SIGMA: Selective-Interleaved Generation with Multi-Attribute Tokens

Supplementary Material

This appendix provides additional details and analyses for SIGMA. Section 7.1 presents an extended description of the interleaved dataset, including task coverage, attribute design, and the construction pipeline with special tokens and interleaving strategies. Section 7.2 further explains how generated data and existing corpora are combined into a unified training set. Section 8 introduces the evaluation protocols, including XVerseBench and a comprehensive benchmark covering compositional generation, selective generation, style and relation transfer, and layout-based generation. Section 9 describes the evaluation and training protocols, detailing how multi-condition inputs are handled for all baselines and providing additional implementation details. Section 10 reports the user study setup and analyzes human preferences across different tasks, showing that SIGMA is consistently favored over strong baselines. Section 11 presents representative failure cases and discusses the conditions under which SIGMA may struggle. Section 12 provides additional qualitative results across various tasks. Finally, Section 13 provides additional qualitative results and visual comparisons across various tasks.

7. Dataset

7.1. Dataset Overview

To make SIGMA actually learn which attribute should be taken from which reference, we curate a large-scale, text–image interleaved dataset with explicit attribute annotations. The final corpus contains about **700K** interleaved sequences, each mixing text spans, special tokens, and one or more condition images. The data covers six major task families: *compositional generation* (100K), *selective extraction* (226K), *style transfer* (153K), *relation transfer* (41.6K), *image editing* (70K), and *conditional layout generation* (110K). This diversity is important: SIGMA is expected to see not only "one prompt + one image" cases, but also multi-image, cross-attribute, and asymmetric-reference cases at training time.

Pure vision–text pairs are not enough for that goal. While vision–text paired data provides useful supervision, it falls short when the model must track multiple visual sources, resolve which text span refers to which image, or propagate an intermediate relation (e.g., "make C look like A stylized as B"). Models trained on such data tend to blur visual and semantic correspondences across modalities, producing less coherent multi-source generations. Our dataset therefore switches the supervision unit from "caption $\leftrightarrow$ image" to "interleaved sequence $\leftrightarrow$ image": every sequence explicitly tells the model when a new visual source appears and what semantic role it plays.

To make these roles machine-readable, we introduce 14 special tokens that indicate specific visual entities or attributes (identity, subject, clothing, style, texture, lighting, emotion, pose, background, environment, perspective, layout, text, and select), together with a small functional token for explicit extraction of designated objects. Instead of describing "a red bag next to the person" only in free-form text, we insert the corresponding token right before the textual mention and then place the referenced image in the sequence. This tokenization does two things at once: it anchors the text span to the correct image group, and it disambiguates which attribute we want to pull from that image. In compositional generation, where several reference images may contain overlapping content or similar colors, this explicit anchoring is crucial. Otherwise, the model easily drifts and copies the wrong source.

Another key design choice is that the dataset is intentionally attribute dense. Many samples contain multiple attribute tokens within the same image, for example `<subject>` `<clothing>` and `<background>` appearing together. This reflects the realistic setting of interleaved generation because when users provide several reference images, the ambiguity of what should be transferred increases very quickly. Training on such densely annotated and multi attribute cases helps SIGMA learn to extract only the factors that are explicitly requested by the instruction and to take them from the correct reference image rather than mixing all visual cues together. Together with the scale of the dataset, which contains 700K samples, and the diversity of included tasks, this design enables SIGMA to maintain strong controllability at inference time even when the input conditions are heterogeneous and user specified.

7.2. Dataset Construction

Data Sources and Generation. Our interleaved dataset combines both newly generated and adapted sources to ensure broad coverage of compositional and conditional generation scenarios. We create large-scale synthetic data for compositional, stylization, and editing tasks, and further incorporate high-quality existing corpora augmented with explicit image–token bindings.

We generate about 220K compositional samples using GPT-4o [21] and Nano-Banana [7], covering diverse human–object–scene combinations such as indoor/outdoor environments, product displays, and clothing replacement. A 30K stylization subset and a 41.6K style-relation transfer subset capture style and relation mappings; the latter trains the model to apply a relation transformation $(A \rightarrow B)$

to a new content image C, with style references collected from PromptsRef and de-styled variants produced by Gemini 2.5 Pro. We additionally include 20K image-editing samples generated with Nano-Banana for customized content replacement and object insertion, verified for visual coherence. For selective-content extraction, we adapt Echo-4o [63] by treating each compositional output as input and identifying target objects via GPT-4o analysis, producing prompts such as "Extract the {object} from the image." The conditional-layout subset includes two types of samples: those using explicit layout inputs such as canny or depth maps, and those where one image from a compositional sequence is designated as the layout condition. Depth and structural cues are extracted using MiDaS [43], ensuring consistent geometric control across diverse scenes.

Existing datasets including Nano-150K [63], Echo-4o [63], X2Edit [32], and ShareGPT-4o [6] are also integrated after token-based alignment, forming a unified interleaved corpus for multi-condition learning.

