MICON-Bench: Benchmarking and Enhancing Multi-Image Context Image Generation in Unified Multimodal Models

Supplementary Material

7. Limitations

Despite the promising results, our current evaluation pipeline inherently relies on the underlying MLLM's perception capability, which is itself susceptible to hallucinations. Consequently, erroneous model predictions may propagate into the evaluation process, potentially biasing the reported performance. Moreover, the effectiveness of the proposed DAR-based re-weighting mechanism remains constrained by the quality of the original attention maps. When the base model fails to correctly interpret the reference image—e.g., missing fine-grained semantics or mislocalizing key regions—the re-weighted attention may still amplify incorrect signals rather than rectify them.

8. Benchmark Detail

8.1. Data Statistics

Dataset Composition. MICON-Bench contains six types of multi-image context generation tasks: Object Composition, Spatial Composition, Attribute Disentanglement, Component Transfer, Foreground/Background (FG/BG) Composition, and Story Inference. The per-task number of cases is visualized in Figure 5 (a). Beyond case counts, we further summarize the total number of images and how many cases use two or three reference images, as shown in Table 6. Overall, the benchmark comprises 1,043 cases and 2,518 images, among which 611 cases use two reference images and 432 cases use three reference images.

Table 6. Statistics of MICON-Bench, including the number of cases, images, and the distribution of two- and three-reference-image settings for each task.

Task	#Cases	#Images	#2 Ref	#3 Ref
Object	200	482	118	82
Spatial	200	498	102	98
Attribut	100	300	0	100
Component	240	601	119	121
FG/BG	200	400	200	0
Story	103	237	72	31
Total	**1,043**	**2518**	**611**	**432**

Evaluation Checkpoints. For automatic evaluation, we adopt the Evaluation-by-Checkpoint paradigm described in the main paper, where an MLLM verifies whether each generated image satisfies a set of fine-grained visual and semantic conditions. For each task, we design five evaluation dimensions (e.g., instruction following, identity preservation, structural consistency, cross-reference consistency, and overall usability) and instantiate several binary checkpoints under each dimension. Figure 5 (b) summarizes the percentage of checkpoints per dimension for all six tasks.

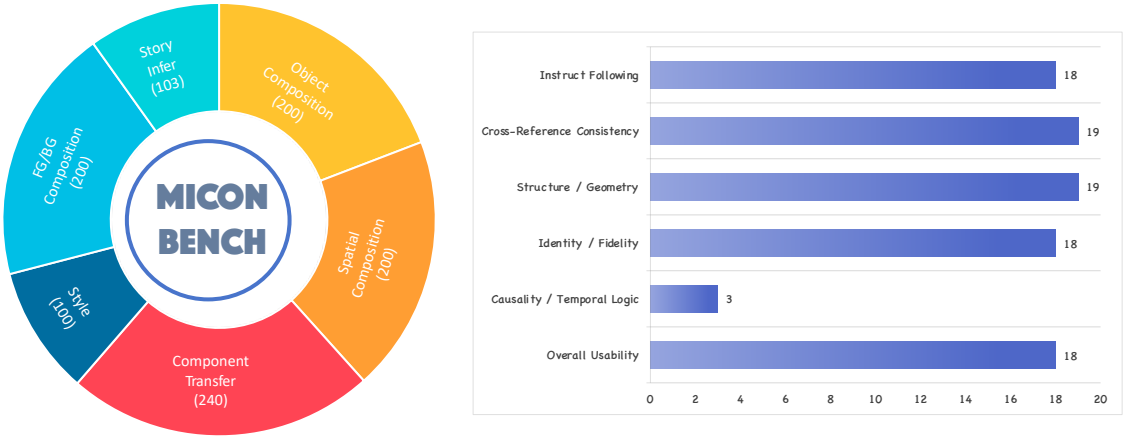
8.2. Benchmark Construction

Image Generation. We employ Qwen-Image to generate image data for five fundamental compositional tasks: Object Composition, Spatial Composition, Attribute Disentanglement, Component Transfer, and Foreground/Background (FG/BG) Composition. We first construct comprehensive collections of subjects, attributes, spatial relations, and scenes. From these collections, we sample elements to fill the reference image prompt templates, which guide Qwen-Image in generating the reference images for each task. Simultaneously, we use the same sampled elements to populate the task prompt templates, thereby forming the task prompts used in our MICON-Bench. This approach ensures that the reference images and corresponding task prompts are naturally aligned and rooted in a shared, well-defined compositional space, providing consistency between the data generation and evaluation processes. The prompt templates for each task are shown in Table 13.

