

## 1 Appendix

### 2 1 Broader Impact Statement

LayerCraft significantly advances text-to-image (T2I) generation by providing precise control over composition and object integration, driven by Chain-of-Thought (CoT) reasoning. This research promises substantial positive societal impact, primarily by **democratizing visual content creation** for non-experts, making sophisticated design tools accessible to a broader audience. It will **revolutionize creative and professional workflows** in industries like advertising and gaming, drastically accelerating content creation and fostering innovation. This capability also **catalyzes new forms of digital storytelling and education**, enabling richer visual narratives. While acknowledging risks like misuse for misinformation or bias propagation, which we condemn and will address through ethical guidelines and further research, our core focus remains on LayerCraft’s transformative power to empower human creativity and broadly benefit society.

### 2 2 More Examples on Batch Collage Editing

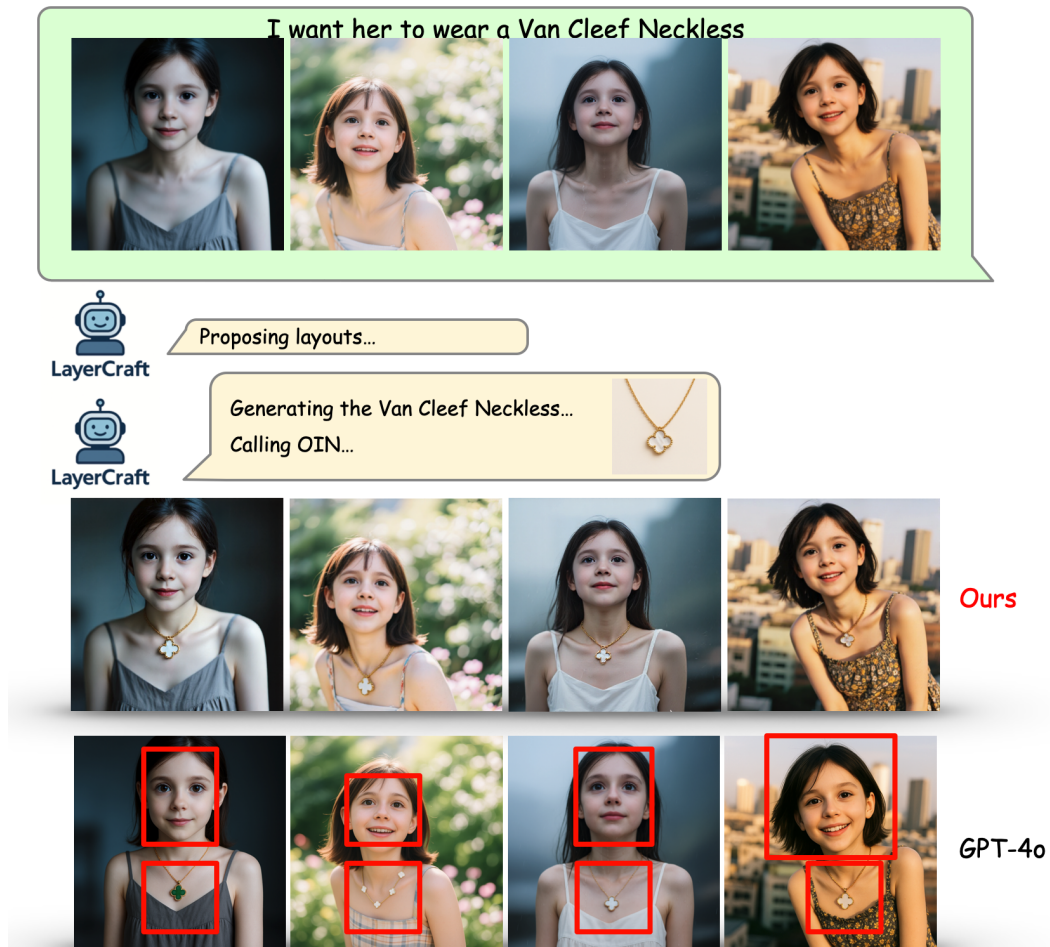
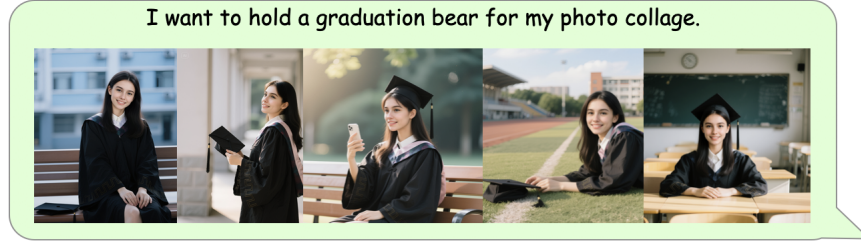