Interleaving and Token Injection Pipeline To unify data from diverse sources under the interleaved paradigm, we design a structured token-injection pipeline that transforms each original text–image pair into an interleaved multimodal sequence. Given a caption describing multiple entities or attributes, we first parse the sentence to locate phrases referring to concrete visual elements (e.g., "the man," "the red car," "the city background," "the jacket"). Each identified phrase is prepended with a corresponding `<special_token>`, such as `<id>`, `<subject>`, `<clothing>`, or `<background>`, which specifies the intended attribute. The referenced image is then inserted immediately after the phrase, forming a locally bound text–image pair. Formally, a standard caption such as "Make the man wear the blue jacket and stand beside the car." is converted into "Make `<id>`the man in <portrait image> wear `<clothing>`the blue jacket in <jacket image> and stand beside `<subject>`the car in <car image>" This procedure yields an interleaved sequence that the diffusion transformer can process directly, with each token–image group representing one condition scope.

For existing datasets without explicit entity information, we employ a large-language-model parser to infer likely correspondences between nouns and reference images. Each inferred entity is assigned an appropriate token from the 14-token vocabulary, ensuring consistent semantics across heterogeneous data sources. We further enforce quality control by discarding cases with ambiguous references or unresolved entities.

This standardized interleaving pipeline transforms traditional vision–language datasets into structured multimodal supervision that explicitly encodes cross-condition relationships. By training on such sequences, SIGMA learns to associate token semantics with their corresponding visual sources, enabling robust multi-attribute reasoning and fine-grained controllability during generation.

8. Benchmarks

We evaluate our method and baselines and cover different aspects of controllable image generation. XVerseBench [3] is used for compositional generation, comprising 20 distinct human identities, 74 unique objects, and 45 different animal species or individuals. The compositional generation split contains 210 cases, each involving 1–3 input combinations, allowing us to evaluate the model's ability to handle multi-entity reasoning.

We further introduce a new **comprehensive benchmark** that jointly evaluates controllability, compositional reasoning, and structural alignment across four representative tasks: *compositional generation*, *selective generation*, *style transfer*, and *layout-based generation*. All benchmark samples are constructed from the held-out portion of our corpus that is never used for training, ensuring a strict separation between training and evaluation data. Specifically, the compositional subset contains 100 samples with 2 to 6 inputs that mix humans, objects, and scenes. The selective subset includes 100 samples that focus on extracting or generating specific subjects. The style transfer task contains 60 samples, and the relation transfer task also contains 60 samples, each designed to evaluate a different aspect of style/relation control. The layout based subset includes 70 samples that are guided by edge based conditions, where 17 use a single canny map and 53 combine canny maps with additional reference images.

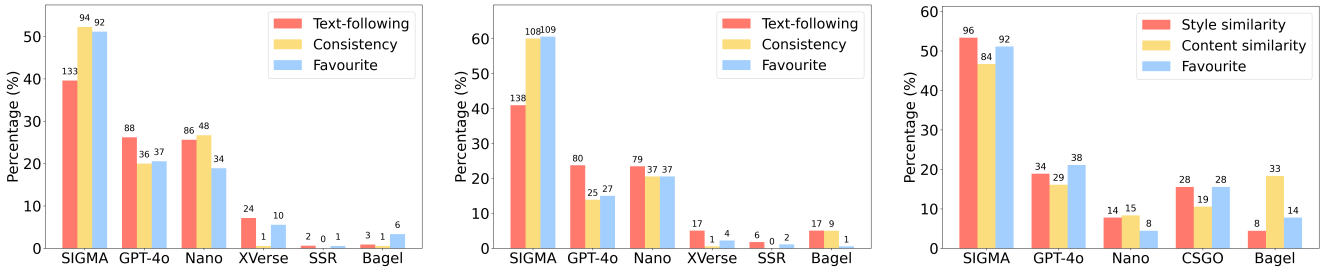
Together, XVerseBench and our comprehensive benchmark provide complementary perspectives. The former serves as a public and standardized evaluation protocol, while the latter delivers a broad and fine-grained assessment of controllability across diverse multi-condition settings. These two benchmarks ensure that the evaluation captures both external benchmark performance and generalization across rich, realistic, and systematically varied generation scenarios.

9. Evaluation and Training Protocols

9.1. Evaluation Protocol for Baselines

To ensure a fair comparison across all methods, we provide each baseline with multi-condition inputs in the format natively supported by that model, without altering its original conditioning mechanism.

For both SIGMA and the Bagel [8], inputs follow the text–image interleaving format used during post-training. Each reference image is inserted as an image token block, and SIGMA additionally prepends the corresponding special attribute token to explicitly specify the role of each ref-



(a) Compositional generation (b) Selective generation (c) Style transfer

Figure 7. User study preference across the three tasks.

erence. Bagel receives the same interleaved sequence but without attribute tokens, resulting in a purely text–image alternating structure. For SSR [66] and XVerse [3], each reference image is provided together with an associated subject identifier, following the native conditioning design of these models. For GPT-4o [21] and Nano [7], only standard text and image inputs are supported. Since neither model accepts interleaved multimodal sequences or explicit attribute tokens, we provide the full textual instruction followed by all reference images in a flat, unordered format. Generation is then performed through a single forward pass without any model-specific tuning or adaptation. This protocol evaluates each model strictly under its native interface, ensuring that differences in performance arise from controllability rather than input formatting constraints.