For the **Story Inference** task, we construct a dataset of short visual narratives that require causal reasoning over multiple reference images. The visual stories are categorized into three types: (1) *physical property changes* (e.g., a candle being blown out), (2) *causal commonsense* (e.g., everyday cause–effect scenarios), and (3) *action continuation* (e.g., predicting the next step of an ongoing action).

For each accepted script, we generate a four-panel comic using image generation models and manually crop each panel to obtain clean reference images. We further apply DiffBIR-based restoration to enhance sharpness and reduce artifacts, resulting in the final Story Inference reference sets used in our benchmark. The model is then asked to infer a plausible continuation of the story given two or three reference images.

Data Filtering. To ensure high-quality and semantically consistent dataset construction, we employ a two-stage filtering pipeline that combines automatic and manual reviews. First, a Multimodal Large Language Model



(a) Case Statistics of MICON-Bench

(b) Checkpoint Percentage of MICON-Bench (%)

Figure 5. The data statistics of MICON-Bench.

(MLLM) is utilized to perform coarse-grained automatic filtering by verifying whether generated images faithfully adhere to the corresponding textual prompts. The MLLM evaluates attributes such as object completeness, spatial relationships, style consistency, and component correctness, effectively identifying blatant failures or inconsistencies at scale. Subsequently, filtered samples undergo meticulous human inspection to catch subtle issues beyond automated detection, including unnatural compositions, visual artifacts, and nuanced spatial or attribute violations. This dual-layer filtering protocol helps ensure that the final dataset achieves both diversity and high fidelity, providing a reliable benchmark for compositional visual understanding and generation.

8.3. Evaluation Detail

Checkpoints Generation. We adopt a fine-grained, checkpoint-based evaluation protocol powered by an MLLM judge. Across all tasks, we define seven evaluation dimensions:

- **A. Instruction Following** – whether the generated image follows the textual instruction and task specification.
- **B. Identity / Fidelity** – whether object identities and key attributes match their references.
- **C. Structure / Geometry** – whether spatial layout, geometry, and physical plausibility are preserved.
- **D. Cross-Reference Consistency** – whether information from multiple reference images is integrated without contradiction.
- **E. Causality** – whether temporal progression and cause–effect relations are reasonable.
- **F. Text Grounding** – whether textual content (e.g., overlaid text) is correctly rendered and grounded.
- **G. Overall Usability** – whether the final image is natural, coherent, and visually usable.

Different tasks activate different subsets of these dimensions according to their nature (e.g., Foreground/Background Composition focuses on A/B/C/D/G, while Story Inference emphasizes A/B/C/D/E/G). Within each active dimension, we design 2–4 concrete *checkpoints* that specify what the MLLM should verify. For each test case, we feed the instructions, reference images, and the generated image into the MLLM, and ask it to decide, for every checkpoint, whether the requirement is satisfied (pass/fail) along with a short justification.

We further mark one key checkpoint per dimension as a **hard constraint**, denoted by $(\mathrm{H}x)$ (e.g., A_check_1 (H1), B_check_1 (H2)). If a hard-constraint checkpoint is judged as failed, the score of that dimension is capped at 0.4 (on a $[0, 1]$ scale), regardless of other checkpoints in that dimension. For each sample, we convert the proportion of satisfied checkpoints in each dimension into a dimension score in $[0, 1]$, aggregate over all active dimensions, and linearly rescale the result to obtain a final score in $[0, 100]$.

Example of Checkpoints. The checkpoints of the Object Composition task are shown in Table 14. Other tasks (e.g., Spatial Geometric Constraints, Foreground/Background Composition, Story Inference, and Text-based Editing) are defined analogously, each with a tailored subset of dimensions and a small set of task-specific checkpoints under the same unified framework.