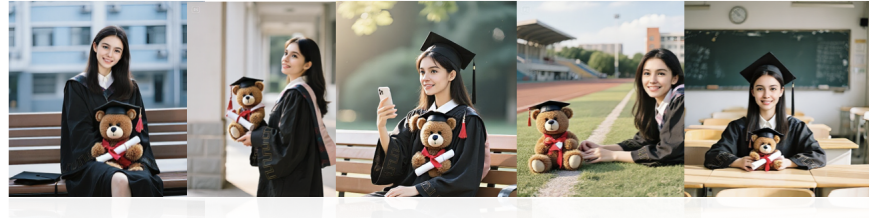
Figure 1: An example of batch collage image editing. LayerCraft effectively proposes bounding boxes for necklace placement and generates a consistent reference image, leading to seamless integration across multiple images with a single prompt. In contrast, GPT-4o fails to preserve facial identity and generates inconsistent necklaces as highlighted by the red boxes.



Generating the graduation bear...



Analyzing Images...  
Proposing Layouts and prompts...  
Calling OIN...  
Here is the result!



Proposing layouts...



Generating the white blazer...  
Calling OIN...



Figure 2: This figure provides more user scenario, the upper part is a larger and clearer demo for the teaser image. The lower panel demonstrates an outfit modification, showing a black man's attire seamlessly changed to a white blazer. These examples highlight the robust capabilities of our model.

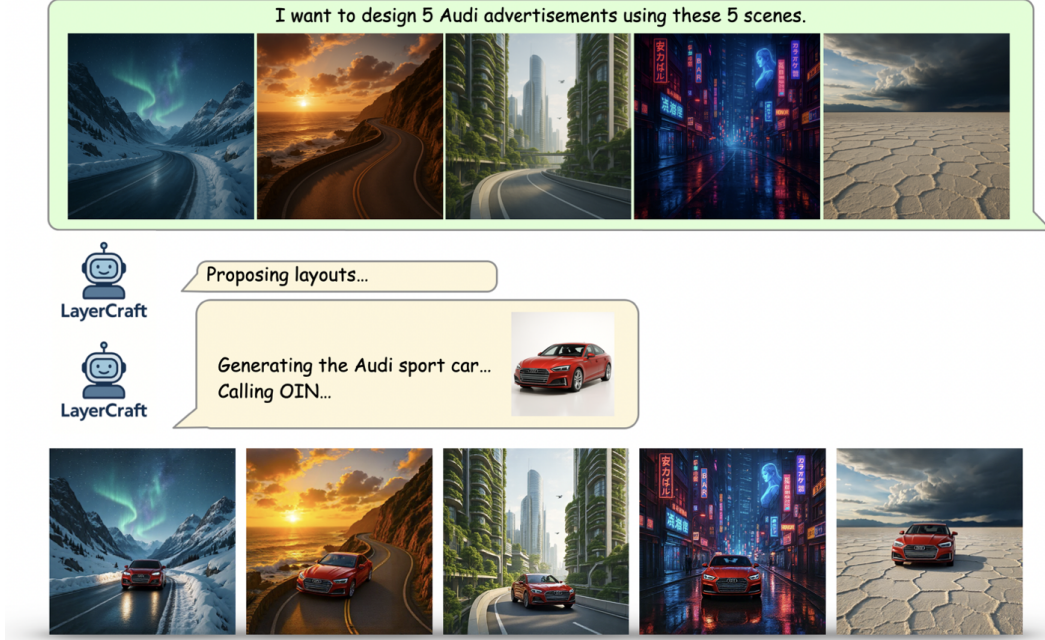


Figure 3: Further examples of batch collage image editing and generation. It illustrates the generation of a consistent Audi advertisement featuring a single car across five distinct scenes.

In this section, we present additional examples of batch collage image editing. Figure 1 illustrates LayerCraft’s ability to seamlessly integrate a Van Cleef necklace across multiple photos of a girl. Our model first intelligently identifies optimal placement bounding boxes for the necklace, then generates a consistent reference image to ensure uniformity throughout the process before engaging the Object Integration Network (OIN) for the final result. In contrast, GPT-4o struggles with this task, failing to preserve facial identity and generating inconsistent necklaces, as highlighted by the red boxes. Figure 2 provides further demonstrations of our model’s robust capabilities. The upper panel shows the zoomed-in version of the teaser image. The lower panel showcases a striking outfit modification, seamlessly changing a black man’s attire to a white blazer. Figure 3 effectively illustrates the generation of a cohesive Audi advertisement, featuring a single car consistently integrated across five distinct scenes.