9.2. Extended Implementation Details

All images are processed at the native spatial resolution of the Bagel backbone. We employ the backbone's default diffusion sampler and scheduling strategy, and use classifier-free guidance with a fixed guidance scale throughout training and evaluation. During both training and inference, no test-time optimization, prompt rewriting, or auxiliary adapters are used. For all baselines, we directly follow their original sampling hyperparameters and native inference settings to avoid introducing differences unrelated to model capability.

9.3. CLIP-ES Metric

CLIP-ES measures the similarity between the generated attribute region and non-source reference images using CLIP embeddings. Specifically, we first crop the region corresponding to the evaluated attribute from the generated image. The CLIP embedding of this region is then compared with the embeddings of non-reference inputs. This metric evaluates whether the generated attribute is correctly aligned with its intended source and not mistakenly copied from other references.

10. User Study

To complement automatic metrics, we conduct a user study to assess the perceptual quality of SIGMA in compositional generation, selective generation, and style transfer.

10.1. User Study Setup

Task design. We construct a user study with a total of 15 groups of samples. Each of the three tasks contains 5 samples in total. For each sample, the images produced by different methods are anonymized and randomly shuffled before being shown to users. The compositional and selective generation tasks include 5 additional methods in total, namely, GPT-4o [21], Nano [7], XVerse [3], SSR [66], and Bagel [8], while the style-transfer task evaluates 4 additional methods, with CSGO [61] replacing XVerse and SSR. Each sample is associated with three multiple-choice questions, leading to 45 question blocks in total. All tasks share a common question that asks users to select their *overall favorite* image among all candidates. For compositional generation, the remaining two questions are: (i) "Select the result that follows the textual instruction," which evaluates text-following ability; and (ii) "Select the result where the objects are most visually consistent with the corresponding objects in the input images," which evaluates visual consistency and correct attribute binding. For selective generation, we use the same two task-specific questions as compositional generation to access the ability of text following and keeping consistency respectively. For style transfer, the two task-specific questions are: (i) "Select the result whose *content* is most similar to the content input image," which measures content preservation; and (ii) "Select the result whose *style* is most similar to the style input image," which measures style similarity. The text-following questions allow users to select more than one result when several outputs follow the instruction equally well. In all other cases, users provide a single choice per question.

Participants. We collected 36 valid responses in total after discarding incomplete questionnaires. The participants include both researchers and university students, covering

people with prior experience in computer vision and image generation as well as non-expert users. Each participant answered all 15 samples, resulting in $36 \times 15 \times 3$ individual choices overall.

10.2. Results and Analysis

To quantify human preference across different aspects of generation quality, we aggregate user votes by task, question type, and method. For each task, we sum all votes belonging to the same evaluation question (instruction following, consistency, or style/content similarity), and normalize the counts to obtain preference percentages for each competing method. Figure 7 visualizes the results for the three tasks. Within each subplot, bars of the same color correspond to the same evaluation aspect, the bar height indicates the normalized preference, and the number above each bar denotes the raw vote count.

As shown in Figure 7a and Figure 7b, SIGMA receives the highest user preference in both compositional generation and selective generation across all evaluation aspects. In particular, SIGMA achieves the largest proportion of instruction-following votes and a markedly higher proportion of consistency votes than all baselines, indicating that users perceive SIGMA as better at preserving object identity and maintaining correct attribute binding during generation. The consistently strong performance on these two tasks suggests that SIGMA more reliably extracts object-level cues from input references and integrates them into the final images in a faithful and coherent way. Figure 7c shows that SIGMA also performs strongly in the style-transfer task. SIGMA attains the highest preference in both style similarity and content similarity, demonstrating its ability to model complex style cues while avoiding conflicts between transferred style and preserved content. The strong user preference in the "favourite" question further suggests that SIGMA produces visually well-balanced results that users find both natural and aesthetically appealing.

11. Failure Cases

Figure 8 illustrates two typical failure cases encountered in our framework, covering compositional generation and selective extraction. The first example is a high-complexity compositional generation task. The prompt requires simultaneously binding multiple heterogeneous references, including a person, hat, dress, shoes, a reading stand, and a vehicle. When the number of input entities becomes large, the model occasionally struggles to maintain natural interactions across all referenced objects. This results in less coherent spatial arrangements as well as incomplete rendering of certain elements, such as the partially blurred rear of the car. These issues reflect the challenge of handling dense multi-entity reasoning when the reference set grows beyond the typical range seen during training. The second example shows a selective generation failure, where the model is asked to isolate and render a specific object from the reference image containing a complex, cluttered background. When the target object is small relative to its surroundings and embedded within highly textured environments, the attribute-specific token fails to fully suppress distractors, leading to geometric deformation or leakage of irrelevant background features. This indicates that selective extraction becomes more difficult when the foreground object is visually entangled with the background, particularly in cases where object boundaries are not sharply separable.

