IDEA-Bench: How Far are Generative Models from Professional Designing?

Supplementary Material

7. Implementation Details

In this section, we detail the methods used for IDEA-Bench construction and experimental anlyses to ensure reproducibility. Sec. 7.1 provides example instructions for utilizing GPT-4o [33] in the construction of IDEA-Bench , while Sec. 7.2 outlines the experimental configurations.

7.1. IDEA-Bench Construction Instruction

Instruction for prompt rephrasing As mentioned in Sec. 4.1.1, to closely align with real design scenarios, IDEA-Bench includes multi-image generation tasks that most existing models do not support. To thoroughly evaluate current generative models' capabilities in these tasks, we utilize one of the most advanced MLLMs, GPT-4o [33], to rephrase multimodal inputs (which may include multiple images and complex long texts) into several text-to-image prompts. The specific rephrasing instruction is illustrated in Fig. 6. However, transforming tasks through rephrasing is merely a workaround, as text alone cannot capture all the details of the given images. Human designers have the ability to autonomously extract information from images and transform it into outputs in a freeform manner. We aim for IDEA-Bench to drive future generative models to acquire this capability.

Instruction for evaluation question construction After collecting the task data, we generate evaluation questions in bulk by combining task keywords provided by human annotators with GPT-4o [33]. Fig. 7 illustrates an example of the instruction for generating evaluation questions for **image(s)-to-images** tasks. In Fig. 7, the red sections indicate prompts that need to be customized for each specific task, while the JSON format templates are omitted. Within the fixed prompts, we first outline the basic requirements for the evaluation questions, such as multi-level standards, the exclusive use of objective judgment questions, and the convention that a score of 1 signifies a better result compared to 0. After incorporating the fundamental task definitions provided by annotators, the prompts also include frequently occurring evaluation capability keywords specific to multi-image generation tasks. This ensures that the evaluation questions defined by GPT-4o maintain a professional standard.

7.2. Inference Configuration

Tab. 9 details the configurations applied during inference for all models. To ensure fairness, all diffusion-based models employ 50 sampling steps (DALL-E 3 [40] utilizes the official API and is therefore excluded from the statistics). Notably, Anole's visual decoder is not diffusion-based [8]; instead, it employs a diffusion-free, token-based architecture. We adhere to the text guidance scale and image guidance scale recommended by the official project codes, as illustrated in Tab. 9.

Table 9. **Inference details of the models being tested.** "–" indicates either an API call or the absence of relevant parameters.

Method	Param.	DiT based	Text Guid. Scale	Image Guid. Scale	Steps
FLUX-1 [25]	12B	✓	3.5	–	50
DALL-E 3 [40]	12B	✗	–	–	–
SD 3 [13]	2B	✓	7.0	–	50
Pixart [7]	0.6B	✓	7.0	–	50
InstructPix2Pix [5]	1B	✗	7.5	1.5	50
MagicBrush [60]	1B	✗	7.5	1.5	50
Emu2 [47]	37B	✗	3.0	–	50
OmniGen [57]	3.8B	✓	3.0	1.6	50
Anole [8]	7B	–	–	–	–

8. Statistical Analysis

Fig. 1 visualizes the distribution of all subtasks across categories. In this section, we further conduct statistical analyses on the composition of the prompts and evaluation criteria of IDEA-Bench.

Distribution of prompt length. In Fig. 4, we present the distribution of prompt lengths across the five task categories using histograms. According to the statistics in Tab. 1, IDEA-Bench's prompts have an average length of approximately 139 words. Prompts shorter than the average are primarily found in the **image-to-image** and **images-to-image** tasks, as these tasks rely heavily on input images to guide the final generation, reducing the need for extensive textual descriptions. However, the prompt lengths for these two categories still significantly exceed those of other benchmarks [9, 17, 20, 23, 43, 45]. Additionally, both **text-to-image** and **image-to-images** tasks feature excessively long prompts, due to the requirements for complex and rich visual elements or detailed descriptions for multiple generated images.