Evaluation of Story Generation. For the complex Story Inference task, which involves causal reasoning over multiple images, we adopt a hybrid evaluation scheme that combines traditional MLLM checkpoint scoring with a human-constructed answer set. Concretely, the final score

Human-Annotated Answer Set


```
"cases": [
  {
    "case_id": "case_00xx",
    "background_summary": "This is a sequence showing a child gradually blowing
on a candle flame.",
    "positive_candidates": [
      "The candle flame flickers under the child's breath.",
      "The flame weakens and eventually goes out, leaving smoke from the wick.",
      "The child continues blowing until the candle is extinguished."
    ],
    "negative_candidates": [
      "The flame stays perfectly steady without reacting to the blowing.",
      "The child blows, but the flame grows taller instead of weakening.",
      "The candle shows no change despite repeated blowing."
    ]
  },
```


Figure 6. Human-annotated answer set for the Story Inference task. Each example pairs the reference image sequence with canonical human-written descriptions of what should happen next, which are used as targets for MLLM-based evaluation.

is a weighted sum of two components: (i) the standard checkpoint-based MLLM score described above (40%), and (ii) a human-answer-set score (60%).

For the human-answer-set component, we construct a small set of human-written targets for each story, including a concise narrative description, plausible predictions of what is likely to happen next, and counterfactual outcomes that are unlikely to occur. During evaluation, we feed the MLLM with the reference images, the model-generated continuation image, and the corresponding human answer set, and ask it to assess how well the generated image matches these canonical outcomes in terms of causal progression, visual evidence, and overall story resolution. An illustration of the human answer set and its usage in the evaluation prompt is shown in Fig. 6.

This hybrid design leverages human-authored answers as semantically precise, high-level anchors for what should happen next, reducing ambiguity and overly lenient judgments compared to using only low-level checkpoints. At the same time, combining checkpoint-based verification with answer-set matching encourages models to be accurate both at the local level and at the global level of causal consistency and narrative coherence, yielding a more robust and faithful evaluation of story generation quality.

9. Additional Experiments

9.1. Ablation Study

Inference Cost. Table 7 presents the average inference time measured on the Object Composition dataset using a single NVIDIA A800-40GB GPU. Our method (+OURS) introduces only minimal overhead compared to the respective baselines for both BAGEL and Omnigen2 across different numbers of reference images. Specifically, the runtime increase remains within a small margin (approximately 5–10%), demonstrating that the performance gains come at a low computational cost. This demonstrates the efficiency of our approach and its practicality for real-world deployment, where inference speed is critical.

Table 7. Average inference time (seconds) on Object Composition dataset measured on a single NVIDIA A800-40GB GPU.

# Reference Images	BAGEL		Omnigen2	
	Baseline	+OURS	Baseline	+OURS
2	57.13	61.70	72.01	74.71
3	69.72	77.01	97.80	101.58

Ablation of Weight Factor γ. Table 8 presents the performance of our model across six subtasks of MICON-Bench under varying weight factors γ. The results indicate that $\gamma = 0.15$ achieves the best overall performance, obtaining the highest accuracy in Object, Spatial, FG/BG, and Story tasks. While a smaller $\gamma = 0.05$ yields competitive results in Attribute and Spatial tasks, it underperforms notably in FG/BG and Story. As γ increases beyond 0.15, the performance across all tasks drops sharply, suggesting that overly strong weighting negatively impacts learning. Notably, at $\gamma = 0.55$, all task accuracies fall to very low levels, indicating that the model struggles to generalize due to excessive regularization. These findings demonstrate the importance of carefully tuning γ to balance the contributions of different components during training, ensuring robust performance across diverse subtasks.

Table 8. Ablation study on the effect of different weight factor γ.

γ	Object	Spatial	Attribute	Component	FG/BG	Story
0.05	86.11	91.68	**92.24**	**56.07**	59.29	64.75
0.15	**88.04**	**91.88**	90.76	56.06	**71.24**	**66.34**
0.25	76.58	85.41	75.43	44.36	64.41	62.86
0.35	53.36	45.78	37.84	23.76	65.14	58.96
0.55	8.39	18.30	0.67	8.39	22.77	40.44

9.2. Prompt Sensitivity Check

To rigorously validate the reliability of our Evaluation-by-Checkpoint framework, we investigate whether the MLLM-based verifier is overly sensitive to the specific phrasing of the evaluation checkpoints. We conduct a prompt sensitivity ablation study using the BAGEL model. We randomly sample a subset of 180 cases from the MICON-Bench dataset, ensuring an even distribution across all six task categories (30 cases per task). For each case, we design two distinct sets of evaluation checkpoints:

- **Generic Checkpoints:** These utilize abstract, template-based descriptions (e.g., *"Does the image include all specified objects?"*).