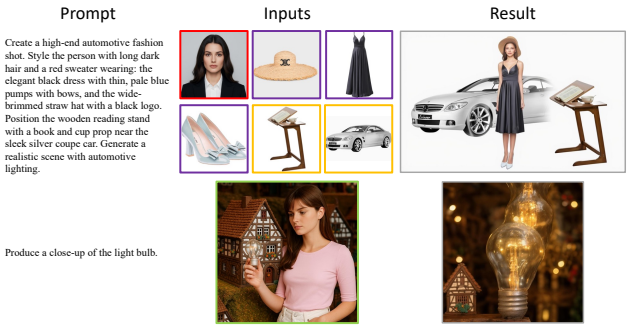


Figure 8. Two representative failure cases of SIGMA. Top: a compositional generation example with a large number of heterogeneous references. Bottom: a selective generation case where the reference image contains complex background structures.

These cases highlight limitations that arise mainly in extreme or visually challenging scenarios, either when the number of referenced entities exceeds the model's typical operating regime or when selective extraction is applied to cluttered scenes with non-salient targets. Improving robustness under these conditions remains an interesting direction for future works.

12. Additional Quantitative Evaluation

12.1. Style Transfer Evaluation

To quantitatively evaluate style and content preservation, we adopt CLIP-I similarity together with GPT-4o pairwise preference evaluation. CLIP-I measures the similarity between generated images and attribute-specific regions cropped from the corresponding reference images. Table 5 summarizes the quantitative results on style transfer using CLIP-I similarity and GPT-4o pairwise preference. From the CLIP-I scores, SIGMA achieves competitive performance on both style application and content preservation. More importantly, the GPT-4o preference results show that SIGMA is preferred over Bagel across both benchmarks, indicating that its outputs better preserve the intended style–content decomposition while maintaining overall visual quality.

Bench	Metric	CLIP-I Score GPT	Nano	CSGO	Bagel	SIGMA	GPT-4o Pref. (SIGMA vs.) CSGO	Nano	Bagel
Ours	Style	63.9	**68.4**	64.0	62.0	64.3	55.0	43.3	71.7
	Content	78.1	**81.3**	75.2	68.2	80.0			
XVerse	Style	**73.8**	69.4	64.1	54.1	72.8	46.3	41.3	71.3
	Content	74.6	81.6	80.1	**94.8**	74.3			

Table 5. Style/Content similarity (CLIP-I↑) & GPT-4o Preference.

Method	Layout Adherence (%)	Visual Quality (%)
SIGMA vs Bagel	65.38	54.85
SIGMA vs EasyControl	88.46	92.31
SIGMA vs Nano	61.54	69.23
SIGMA vs GPT-4o	69.23	46.15

Table 6. Human A/B preference results for layout-conditioned generation. Values indicate the percentage of cases where SIGMA is preferred over the competing method.

Comparison	Preference Rate (%)
SIGMA vs Bagel	87.5
SIGMA vs XVerse	96.15
SIGMA vs Nano	57.69
SIGMA vs CSGO	76.92

Table 7. Pairwise human preference results across multiple tasks. Values indicate the percentage of cases where SIGMA is preferred.

12.2. Pairwise A/B Preference Study

To further evaluate generation quality and attribute binding correctness, we conduct human A/B preference studies comparing SIGMA with several strong baselines. Participants are asked to select the result that better satisfies the given instruction.

Layout-only generation. Since automatic metrics such as FID may not always align with human perception of spatial alignment and visual realism, we conduct a dedicated A/B study for layout-conditioned generation and the results are shown in Table 6. Participants compare outputs produced by SIGMA and competing methods and select the image that better follows the given spatial layout.

Attribute Binding Evaluation. We further evaluate source correctness and selective attribute binding across four representative tasks: compositional generation, selective generation, style transfer, and layout-conditioned generation. For each pairwise comparison, participants choose the result that better follows the instruction and correctly preserves attribute sources. The results are shown in Table 7.

Overall, the results consistently favor SIGMA over competing approaches, indicating that the proposed token-based attribute binding and selective generation mechanism improves both controllability and perceptual generation quality.

12.3. Computational Costs

With 49 sampling steps, SIGMA has comparable inference cost to the Bagel backbone. Specifically, SIGMA requires 77.47s and 30.4GB GPU memory per generation, compared to 74.37s and 29.8GB for Bagel. The additional overhead mainly comes from the attribute-token processing and group-scoped attention masking. The post-training stage runs for 50K steps and takes approximately 120 hours on our training setup.