Distribution of evaluation ability. We conduct a statistical analysis of the evaluation dimensions involved in each subtask within every category, with the results illustrated in the figure. In Fig. 5, a higher value for a dimension indicates that the category places greater emphasis on assessing the model's capabilities in that dimension. The analysis reveals that all five categories prioritize the evaluation of aesthetic aspects and the quality of the association between the

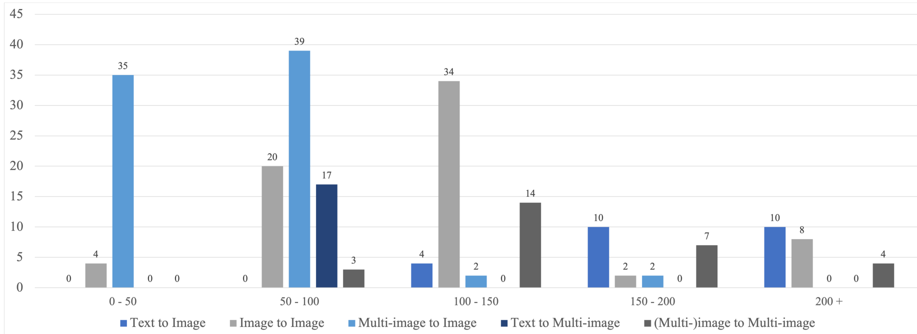


Figure 4. **Statistics of prompt lengths for all tasks in IDEA-Bench.** Each of the five task categories is represented by a distinct color. Prompt lengths are divided into five intervals, and the y-axis shows the number of tasks that fall within each interval.

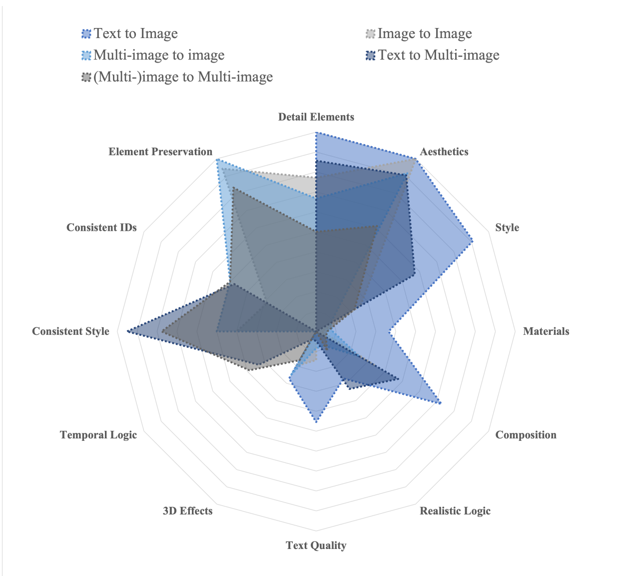


Figure 5. **Statistics of evaluation dimensions for all tasks in IDEA-Bench.** Each of the five task categories is represented by a distinct color. A total of 12 evaluation dimensions are analyzed, with the radar chart values indicating the proportion of evaluation questions related to each dimension within each category.

generated images and the details in the prompts. Specifically, **text-to-image** tasks emphasize assessments of style, image composition, and text quality. In contrast, **image-to-image** and **images-to-image** tasks focus on evaluating the retention of elements between the input and output images. Meanwhile, **text-to-images** and **image(s)-to-images** tasks, which involve generating multiple images, concentrate on evaluating dimensions such as ID consistency and style consistency among the generated images.

9. Additional Experiments

Supplementary results on image(s)-to-images Due to space constraints, we do not include all experimental results for the **image(s)-to-images** category in Tab. 7. Supplementary results are provided in Tab. 10. The current abilities of all models to achieve inter-image associations like ID consistency and style consistency stem from GPT-4o's [33] detailed rephrasing of each prompt, akin to the group image descriptions in GDT [22]. However, GDT employs a design where image tokens are concatenated during attention computation, whereas solely using MLLM rephrasing does not facilitate inter-image association modeling in the latent space. In the future, to enable multi-image generation tasks with complex associations, models will need to consider parallel generation of multiple images or utilize partially generated images as input conditions to guide the generation of subsequent images.