### 3 Additional comparisons on T2I generation with other SOTA methods

In this section, Figure 9 presents a detailed qualitative comparison with state-of-the-art methods, including expanded versions of examples from the main paper for clearer visualization. As demonstrated, our model consistently yields better results, particularly in terms of object numeracy and accurate spatial relationships. Furthermore, LayerCraft exhibits significantly fewer artifacts compared to other agent-based methods. For instance, LLM Blueprint generates an anomalous red object beneath the table in the hot dog example. GenArtist, even in its teaser image, struggles with perspective accuracy: while the hotdogs are in focus, the distant car and bike remain sharply defined despite the blurry far end of the table, diminishing overall realism. Our method, conversely, avoids such inconsistencies, producing more coherent and realistic compositions.

## 4 Additional Analysis on Object Integration Network

### 4.1 Ablation on Attention Mixing

In our work, attention outputs are derived through the strategic blending of dual attention maps utilizing mask indices in the latent space. Specifically, our approach computes attention outputs

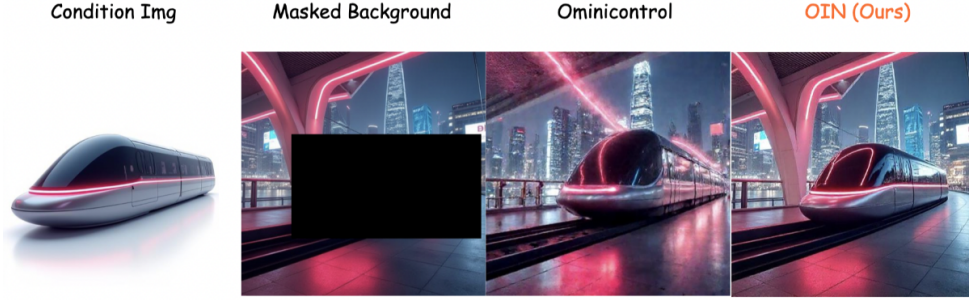


Figure 4: Failure case for direct attention mix in Section 4.1“Analysis of Objection Integration Network”. One can see the background is changed and the boundary is easy to see.

independently for each branch and subsequently integrates the hidden states according to a latent mask derived from the original masked region. This strategic integration ensures that the model effectively learns optimal object placement within the background while preserving background integrity. Furthermore, since the Multi-Modal attention mechanism within the FLUX architecture processes textual and image tokens concurrently, we address the cross-modal correlations by implementing a weighted average of textual tokens from both branches. This dual-branch integration synergistically enhances the model’s comprehension of the conditional inputs. To validate our approach, we conducted comparative analyses against two alternative methodologies. The first alternative, inspired by OmniControl [2], involves extending the input sequence and computing the attention matrix for the entire augmented sequence. However, this methodology encountered significant convergence challenges during optimization. The second approach implements a weighted summation of attention outputs, which resulted in artifacts characterized by pronounced boundaries in the masked regions and unintended modifications to background elements. Figure 4 presents visual evidence of these failure cases comparing with the successful results obtained using our method. Our proposed methodology demonstrates superior performance compared to these alternatives, as illustrated in more results of OIN in Figures 7 and 8.

## 4.2 Comparisons with Concurrent Subject-driven Inpainting Methods

In addition, we provide qualitative comparisons with concurrent approaches that support subject-driven inpainting to contextualize our contributions as being on par with, or even surpassing, the current state of the art [1, 3]. Figure 5 displays these results. A detailed examination reveals that our Object Integration Network (OIN) excels in capturing finer details, as exemplified by the orange bottle where text and logos are well-preserved, unlike the blurred text in ACE++’s output or the reverted logo in EasyControl’s result. Another instance is the violin, whose features (e.g., the chin rest) are maintained only by OIN, while they disappear in the results from ACE++ and EasyControl. Similarly, the lighted earring on the lady vanishes in EasyControl’s output. While these concurrent methods could theoretically replace OIN within our pipeline, OIN offers superior efficiency by requiring only a single loading and unloading of models, as it only necessitates loading the correct LoRA weights. In terms of speed, OIN and ACE++ generate results within one minute, depending on GPU type, whereas EasyControl is two to three times slower, further underscoring OIN’s significant contribution to the overall pipeline.

## 5 Handling Difficult Bounding Box Proposal.

To propose a correct bounding box using LLM only will be difficult sometimes, so we adopt a iterative refinement process to solve this problem. Figure 6 demonstrates the process. Initially, the ChainArchitect proposes an initial bounding box. This proposal is then visualized directly on the image, allowing the LLM to iteratively refine its size and position. This refinement process proves highly effective in achieving appropriate bounding box localization.

