "text": "You are a visual style analysis expert."},
      {"type": "text", "text": (
          "You will be given two images:\n"
          "1. Style image A\n"
          "2. Style image B\n\n"
          "Your task is to determine whether these two images share the *same artistic
              style*.\n"
          "By 'same style', we refer to having similar **texture**, **line quality**, and
              **material or rendering characteristics**, "
          "not merely similar colors, lighting, or atmosphere.\n\n"
          "Answer only 'YES' if both images have the same style, otherwise 'NO'. "
          "Then briefly explain why."
      )},
      {"type": "text", "text": "Style image A:"},
      {"type": "image", "image": style_path},
      {"type": "text", "text": "Style image B:"},
      {"type": "image", "image": output_path},
   ],
}
```

Listing 2. The prompt used to generate initial, content-only captions. This prompt instructs the VLM to describe only the main objects and their spatial relationships, while explicitly forbidding all stylistic, color, or lighting descriptors.

```
{
   "role": "user",
   "content": [
      {"type": "image", "image": image_path},
      {"type": "text", "text": (
         f"Please generate a caption for this image, "
         f"DO NOT mention style, color, texture, material, lighting, atmosphere, "
         f"or any words like 'red', 'blue', 'green', 'shiny', 'dark', 'bright', "
         f"'realistic', 'light-colored', etc. "
         f"ONLY include main objects and their spatial relationships. "
         f"No more than 50 tokens."
      )},
   ],
}
```

Listing 3. The prompt for the caption rewriting step. This instructs Qwen3 to remove all style, color, and texture descriptors from an input caption (`caption`).

```
{"role": "system", "content": "You are an assistant that edits text."},
{
   "role": "user", "content": "
   Keep all factual and visual details about objects, people, scenes, and actions.
   Remove all references to style, color, texture, material, lighting, or atmosphere, such
       as:
   red, blue, green, shiny, dark, bright, realistic, light-colored, etc.
   Do not add new information.
   Output only the cleaned description.

   Based on the rules above, rewrite the following description in English:
   {caption}
   "
}
```

Listing 4. The prompt for the content-matching filter. This instructs the VLM to verify if the stylized image (`img_path`) strictly matches the clean content caption (`caption`).

```
{
{
   "role": "user",
   "content": [
      {"type": "text", "text": "Here is a stylized image:"},
      {"type": "image", "image": img_path},
      {"type": "text", "text": (
         f"Check if the content of this image strictly matches the caption: "
         f" '{caption}'."
         f"Ignore color, material, and things related to style in caption, "
         f"answer only 'YES' if it strictly matches, otherwise 'NO'. "
         f"Then briefly explain why."
      )},
   ],
}
```

Listing 5. The prompt for the person-consistency filter. This instructs the VLM to verify that the stylized image (`img_path`) and the source image (`src_path`) contain the same number of people and that their genders strictly match.

```
{
   "role": "user",
   "content": [
      {"type": "text", "text": "This is image 1:"},
      {"type": "image", "image": img_path},
      {"type": "text", "text": "This is image 2:"},
      {"type": "image", "image": src_path},
      {"type": "text", "text": (
         f"Check if the two images"
         f"1. Contains the same NUMBER of objects (\eg, people). Variations that reflect
             intentional youthful stylization should be recognized as the same person."
         f"2. GENDER OF EACH PERSON strictly matches"
         f"YOU MUST IGNORE ANY OTHER DETAIL CHANGE AND FOCUS ONLY ON NUMBER & EACH PERSON'S
             GENDER."
         f"Answer only 'YES' if the number & EACH PERSON's gender met, otherwise 'NO'."
         f"Then briefly explain why."
      )}
   ],
}
```