13. Additional Qualitative Evaluation

We further provide an extensive set of qualitative results to illustrate the generality and robustness of SIGMA across a wide range of generation scenarios. For compositional generation, Figure 9 shows additional comparisons against all baselines, where SIGMA consistently produces coherent combinations that preserve object identity and correctly bind attributes across references. In selective generation (Figure 10), SIGMA reliably extracts only the instructed regions or objects while maintaining the appearance and structure. Additional style-transfer results in Figure 11 show that SIGMA captures complex stylistic cues while preserving content, producing clean and visually stable outputs. We also include layout-only generation results in Figure 12, where SIGMA adheres faithfully to coarse spatial layouts while generating high-quality content, as well as layout + reference generation in Figure 13. Relation-transfer examples are provided in Figure 14, illustrating SIGMA's ability to reproduce relational configurations such as relative poses or style. Finally, Figure 15 presents editing examples, showing that SIGMA performs competitively on diverse edit instructions while preserving overall scene fidelity. These extended qualitative results highlight SIGMA's versatility, its strong attribute and relation binding, and its overall stability and fidelity across a broad spectrum of generation and editing tasks.

Figure 9. Additional compositional generation results. SIGMA preserves object identity, maintains attribute binding across references, and produces coherent global structures, while baselines often struggle with object fusion, spatial arrangement, or cross-object consistency.

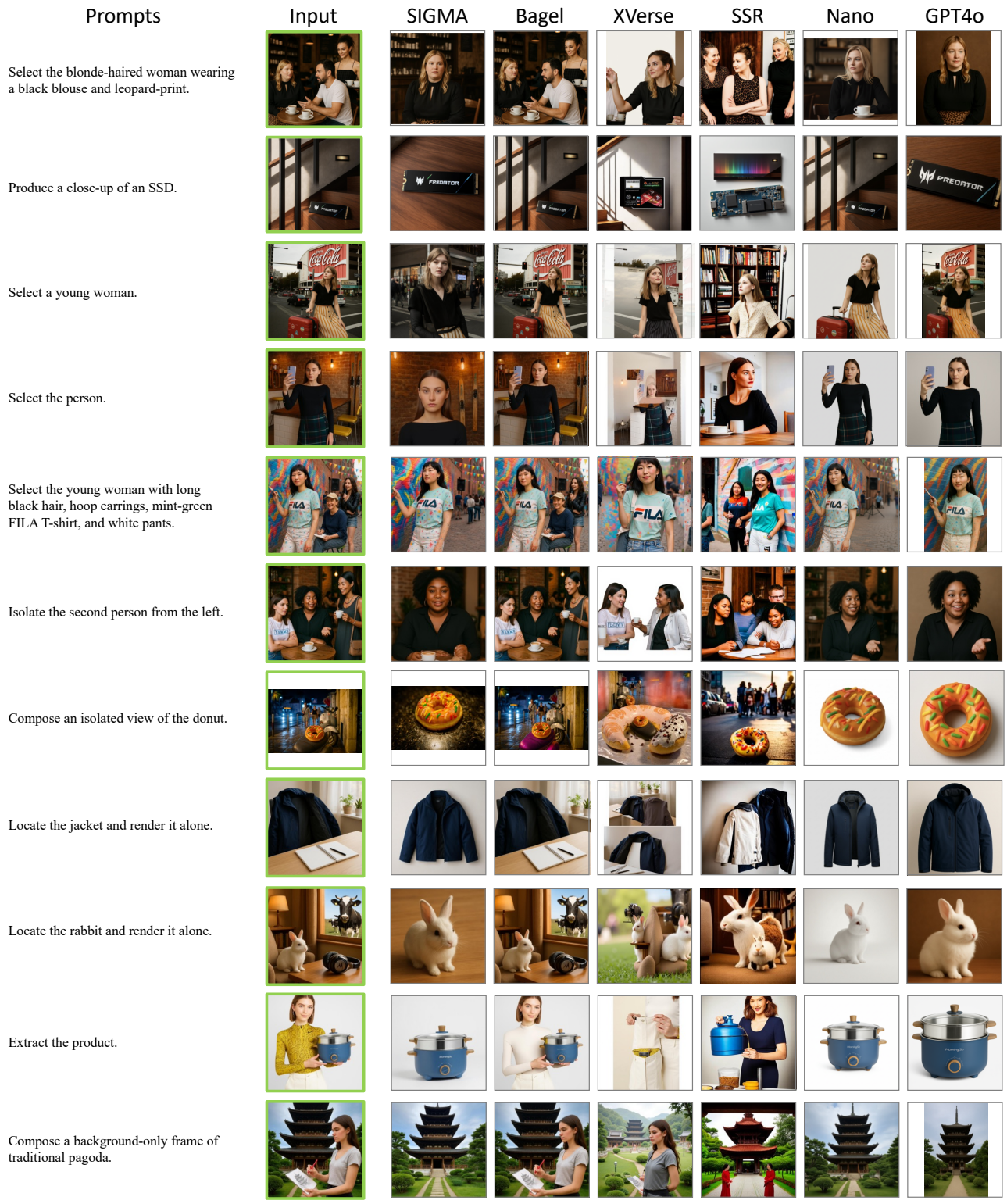


Figure 10. Additional selective generation examples. Given an instruction that targets only part of the scene, SIGMA reliably extract the specified region while keeping the appearance unchanged. Compared with the baselines, SIGMA achieves stronger localized control with fewer artifacts and better structural preservation.