Selecting the Evaluation Model On a subset of the dataset, we select Gemini 1.5 Pro [49] to score the images generated by models based on the refined evaluation questions. However, MLLMs produce free-form textual outputs, making it challenging to ensure binary scores of 0 or 1 as human annotators do, potentially resulting in scoring failures. In Tab. 11, we report the failure rates of Gemini 1.5 pro [49] and GPT-4o [33], representing the proportion of evaluation questions where scoring failed. Specifically, we conduct three evaluations per question. If all three attempts do not yield a clear score, the evaluation is considered a failure. Across all models and evaluation questions, Gemini 1.5 pro exhibits a low failure rate of 0.95%, whereas GPT-4o shows a high failure rate of 52.84%, rendering it unsuitable as a reliable automated evaluation model. In practice, GPT-4o frequently responds with phrases such as *"I'm sorry, I can't assist with that"*, whereas Gemini 1.5 pro provides more consistent responses. Additionally, Tab. 12 compares the evaluation results of Gemini 1.5 Pro and Gemini 1.5 Flash, showing that the Pro model exhibits higher consistency with human annotations.

Table 10. **Experimental results on Image(s)-to-Images.** Each task category is averaged across all its subtasks, with the top-ranked model scores for each task type highlighted in **bold**. Task types that a model cannot support are marked with "–". "†" represents the use of MLLM for prompt rephrasing.

Method	Params	Subtasks Score							Avg. Score
		Paint. Undo	Same Pose	Three-view Trans.	Child. Book	Plant Growth	Prod. Usage Scen.	Stop-motion Anim.	
FLUX-1† [25]		0.00	0.00	0.00	**45.83**	41.67	**33.33**	**25.00**	**29.17**
DALL-E 3† [40]		0.00	0.00	0.00	37.50	**58.30**	16.67	16.67	14.44
Stable Diffusion 3† [14]		0.00	**25.00**	0.00	29.17	16.67	16.67	16.67	13.06
Pixart† [7]		0.00	8.33	0.00	37.50	41.67	16.67	16.67	21.39
InstructPix2Pix [5]		–	–	–	–	–	–	–	–
MagicBrush [60]		–	–	–	–	–	–	–	–
Anole [8]		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Emu2 [47]		–	–	–	–	–	–	–	–
OmniGen [57]		–	–	–	–	–	–	–	–

Table 11. **Comparison of evaluation failure rates among different MLLMs.** For each evaluation question, MLLMs score the model-generated outputs three times. If none of the three scores return the required value (0 or 1), the evaluation is considered a failure.

Eval. MLLM	Method								Total
	FLUX-1	DALL-E 3	SD3	Pixart	InstructPix2Pix	MagicBrush	Emu2	OmniGen	
Gemini 1.5 pro [49]	0.33%	1.63%	1.96%	0.00%	0.00%	0.00%	0.67%	1.33%	0.95%
GPT-4o [33]	52.29%	54.58%	52.95%	56.86%	16.67%	26.67%	54.00%	53.33%	52.84%

Table 12. **Correlation between human and automated evaluation using different MLLMs.** "S." and "P." represents the Spearman coefficient and the Pearson coefficient, respectively.

MLLM	T2I		I2I		Is2I		T2Is		I(s)2Is	
	S.	P.	S.	P.	S.	P.	S.	P.	S.	P.
Gemini-1.5-flash	0.95	0.97	-0.42	-0.29	**0.60**	**0.43**	0.91	0.99	0.92	0.97
Gemini-1.5-pro	**0.97**	**0.99**	**0.18**	**0.20**	0.15	-0.06	**0.99**	**1.00**	**0.95**	**0.98**

Quantifying the impact of MLLM assistance To further quantify the impact of MLLM [33] assistance on model capabilities within the benchmark, we generate multi-image tasks that not supported by Emu2 [47] and OmniGen [57] using GPT-4o-rephrased prompts. Despite enhancing Emu2 and OmniGen's capabilities with MLLMs, FLUX-1 [25] remains the top performer, as shown in Tab. 13. Notably, FLUX-1 even surpasses other models that support image input in **image-to-image** tasks. The advantage of T2I models [7, 13, 25, 40] in this experiment lies in their ability to leverage MLLMs to understand different tasks. IDEA-Bench's task definitions are highly specialized, making it difficult for other models to comprehend these tasks without relying on MLLM. Universal generative models also have difficult ensuring the quality of generated images, resulting in lower scores finally. Overall, to achieve high scores across all benchmark tasks, a model must possess both multimodal input-output capabilities and robust MLLM-level multimodal understanding.