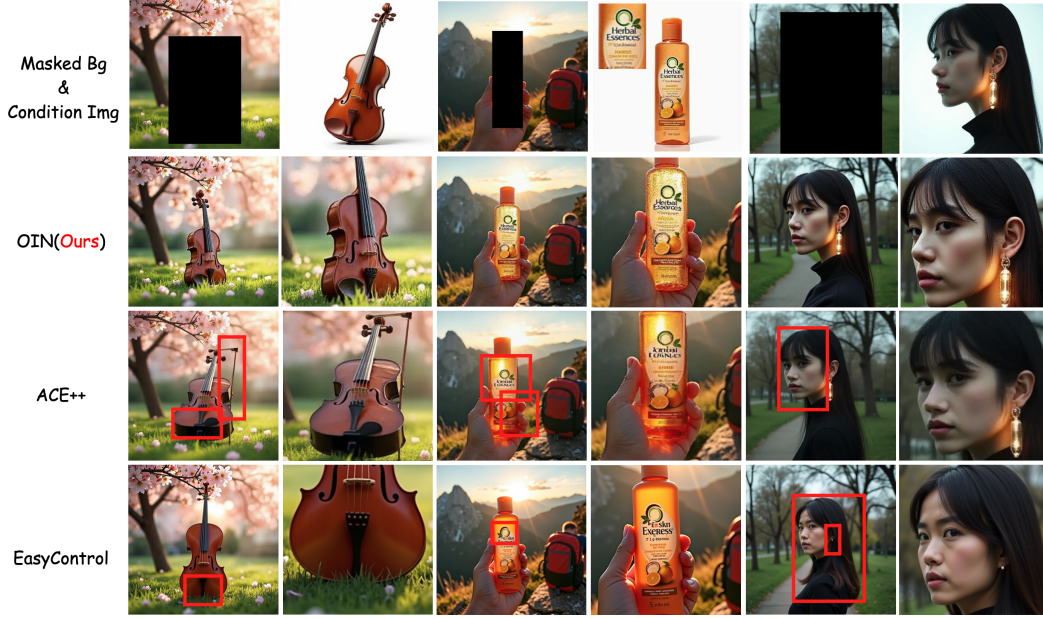


Figure 5: Visual comparisons with concurrent works on subject-driven inpainting task. The first row is our Object-Integration Network, the second row is ACE++ [1], and the last row is EasyControl [3]

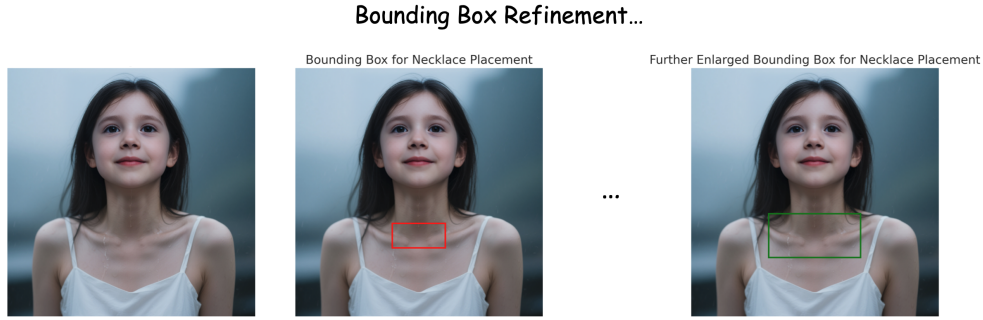
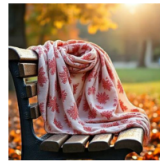
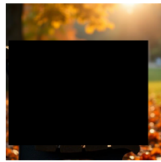
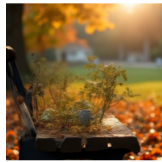


Figure 6: Our approach to handling bounding box proposals in editing tasks demonstrates remarkable effectiveness. Initially, the Large Language Model (ChainArchitect) proposes an initial bounding box. This proposal is then visualized directly on the image, allowing the LLM to iteratively refine its size and position. This refinement process proves highly effective in achieving appropriate bounding box localization.

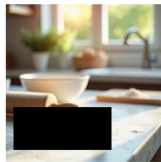
## 75 6 Overall Process Decomposed Demo

76 In this section we will demonstrate how the process happen within the framework. We use the  
 77 example of living room decoration to demonstrate the CoT thinking process of the framework, it is  
 78 demonstrated at the end of the supplementary materials.

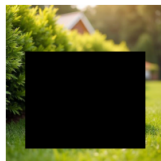
Fill the blank area in the given background with A patterned cotton scarf for diverse fashion styles., with description: Draped gracefully over a sunlit park bench, it rustles gently in the cool autumn...



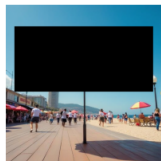
Fill the blank area in the given background with A simple metal measuring scoop., with description: In a brightly lit kitchen, it rests on a marble countertop beside a flour-dusted rolling pin and a m...



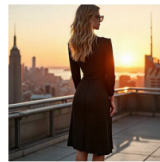
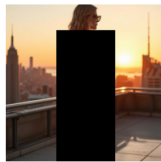
Fill the blank area in the given background with Organizes and safeguards photography gear efficiently., with description: Surrounded by lush green shrubs, it reclines on the vibrant lawn of a country...