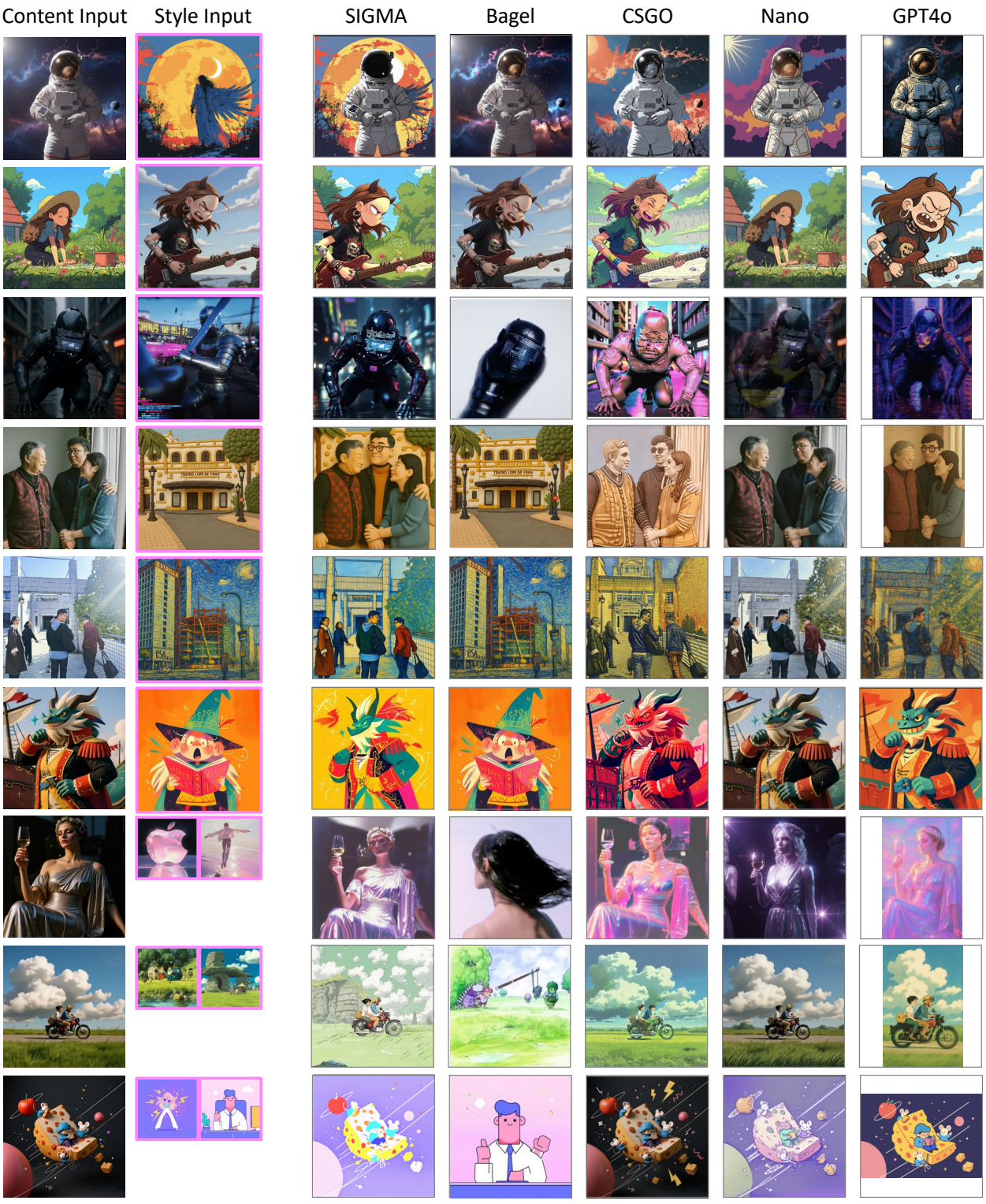


Figure 11. Additional style-transfer comparisons. SIGMA successfully transfers fine-grained style cues from the reference image onto the content image while maintaining structure and detail.