Table 9. Ablation study on prompt sensitivity check for the Evaluation-by-Checkpoint.

Model	Object	Spatial	Attribute	Component	FG/BG	Story	Avg. Score
BAGEL(Generic)	87.72	84.49	87.51	47.33	51.78	63.66	68.52
BAGEL(Specific)	88.17	84.09	85.20	46.09	51.78	63.51	68.01
BAGEL+**DAR**(Generic)	91.22	87.81	89.24	48.49	78.22	68.86	75.84
BAGEL+**DAR**(Specific)	91.72	87.53	89.42	47.67	80.00	67.51	75.92

Table 10. Performance comparison of different models on our MICON-Bench evaluated by InternVL3.5-38B.

Model	Object	Spatial	Attribute	Component	FG/BG	Story	Avg. Score
Nano-Banana [11]	97.88	98.18	72.03	79.41	83.71	79.06	86.63
GPT-Image [14]	98.85	97.25	83.70	85.26	91.08	81.58	90.77
UNO [39]	62.30	57.75	55.11	27.56	28.47	38.17	43.87
DreamOmni2 [40]	88.95	81.69	62.64	65.52	81.01	59.94	75.26
Qwen-Image-Edit-2507 [34]	98.03	97.57	65.05	57.12	74.95	60.74	77.26
BAGEL [7]	89.08	89.32	69.05	59.57	72.39	54.32	73.78
BAGEL + **DAR**	**89.56**	**91.57**	**72.39**	**66.05**	**80.33**	**60.93**	**78.29**
OmniGen2 [35]	88.77	81.91	69.13	47.53	69.98	57.37	69.38
OmniGen2 + **DAR**	**89.50**	**83.06**	**70.71**	**49.79**	**72.26**	**61.15**	**71.22**

- **Specific Checkpoints:** These replace abstract terms with concrete, instance-level entities derived directly from the user prompt (e.g., *"Does the image include a giraffe and a wooden chair?"*).

The quantitative results are summarized in Table 9. The performance of BAGEL evaluated under Generic checkpoints is 68.52, while the score under Specific checkpoints is 68.01. This high degree of consistency across all six tasks demonstrates that our chosen MLLM verifier exhibits robust reasoning capabilities. It successfully comprehends the core intent of the evaluation dimensions without being biased by phrasing variations, thereby confirming that our automatic evaluation metric is both stable and reliable.

9.3. Verifier Generalization

To demonstrate the robust generalization of our Evaluation-by-Checkpoint framework, we replace our default verifier, Qwen3-VL-32B-Instruct, with another state-of-the-art open-source MLLM, InternVL3.5-38B. We re-evaluate all generated images across the entire MICON-Bench dataset (all cases, without sampling) using the exact same checkpoints.

The results are presented in Table 10. As the results show, despite changing the verifier, the relative performance rankings of all evaluated models remain perfectly consistent with our main results. This comprehensive evaluation on the full dataset proves that our framework successfully captures fundamental semantic and visual alignments, rather than idiosyncratic biases of a single MLLM, demonstrating strong robustness to the choice of verifier.

9.4. Human Alignment Check

To ensure that our automated Evaluation-by-Checkpoint metric serves as a reliable proxy for actual human perception, we conduct a rigorous human alignment study. We randomly sample a subset of 120 cases from MICON-Bench, ensuring a strictly balanced distribution of 20 cases for each of the six tasks. Three expert human annotators are instructed to manually evaluate the generated images. To ensure a fair comparison, the human annotators are provided with the exact same binary, hard-constraint checkpoints (e.g., identity preservation, spatial relationships) that were fed to the MLLM verifier.

The human evaluation scores on the sampled subset are reported in Table 11. And the Table 12 provides a comprehensive comparison between human judgments and our default verifier. As illustrated in the summary table, the MLLM evaluator achieves exceptionally high consistency with human judgments. The average deviation from human scores across the models is marginal (only $+0.67$). The relative rankings of the models provided by the automated verifier perfectly mirror the human rankings. These findings confirm that our hard-constraint rubric effectively bridges the gap between automatic metrics and human visual perception, validating the reliability of our benchmark.

Table 11. **Human Evaluation** results on the uniformly sampled subset of the MICON-Bench.