Fill the blank area in the given background with A large, colorful beach shade., with description: Situated on a crowded boardwalk, it provides shade to street performers and curious onlookers, reflec...



Fill the blank area in the given background with A chic and modern formal outfit., with description: Across a rooftop terrace, it is portrayed in the golden hour light, slightly silhouetted against th...



Fill the blank area in the given background with Versatile kitchen appliance with multiple cooking functions., with description: In a modern minimalist kitchen, positioned symmetrically on the island ...



Figure 7: More results of OIN

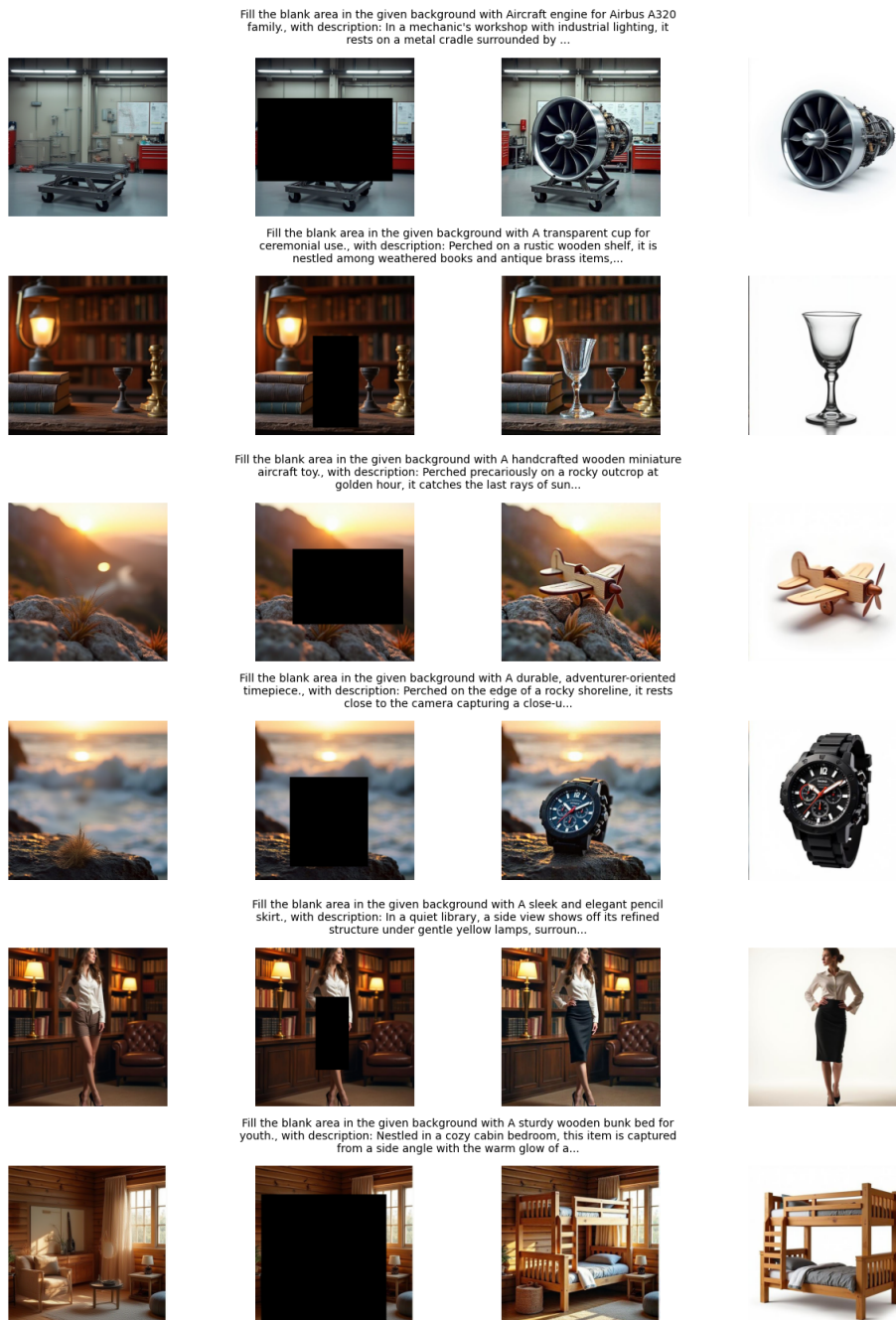


Figure 8: More results of OIN

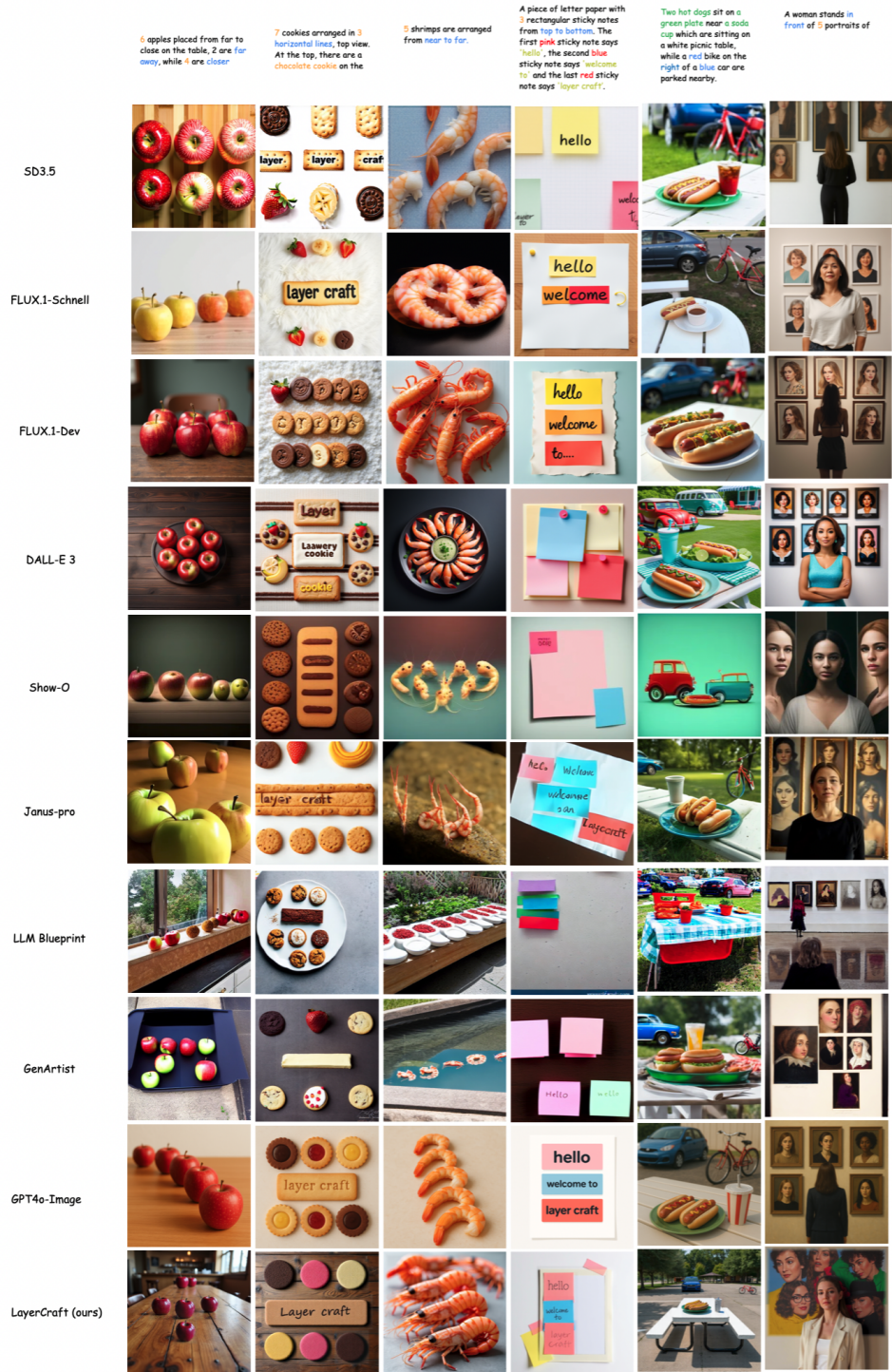


Figure 9: More comparisons with state-of-the-art models. We also include the examples used in the main paper and make them larger for details.

79 USER: I want to decorate this room.

80 LAYERCRAFT COT THINKING PROCESS: This image shows a minimalist, modern living  
81 room with a clean, warm aesthetic. Let me identify the objects and describe the  
82 background in JSON format.

```
83 {  
84   "objects": [  
85     {  
86       "name": "reading nook",  
87       "description": "A built-in wooden seating area with  
88         dark cushions and white pillows nestled within  
89         the wooden shelving unit"  
90     },  
91     {  
92       "name": "pendant light",  
93       "description": "A circular, disc-shaped hanging  
94         light fixture with a diffuse glow suspended  