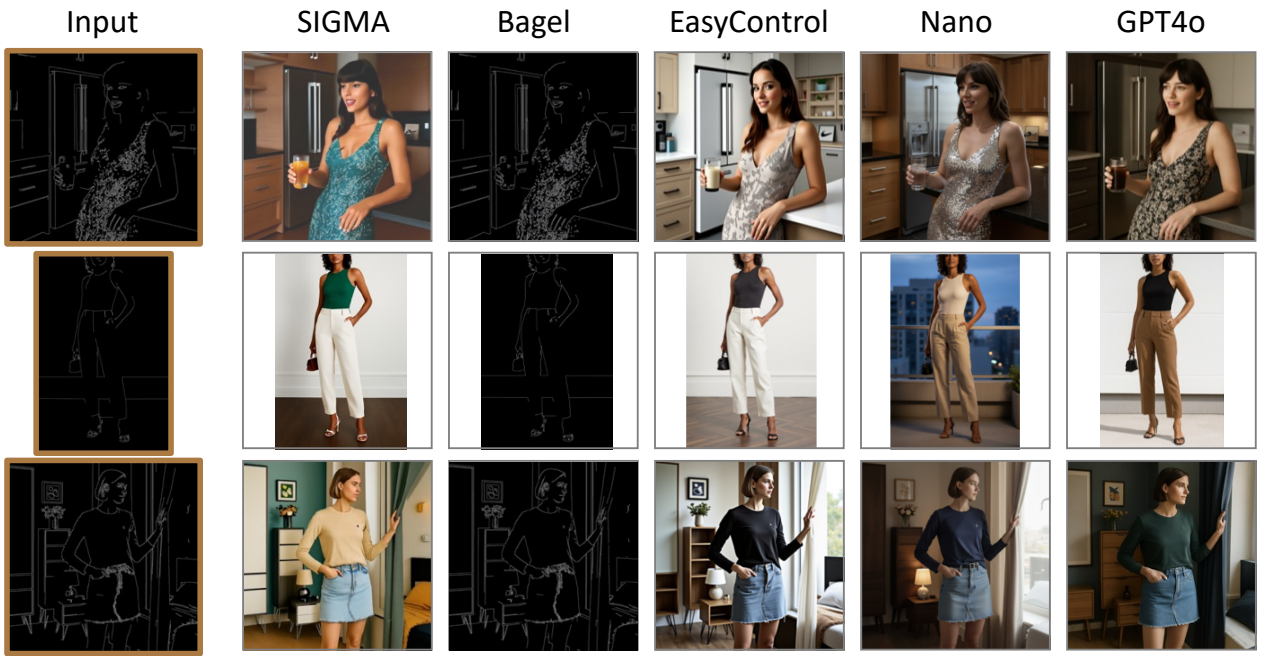


Figure 12. Additional layout-only generation results. Given only a spatial layout as conditioning, SIGMA adheres closely to the prescribed structure while generating high-quality and coherent content.

Figure 13. Additional layout + reference generation examples. SIGMA effectively combines layout constraints with appearance cues from the reference image, producing results that align with both spatial structure and visual identity. Baselines exhibit layout drift or fail to transfer reference-specific details, while SIGMA maintains precise structure and faithful attribute transfer.