Comparison of T2I capabilities across all models We also apply prompt rephrasing to all models in text-to-image generation settings. In this setup, all models have unified input comprehension capabilities, evaluating whether they can accurately translate prompts into high-quality generated images. The results are included in Tab. 13, featuring Emu2 [47], OmniGen [57], and Anole [8], each distinguished by the "-T2I" suffix. FLUX-1 [25] remains the top-ranked model. FLUX-1 demonstrates a strong ability to convert prompts into images, maintaining stable image quality with only rare instances of failure.

Table 13. **Additional experimental results on all categories of IDEA-Bench.** "†" represents the use of MLLM for prompt rephrasing.

Method	Scores on All Categories					Avg. Score
	T2I	I2I	Is2I	T2Is	I(s)2Is	
FLUX-1† [25]	**46.06**	**12.13**	4.79	20.15	**29.17**	**22.46**
DALL-E 3† [40]	24.34	6.95	5.22	14.36	14.44	13.06
Stable Diffusion 3† [13]	24.04	10.79	4.57	21.59	13.06	14.81
Pixart† [7]	14.04	7.75	3.18	17.46	21.39	12.76
Anole† [8]	0.00	0.64	0.00	1.74	0.00	0.48
Emu2† [47]	17.98	7.05	**8.94**	15.53	12.78	12.46
OmniGen† [57]	21.41	8.17	2.11	**23.52**	21.39	14.32
Anole-T2I† [8]	0.00	3.10	1.17	8.98	8.89	4.43
Emu2-T2I† [47]	17.98	3.15	2.15	15.53	12.78	10.32
OmniGen-T2I† [57]	21.41	6.09	4.35	**23.52**	21.39	15.35

10. Data Examples

Fig. 8 - Fig. 12 display additional model-generated results, including the input images and text prompts used. Some text prompts are omitted due to their length. Since different models support a limited number of task categories, we only showcase the models that are capable of handling each respective category in the generation results.

Fig. 13 - Fig. 15 illustrate examples of automated evaluations conducted using Gemini 1.5 pro [49]. Due to the detailed definitions of the generation prompts and evaluation questions, the evaluation process can be effectively transformed into a multimodal understanding task, which MLLM excels at. In both presented examples, the model-generated results fail to fully meet the prompt requirements, resulting in a score of 0.

This is a case from the image generation test set. Please determine how many images in total need to be generated and write a text prompt for each output image in English. Since the task requires a certain degree of correlation between the final output images, this correlation can only be achieved by describing the image details as thoroughly as possible. For example, the original use case might describe generating 4 images in a specific style, and in each generation prompt, you need to re-define the style, as the model cannot see the other generated images. Each image prompt is independent, and you should not include descriptions like “consistent with xxx”.

Here is the original case:

Please generate a series of images based on the following text prompt, showcasing the use of this speaker in different application scenarios. A total of four images need to be generated, each representing a different scene. The first image shows the speaker on a modern-style office desk indoors, with a computer in the background, sunlight streaming through the window illuminating the desk, and the black speaker contrasting with the white and wooden elements of the desk. The second image shows the speaker on the armrest of a soft sofa, with fine fabric textures on the sofa, a warm table lamp nearby, creating a cozy and relaxed atmosphere. The third image shows the speaker outdoors on a tree stump, with sunlight filtering through the leaves in the background, a backpack placed next to the speaker, creating a natural and lively atmosphere. The fourth image shows the speaker placed on an outdoor wooden table, with a smartphone next to it playing music, autumn leaves scattered on the table, blending nature with technology.

Output in JSON format. [omitted here]

{
“prompt_1”: {
Showcase a modern-style office desk indoors. Place a sleek black speaker prominently on the desk, with a computer visible in the background. Sunlight should stream through a nearby window, creating a bright and inviting atmosphere that highlights the contrast between the black speaker and the white and wooden elements of the desk.
}
“prompt_2”: {
Illustrate a soft sofa with fine fabric textures, featuring the black speaker resting on the armrest. Nearby, a warm table lamp casts a gentle glow, contributing to a cozy and relaxed ambiance. The scene should evoke comfort and leisure.
}
“prompt_3”: {
Depict a natural outdoor scene featuring a tree stump as a pedestal for the black speaker. Sunlight filters through the leaves above, creating dappled light on the ground. A backpack is placed next to the speaker, emphasizing a lively and adventurous atmosphere in a serene natural setting.
}
“prompt_4”: {
Present an outdoor wooden table adorned with autumn leaves scattered across its surface. The black speaker sits on the table alongside a smartphone that is actively playing music. The scene should blend elements of nature and technology, capturing a harmonious outdoor moment.
}
}

Figure 6. **An instruction example for prompt rephrasing.** The red sections indicate customization for different subtasks. The JSON file format templates within the instructions are not shown.