Model	Object	Spatial	Attribute	Component	FG/BG	Story	Avg. Score
Nano-Banana-Pro [11]	90.00	93.20	92.50	85.31	95.29	87.33	90.61
Nano-Banana [11]	92.45	90.34	89.23	76.35	87.67	75.00	85.17
GPT-Image [14]	92.90	93.46	91.73	79.80	87.12	90.00	89.17
UNO [39]	56.03	54.12	60.10	15.73	25.00	40.09	41.84
DreamOmni2 [40]	90.00	81.25	83.30	51.98	78.20	60.35	74.18
Qwen-Image-Edit-2507 [34]	89.40	84.30	75.00	44.76	77.95	59.50	71.82
BAGEL [7]	88.24	84.54	63.56	54.12	70.67	65.30	71.07
BAGEL + **DAR**	**89.67**	**88.36**	**63.45**	**58.34**	**74.95**	**69.73**	**74.08**
OmniGen2 [35]	86.20	73.89	52.45	41.03	60.17	56.79	61.75
OmniGen2 + **DAR**	**88.76**	**75.57**	**54.35**	**42.00**	**64.07**	**58.56**	**63.88**

Table 12. **Reliability Analysis**

Target Model	Human Score	Automatic Verifiers	
		Qwen3-VL	InternVL3.5
Nano-Banana-pro	90.61	92.42	90.93
GPT-Image	89.17	89.15	89.05
Nano-Banana	85.17	85.68	85.72
BAGEL + DAR (Ours)	**74.08**	**75.62**	**75.57**
DreamOmni2	74.18	74.58	73.36
BAGEL	71.07	71.42	71.08
Qwen-Image-Edit	71.82	72.26	72.35
OmniGen2 + DAR (Ours)	**63.88**	**64.58**	**63.32**
OmniGen2	61.75	62.44	61.12
UNO	41.84	41.62	41.44
Avg. Deviation from Human	-	*+0.67*	*+0.54*

9.5. Stability and Reproducibility of the MLLM Evaluator

To ensure that our Evaluation-by-Checkpoint framework serves as a robust scientific metric, we conduct a stability test by repeatedly evaluating the generated images of a representative model (BAGEL) using our default verifier, Qwen3-VL. Specifically, we performed five independent evaluation runs on the images generated by the BAGEL model. The final average scores for BAGEL across the five trials are remarkably consistent: Run 1 yielded 68.47; Run 2: 68.51; Run 3: 68.48; Run 4: 68.51; and Run 5: 68.47. The maximum score discrepancy across all five trials is a mere 0.04. This exceptionally low variance indicates that our strictly defined, binary-choice checkpoint mechanism effectively constrains the generative randomness typically associated with MLLMs.

9.6. More Examples

We provide two additional visual examples that further demonstrate the robustness and generalizability of our Dynamic Attention Rebalancing (DAR) approach across diverse multi-image composition scenarios. As shown in Figure 8 and Figure 9, these examples reinforce DAR's capability to precisely integrate multiple reference inputs, effectively suppressing irrelevant attention while emphasizing critical details from each source image. As shown, our method successfully avoids common pitfalls seen in baseline models, such as attribute leakage and misplaced spatial arrangements, by maintaining clear object boundaries and consistent attribute rendering. The supplemental figures also highlight DAR's flexibility in handling various composition challenges, including complex foreground-background interactions and nuanced component transfers. Collectively, these extended visualizations provide compelling qualitative evidence supporting the scalability and effectiveness of DAR in resolving cross-image inconsistencies and achieving fine-grained visual alignment in multi-image tasks.

Table 13. **Prompt Templates for Benchmark Construction.** We detail the reference image prompts and the specific task prompts used for generation, covering object composition, spatial arrangement, attribute disentanglement, component transfer, background replacement, and story inference.