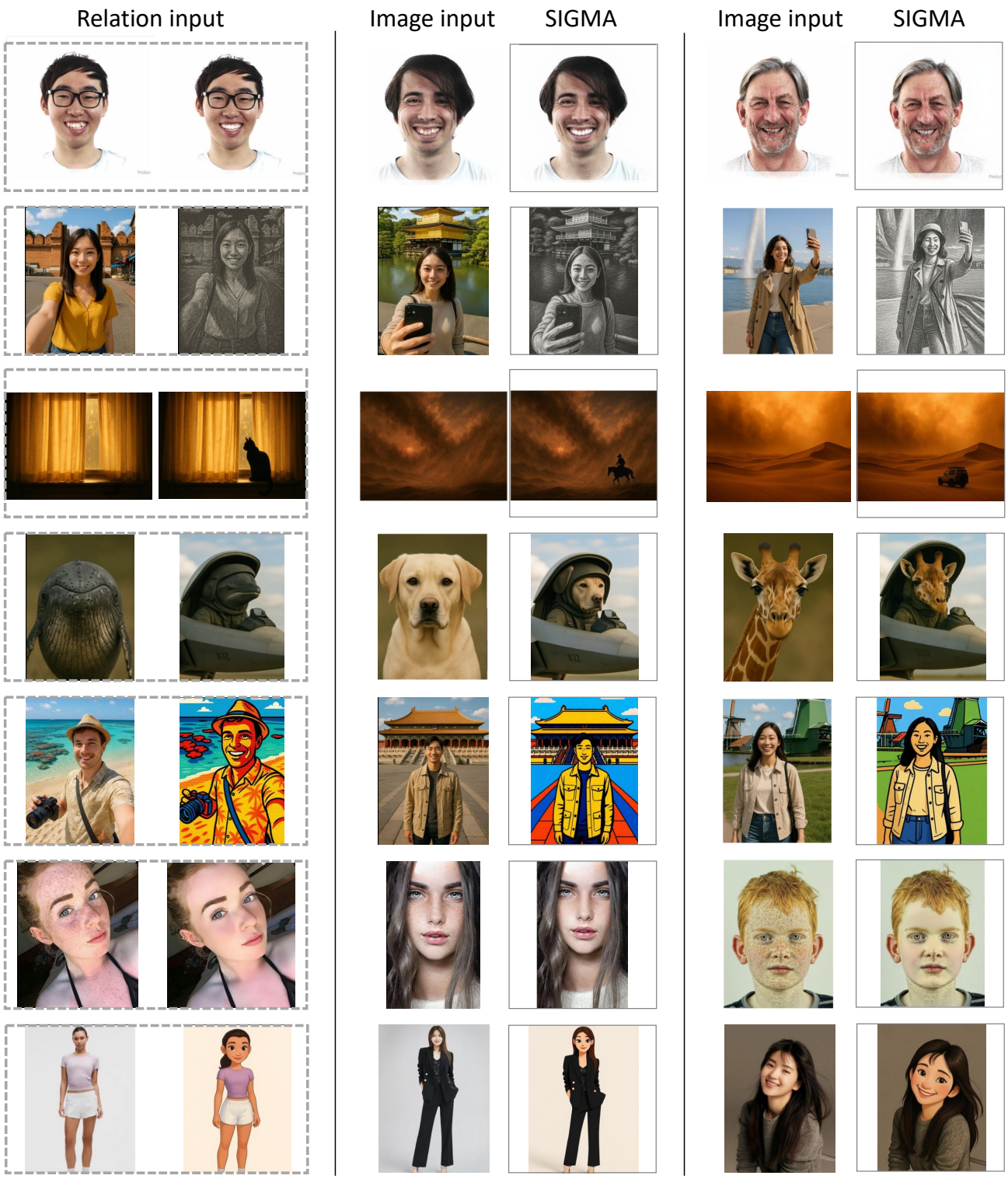


Figure 14. Additional relation-transfer results. SIGMA reproduces the relational configuration between reference objects, such as relative pose, interaction, or style, while allowing appearance variations. These examples demonstrate SIGMA's ability to capture higher-order visual relationships beyond object-level cues.

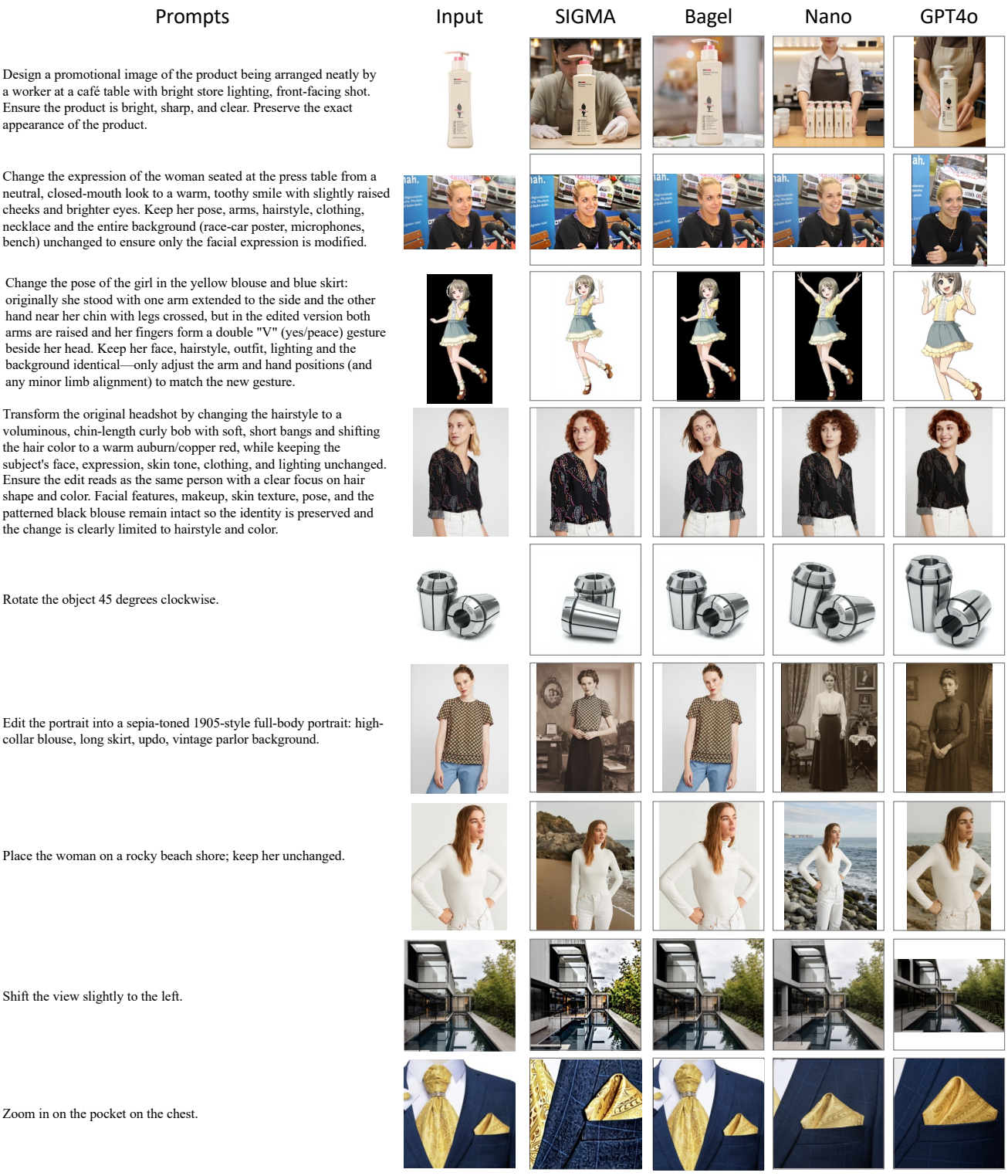


Figure 15. Image-editing results. SIGMA modifies the requested content according to the text prompts while preserving surrounding regions, global scene layout, and overall visual coherence.