11. Limitations & Future Work

Due to the current capabilities of multimodal large language models (MLLMs) still falling short of human performance, we are unable to apply automated MLLM evaluations to all tasks while meeting the evaluation standards of professional designers. Furthermore, the primary goal of IDEA-Bench is to bridge the gap between current generative models and professional tasks, pushing model capabilities toward a professional level. However, there remains a significant distance to match the proficiency of professional designers. In the future, we will focus on updating and maintaining IDEA-Bench, continuously refining automated evaluation methods in line with the real-time advancements of MLLMs, and expanding to more specialized tasks. This will ensure that the benchmark effectively supports the ongoing evolution of generative model capabilities.

You are a professor at a design school, and the following is an exam question. You need to design 6 evaluation questions for a task where the model generates multiple images based on a text description and multiple referencing images to assess the output. The questions should follow these two principles:

a. There should be a total of 6 questions, and each question must be a clear, binary judgment (0 points or 1 point). Each question should first specify the evaluation aspect and details, followed by clear criteria for scoring 0 points and 1 point. A score of 1 point represents a better outcome, and 0 points represent a worse outcome.

b. The first two questions should focus on whether the model output meets the basic definition of the task. The next two questions should further evaluate whether the model output follows specific instructions from the text input. For example, if the text input specifies making the skin smoother, the model needs to clearly fulfill that directive. The last two questions should evaluate the images from a high-level professional perspective, assessing aspects like image detail, aesthetics, and so on.

Please design 6 evaluation questions for the "multi-appearance variant generation" task. This task is to generate multiple appearance color renderings of an object. Ensure the questions are created from a professional and comprehensive perspective, considering the following aspects:

· If the task involves partial modification, does the rest of the image remain unchanged?
· The correlation between input and output images (content, style, ID, etc.)
· Does it meet the requirements of the text description?
· The quality of text editing within the image.
· Chronological logic.
· Consistency of image style.
· Consistency of image ID (recognizable as the same person or object).
· Logical consistency.
· Alignment with the text description.
· Is the output derived from the input image, or is it entirely unrelated?

Don't limit yourself to these aspects, and also consider whether additional criteria are relevant to this task. Please provide the 6 evaluation questions you have defined. Output in JSON format. [omitted here]

{
"questions": [
{
"question": Is the output image clearly derived from the input image, maintaining the core structure and identity of the product?
"0_point_standard": The output image shows significant deviation from the input image, making the product unrecognizable or altering its core structure.
"1_point_standard": The output image maintains the core structure and identity of the product, ensuring it is recognizable as the same item.
}
{
"question": For partial color modifications, does the model ensure that non-targeted areas of the product remain unchanged?
"0_point_standard": Unintended areas of the product are altered, resulting in unnecessary changes in parts of the image that should remain constant.
"1_point_standard": Non-targeted areas of the product remain unchanged, preserving the initial design and details of the image.
}
{
"question": Does the model output strictly adhere to the text description's requirements for specific color variations?
"0_point_standard": The color variations do not match the specifications provided in the text description, with colors appearing inaccurately or not as described.
"1_point_standard": The color variations in the output images are accurate and align precisely with the text description's requirements. }
{
"question": Is the logical consistency of the product maintained across different color variants, ensuring realistic and feasible color application?
"0_point_standard": The color changes result in unrealistic or illogical appearances, such as unnatural shading or color bleeding that affects realism.
"1_point_standard": The color application is logical and realistic, maintaining feasible and natural appearances across all variants.
}
{
"question": Do the generated color variants maintain a consistent image style, ensuring uniformity in lighting, texture, and presentation?
"0_point_standard": There is inconsistency in style across the color variants, with varied lighting, texture, or presentation that disrupts uniformity.
"1_point_standard": The image style is consistent across all color variants, with uniform lighting, texture, and presentation.
}
{
"question": Is the overall aesthetic and visual quality of each color variant maintained, ensuring high image quality and appeal?
"0_point_standard": The color variants suffer from poor image quality, with issues such as pixelation, loss of detail, or lack of visual appeal.
"1_point_standard": Each color variant maintains high image quality, with clear details and an aesthetically pleasing appearance.
}]
}