Task Category	Configuration	Prompt Template
Object Composition	Reference Image	"A photo of {personalized_obj} in {scene}."
	Task (2 Refs)	"Generate an image that contains both the complete {obj_a} and {obj_b} together in {chosen_scene}."
	Task (3 Refs)	"Generate an image that contains all three objects ({obj_a}, {obj_b}, and {obj_c}) together in {chosen_scene}."
Spatial Composition	Reference Image	"A photo of {personalized_obj} in {scene}."
	Task (2 Refs)	"Generate an image that contains both the complete {obj_a} and {obj_b} together in {chosen_scene}, with {obj_a} positioned to the {spatial_relation} of {obj_b}."
	Task (3 Refs)	"Generate an image that contains all three objects ({obj_a}, {obj_b}, and {obj_c}) together in {chosen_scene}, with {left_obj} on the left, {center_obj} in the center, and {right_obj} on the right."
Attribute Disentanglement	Ref A (Subject)	"A photo of a clear {personalized_description} in {background_desc}."
	Ref B (Style)	"A photo of a {style_object} rendered in {style_desc}."
	Ref C (Background)	"A photo of beautiful {specific_background}, empty scene without main objects."
	Task (A+B+C)	"Generate an image of the {main_object} from image A, using the visual style from image B, and placing it in the {specific_background} environment from image C."
Component Transfer	Ref (Single Subject)	"A {subject_type} in {scene}, wearing {clothing_desc}, with {accessories_desc}."
	Ref (Two Subjects)	"A {subject1_type} on the {position1} wearing {cloth1} with {acc1}, and a {subject2_type} on the {position2} wearing {cloth2} with {acc2}, in {scene}."
	Task (Complex Mode)	"Extract {elements_desc} from {source_desc} in Image {source_label}, then apply these elements to {target_desc} in Image {target_label}. Create a new composition showing the target subject(s) wearing/displaying the transferred elements."
	Task (Simple Mode)	"Task: Extract only the {local_element} from the subject in Image A, then apply this element to the subject in Image B. Create a new composition showing the target subject wearing/displaying the {local_element}."
FG/BG Composition	Reference Image	"A photo of {personalized_obj} in {scene}."
	Task (2 Refs)	"Generate an image where you cleanly extract the {obj_a} from image A and replace the {obj_b} in image B. The background from image B should remain unchanged."
Story Inference	Ref Image (4-panel)	"Generate a four-panel comic with logically coherent and causally related events, following one of the specified types (physical property change, causal commonsense, or action continuation)."
	Task (2/3 Refs)	"Given the reference images, infer and generate a realistic photo of what might happen next."

Table 14. **Example of Evaluation Checkpoints.** Detailed checklist for Object Composition task.

Dimension	ID	Evaluation Question
A. Instruction Following	A_check_1 **(H1)**	Does the image contain all specified objects as required by the instructions?
	A_check_2	Are the relative arrangement and requested relations between objects correctly followed?
	A_check_3	Are there no obviously extra or missing salient elements?
B. Identity / Fidelity	B_check_1 **(H2)**	Does each object's identity strictly match its reference (e.g., category, instance)?
	B_check_2	Are the key attributes (e.g., color, texture, shape) of each object well preserved?
	B_check_3	Are the object details accurate and easily recognizable?
C. Structure / Geometry	C_check_1	Are the spatial relationships between objects consistent with the instructions and physically plausible?
	C_check_2	Are the relative sizes, proportions, and perspective of objects realistic?
D. Cross-Reference Consistency	D_check_1	Are objects from different reference images integrated without conflicts or contradictions?
	D_check_2	Are style, lighting, and background consistent across composed objects?
G. Overall Usability	G_check_1	Does the final scene appear natural, coherent, and visually plausible?
	G_check_2	Are lighting, shadows, and global aesthetics of sufficient quality for practical use?

MICON-Bench Evaluation Prompt

You are a professional multimodal evaluation specialist tasked with assessing AI-generated images against specific reference materials and instructions.

The task type is: {Task Type}.

****IMPORTANT - Image Order**:**
You will be shown multiple images in the following order:
1. Reference images (first N images) - these are the input/reference materials
2. Generated image (LAST image) - this is the candidate image to evaluate

The Original Generation Instruction is: {Image Prompt}.

 # Task Requirements
Hard Constraints:
- H1: Must include ALL target objects from reference images
- H2: Objects must maintain their core visual identity from references
Here are the Evaluation Principles:
1. Binary Assessment: For each checkpoint, output only:
- 1 = PASS (requirement clearly met)
- 0 = FAIL (requirement not met or partially met)
2. Hard Constraint Enforcement:
If a hard constraint fails, mark it as 0 in "hard_constraint_results".