95         from the ceiling"  
96     },  
97     {  
98       "name": "area rug",  
99       "description": "A large, neutral-toned rectangular  
100         rug covering part of the wooden floor"  
101     },  
102     {  
103       "name": "shelving unit",  
104       "description": "Built-in light wood shelving  
105         spanning the back wall with various decorative  
106         items displayed"  
107     },  
108     {  
109       "name": "decorative objects",  
110       "description": "Small sculptures, books, and framed  
111         artwork arranged sparsely on the shelves"  
112     },  
113     {  
114       "name": "sliding glass doors",  
115       "description": "Large floor-to-ceiling glass doors/  
116         windows on the left side offering views to the  
117         outdoors"  
118     }  
119   ],  
120   "background": {  
121     "description": "A spacious, minimalist room with  
122       light wooden flooring, white ceiling, and light  
123       wood wall paneling. The room features built-in  
124       shelving along the back wall and large sliding  
125       glass doors that allow natural light to flood the  
126       space. The viewpoint is from the center of the  
127       room looking toward the back wall with the built-  
128       in shelving and seating area, with the glass  
129       doors visible on the left side.",  
130     "included_elements": ["wooden flooring", "white  
131       ceiling", "light wood wall paneling", "natural  
132       lighting"]  
133   }  
134 }  
135  
136
```