Figure 7. **An instruction example for generating evaluation questions.** The red sections indicate customization for different subtasks. The JSON file format templates within the instructions are not shown.

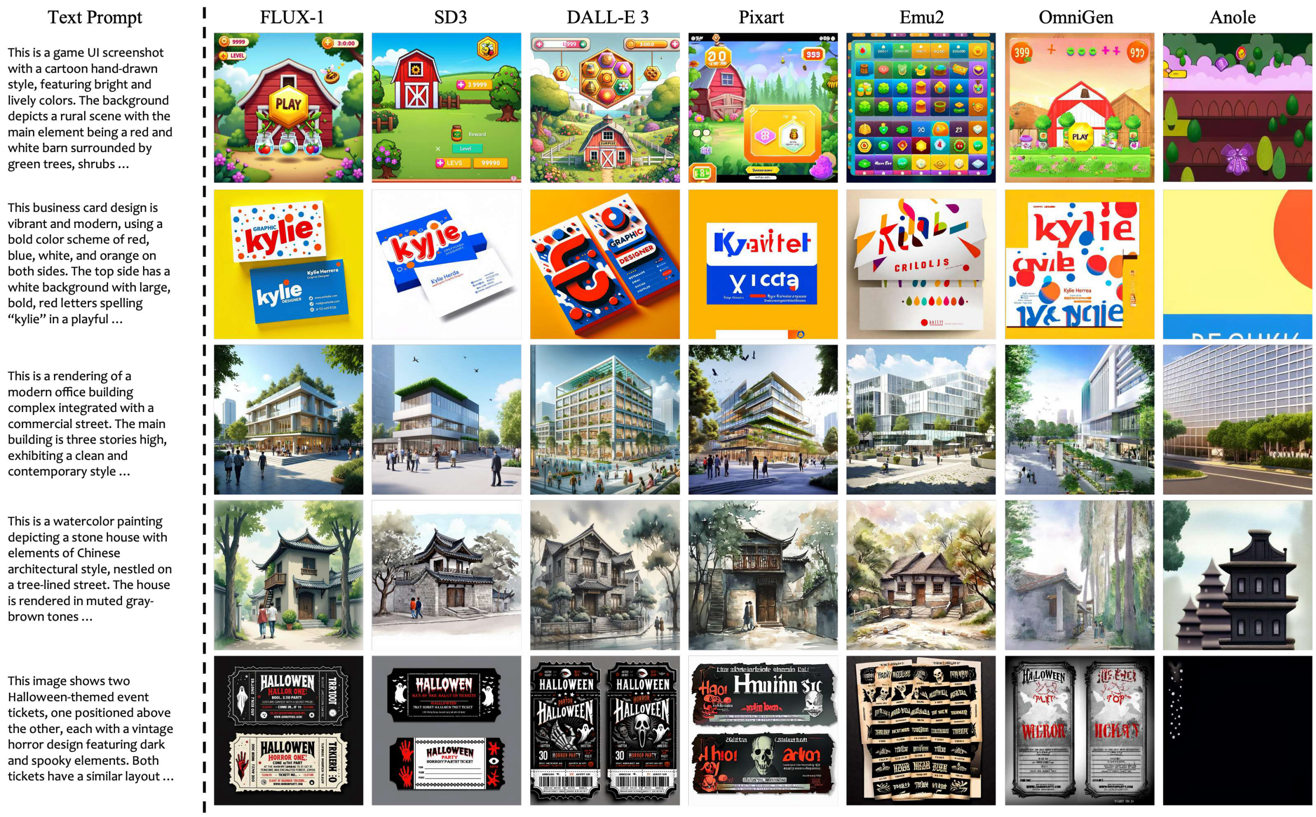


Figure 8. **Generations for selected cases in the text-to-image category.** The displayed task categories, from top to bottom, include *game UI generation*, *business card generation*, *architectural style generation*, *painting generation*, and *ticket generation*.