 # Verification Checkpoints
A. Instruction Following: {Checkpoints List}
B. Identity / Fidelity: {Checkpoints List}
C. Structure / Geometry: {Checkpoints List}
D. Cross-Reference Consistency: {Checkpoints List}
E. Causality: {Checkpoints List}
F. Text Grounding: {Checkpoints List}
G. Overall Usability: {Checkpoints List}

[Output JSON Format]
You MUST output JSON in this exact structure:

```
{
  "checkpoint_results": {
    "A_check_1": {"pass": 0 or 1, "hard_id": "H1"},
    "A_check_2": {"pass": 0 or 1},
    ........
  },
  "hard_constraint_results": {
    "H1": 0 or 1,
    "H2": 0 or 1
  },
  "rationale_short": "<<=200 chars>>",
  ]
}
```

Figure 7. The evaluation prompt used for MLLM scoring.

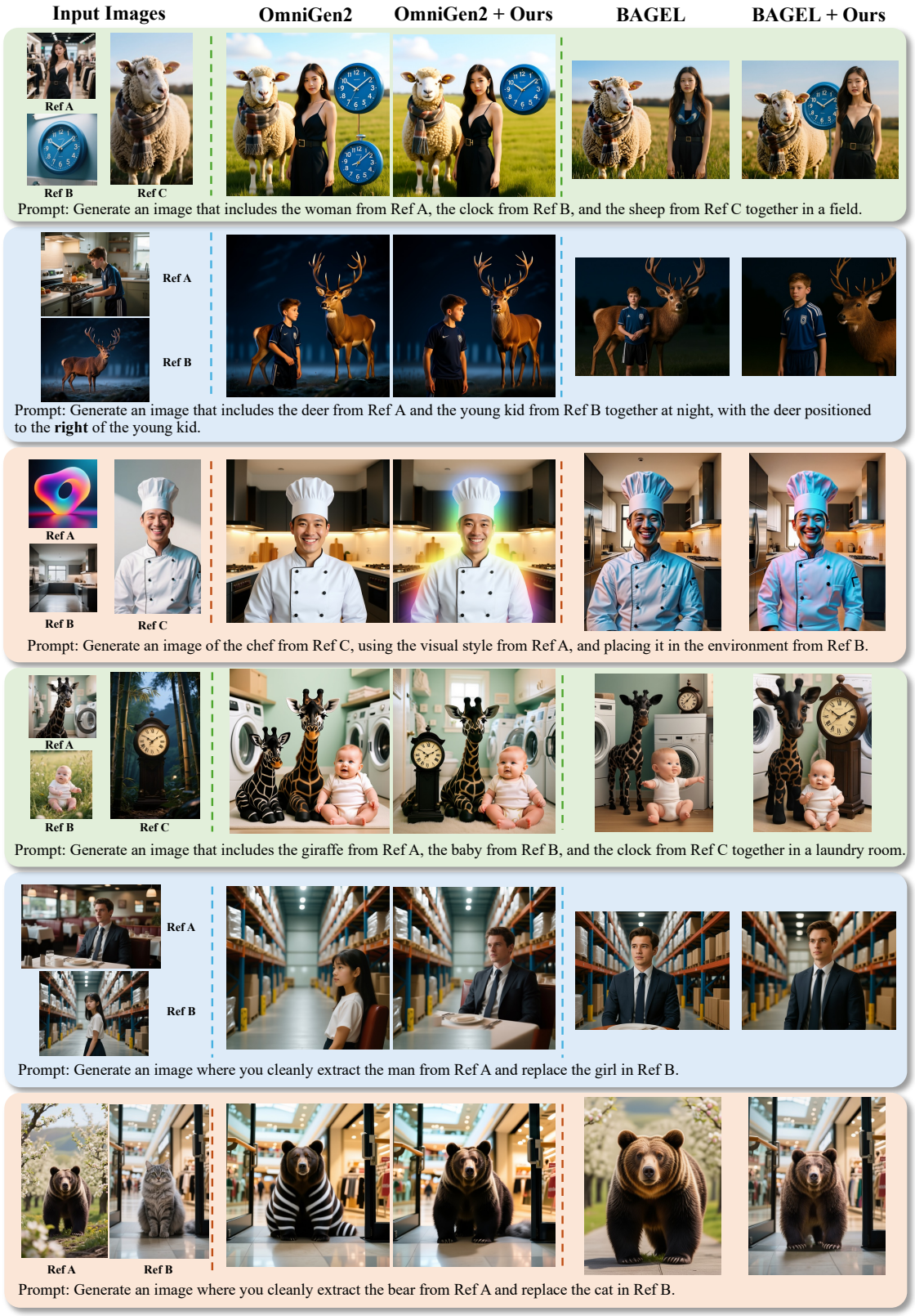


Figure 8. More visualization examples of our method vs. baseline on MICON-Bench.