137 The shelving unit and built-in seating are considered separate objects but could be  
138 viewed as part of the architectural design. I've included the structural elements like  
139 flooring, ceiling, and wall paneling in the background description since they define  
140 the space rather than being foreground objects.

Looking at the image, I can suggest several items that could be added to enhance the minimalist space while maintaining its clean aesthetic. Let me think about what would complement this room:

```
{
  "suggested_additions": [
    {
      "name": "coffee table",
      "description": "A low, rectangular wooden coffee
        table with clean lines, placed in front of the
        reading nook on the area rug"
    },
    {
      "name": "side table",
      "description": "A small, round side table in light
        wood or brass next to the reading nook"
    },
    {
      "name": "floor lamp",
      "description": "A slim, architectural floor lamp
        with warm lighting placed in the corner
        opposite the glass doors"
    },
    {
      "name": "accent chair",
      "description": "A minimalist armchair in a
        complementary neutral tone (tan leather or
        light gray fabric) positioned at an angle to
        the reading nook"
    },
    {
      "name": "throw blanket",
      "description": "A textured throw blanket in a
        subtle pattern or muted color draped over one
        corner of the reading nook"
    },
    {
      "name": "indoor plant",
      "description": "A tall potted plant like a fiddle
        leaf fig or snake plant placed near the glass
        doors to bring in natural elements"
    },
    {
      "name": "decorative bowl",
      "description": "A sculptural bowl or vessel on one
        of the open shelves to add visual interest"
    }
  ]
}
```

These additions would enhance the functionality and warmth of the space while respecting its minimalist design language and natural materials.

Analyzing the image based on the provided grid regions, I'll determine suitable placement areas for the suggested objects while considering perspective and available space:

```
{
  "object_placements": [
    {
      "object": "coffee table",
      "suitable_regions": ["center", "between center and
        center-right", "between lower center and center",
        "between lower center-right and center"],
    }
  ]
}
```

```

204         "reasoning": "The central floor area on the rug
205             provides ample space for a coffee table,
206             positioned in front of the reading nook"
207     },
208     {
209         "object": "side table",
210         "suitable_regions": ["center-right", "between
211             center and center-right"],
212         "reasoning": "A small side table would fit well
213             beside the reading nook on the right side"
214     },
215     {
216         "object": "floor lamp",
217         "suitable_regions": ["lower left", "between lower
218             left and center-left"],
219         "reasoning": "The corner area opposite the glass
220             doors provides good placement for a tall floor
221             lamp without blocking views or pathways"
222     },
223     {
224         "object": "accent chair",
225         "suitable_regions": ["lower center-left", "between
226             lower center-left and center", "center-left", "
227             between center-left and center"],
228         "reasoning": "An accent chair would fit well in
229             this area, creating a conversation zone with
230             the reading nook while maintaining open space"
231     },
232     {
233         "object": "throw blanket",
234         "suitable_regions": ["center-right"],
235         "reasoning": "The throw blanket would be placed on
236             the reading nook which is already in this
237             region"
238     },
239     {
240         "object": "indoor plant",
241         "suitable_regions": ["upper left", "between upper
242             left and center-left"],
243         "reasoning": "Near the glass doors to receive
244             natural light while adding greenery to that
245             corner of the room"
246     },
247     {
248         "object": "wall art",
249         "suitable_regions": ["upper center-left", "upper
250             center", "between upper center-left and center"
251             , "between upper center and center"],
252         "reasoning": "The visible wall space opposite the
253             shelving unit would accommodate wall art while
254             maintaining visual balance"
255     },
256     {
257         "object": "decorative bowl",
258         "suitable_regions": ["center-right", "between
259             center and center-right"],
260         "reasoning": "Could be placed on one of the
261             existing shelves in the built-in unit"
262     }
263 ],
264 "spatial_considerations": {
265     "viewing_perspective": "The image is taken from a
266         position looking toward the back wall with the
267         built-in shelving, with the glass doors on the
268         left side",

```

```

269     "floor_space": "The central and lower areas have the
270         most available floor space for furniture
271         placement",
272     "traffic_flow": "Pathways should be maintained
273         between the entrance (presumed to be behind the
274         viewing position) and the glass doors",
275     "focal_points": "The reading nook and shelving unit
276         already serve as focal points, so additional
277         elements should complement rather than compete
278         with them"
279 }
280 }
281

```

282 This analysis considers both the physical space available in each region and the de-  
283 sign principles of balance, flow, and functionality while maintaining the minimalist  
284 aesthetic of the room.