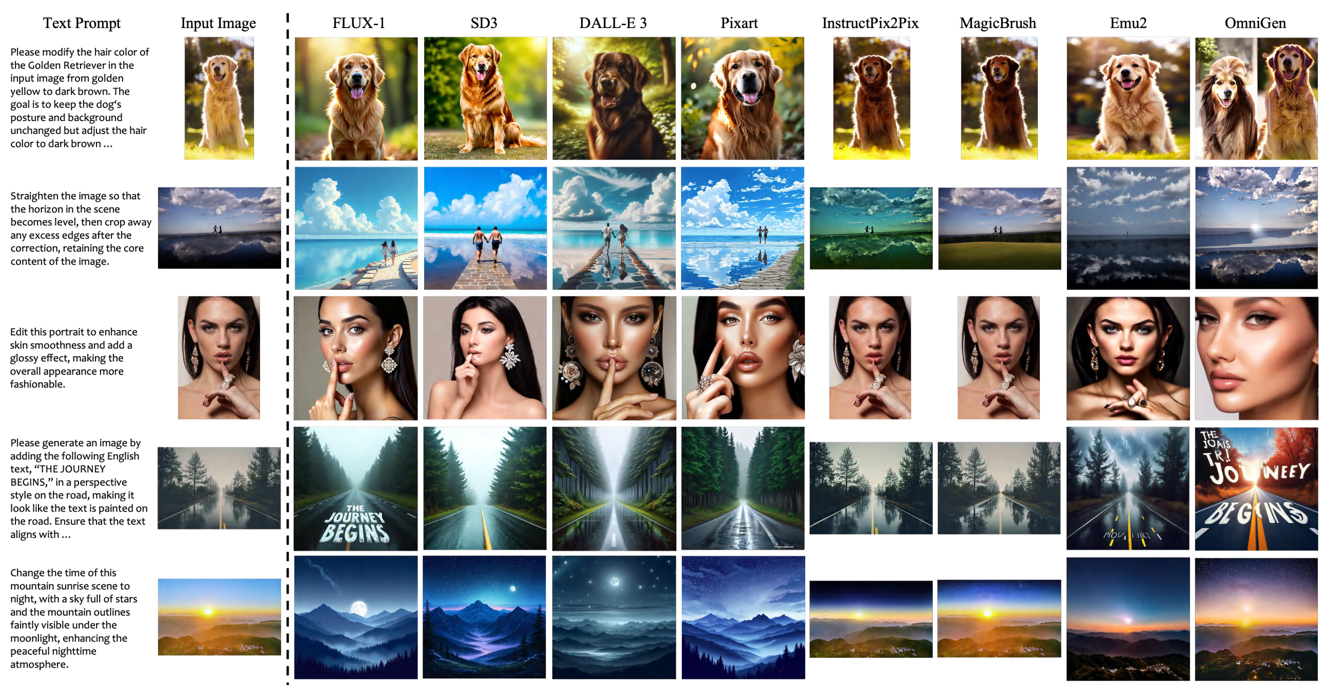


Figure 9. **Generations for selected cases in the image-to-image category.** The displayed task categories, from top to bottom, include *animal hair editing*, *image straighten*, *image retouching*, *text insertion*, and *time editing*.