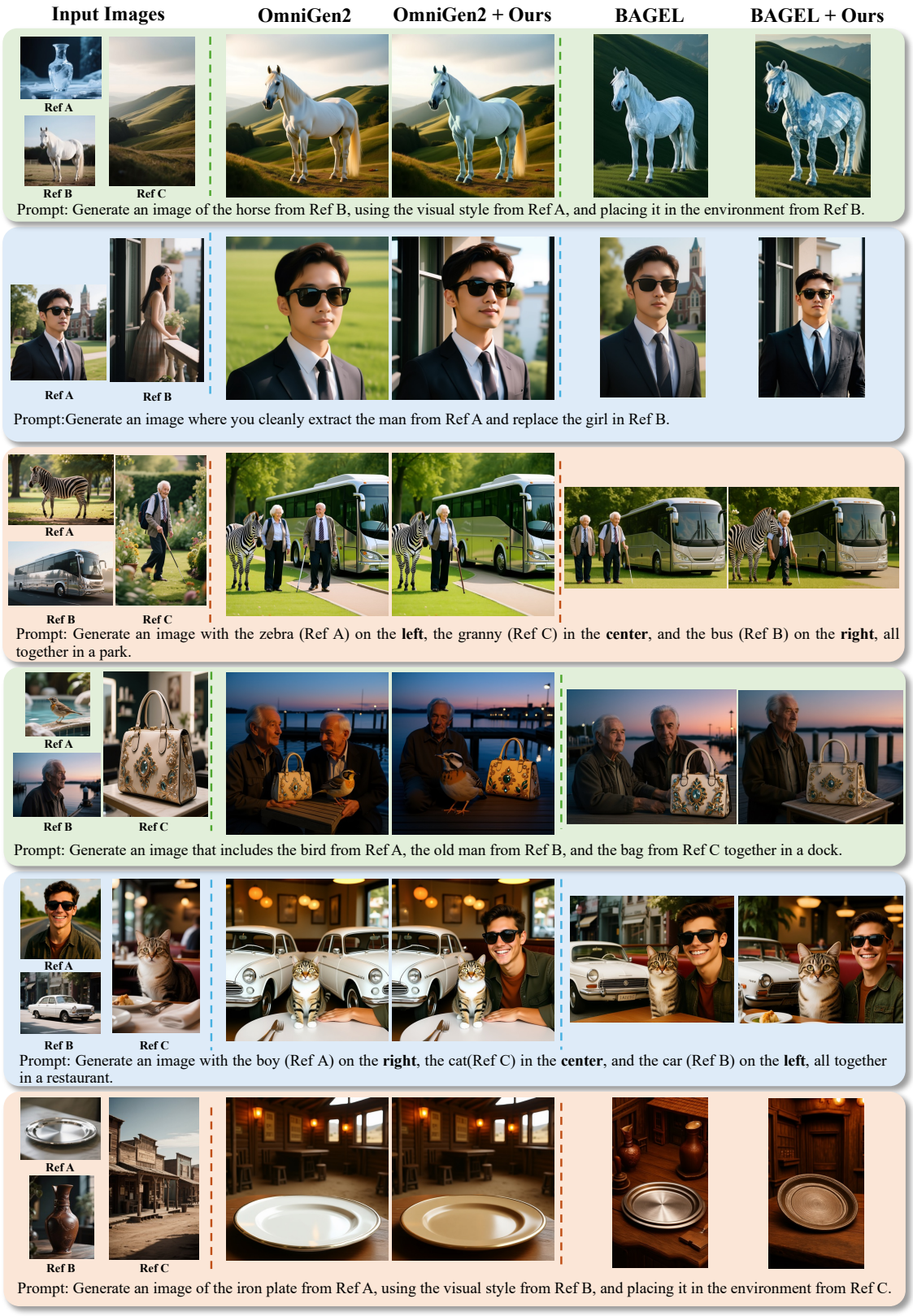


Figure 9. More visualization examples of our method vs. baseline on MICON-Bench.