285 Based on our analysis, here are the bounding boxes and prompts for each suggested  
286 object, ordered from far to close in terms of generation order:

```

287 {
288     "objects": [
289         {
290             "type": "indoor plant",
291             "position": "upper left",
292             "generation_order": 1,
293             "prompt": "A tall, elegant fiddle leaf fig plant
294                 with large, glossy green leaves in a minimalist
295                 ceramic pot, placed near large windows to
296                 receive natural light, rendered in a clean,
297                 contemporary style that complements modern
298                 interior design.",
299             "bounding_box": [50, 150, 150, 350]
300         },
301         {
302             "type": "decorative bowl",
303             "position": "center-right",
304             "generation_order": 2,
305             "prompt": "A sculptural, handcrafted ceramic bowl
306                 in a soft matte finish with subtle organic
307                 texture and asymmetrical form, placed on an
308                 open shelf, rendered in a minimalist
309                 Scandinavian style.",
310             "bounding_box": [420, 220, 470, 260]
311         },
312         {
313             "type": "throw blanket",
314             "position": "center-right",
315             "generation_order": 3,
316             "prompt": "A soft, textured throw blanket in muted
317                 beige with subtle geometric pattern, casually
318                 draped over the corner of the built-in seating
319                 area, rendered in a warm, inviting style that
320                 adds comfort to the minimalist space.",
321             "bounding_box": [420, 250, 500, 300]
322         },
323         {
324             "type": "side table",
325             "position": "between center and center-right",
326             "generation_order": 4,
327             "prompt": "A small, round side table with slender
328                 brass legs and a light oak top, positioned
329                 beside the reading nook at perfect arm's reach
330                 height, rendered in a clean, architectural
331                 style.",
332

```

```

333     "bounding_box": [350, 250, 420, 320]
334   },
335   {
336     "type": "floor lamp",
337     "position": "between lower left and center-left",
338     "generation_order": 5,
339     "prompt": "A tall, architectural floor lamp with a
340               slender brushed brass stem and minimal white
341               shade casting a warm glow, positioned in the
342               corner of the room, rendered in a contemporary
343               style that emphasizes clean lines.",
344     "bounding_box": [70, 320, 150, 450]
345   },
346   {
347     "type": "accent chair",
348     "position": "center-left to between center-left and
349               center",
350     "generation_order": 6,
351     "prompt": "A minimalist lounge chair with gentle
352               curves, upholstered in natural tan leather with
353               a light wooden frame, positioned at an
354               inviting angle to create conversation space,
355               rendered in a Scandinavian modern style.",
356     "bounding_box": [100, 250, 250, 380]
357   },
358   {
359     "type": "coffee table",
360     "position": "center to between lower center and
361               center",
362     "generation_order": 7,
363     "prompt": "A low, rectangular coffee table with
364               clean lines and rounded corners, crafted from
365               light oak with a subtle grain pattern and
366               minimalist design, positioned centrally on the
367               area rug, rendered in a contemporary style that
368               balances form and function.",
369     "bounding_box": [200, 300, 350, 400]
370   }
371 ]
372 }
373

```

374 This arrangement considers depth perception (generating background objects first),  
375 spatial relationships, and the existing elements in the room. The bounding boxes  
376 are designed to place objects in appropriate locations while maintaining a balanced,  
377 functional layout that enhances the minimalist aesthetic of the space.

378 USER: I've done the previous decorating and want to put this cute teddy bear in.

379 LAYERCRAFT: Looking at the new item you want to add to the room, I can help place  
380 the teddy bear in the previously discussed space.

```

381 {
382   "additional_object": [
383     {
384       "type": "teddy bear",
385       "position": "lower center-right",
386       "generation_order": 8,
387       "prompt": "A small, golden-brown teddy bear with a
388                 smiling face and soft plush texture, sitting
389                 upright on one of the built-in shelves, adding
390                 a touch of warmth and playfulness to the
391                 minimalist space.",
392       "bounding_box": [290, 300, 480, 490]
393     }
394   ]
395 }

```

396  
397

}

398

These placements would work well with your previous decorating plans.

399 **References**

- 400 [1] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou.  
401 Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint*  
402 *arXiv:2501.02487*, 2025.
- 403 [2] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and  
404 universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024.
- 405 [3] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and  
406 flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025.