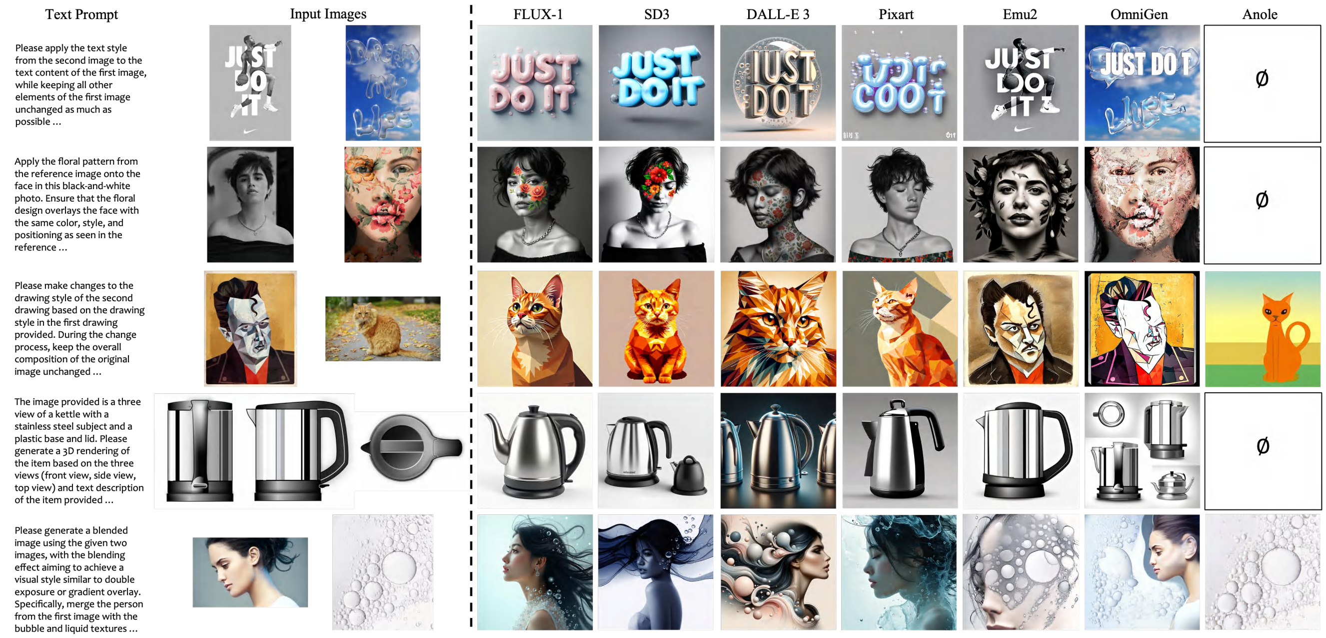
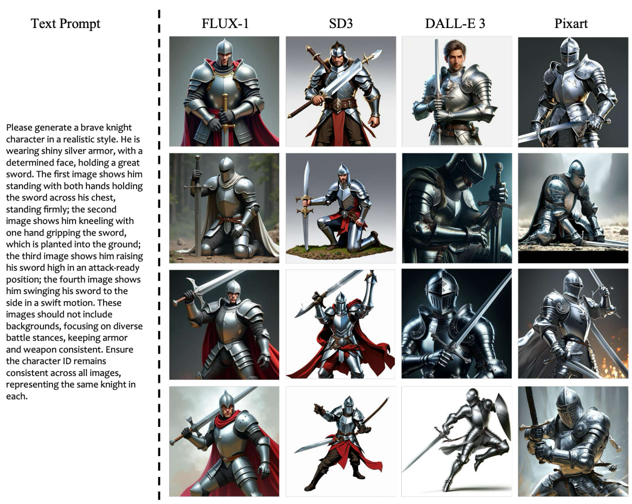
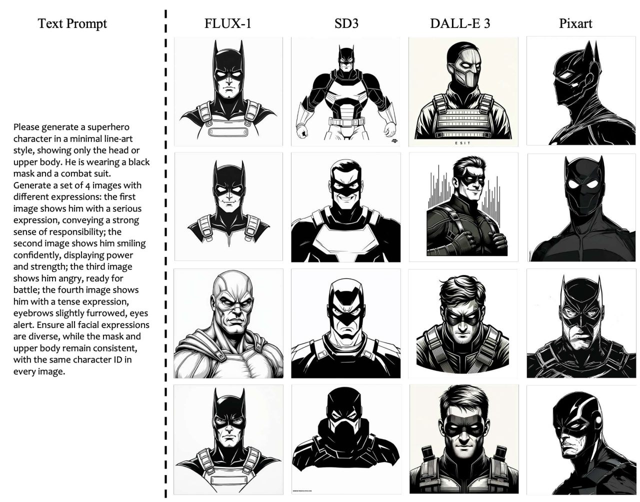


Figure 10. **Generations for selected cases in the multi-image-to-image category.** The displayed task categories, from top to bottom, include *text style transfer*, *body painting transfer*, *art style transfer*, *3D rendering*, and *double explosure*.



(a) Generated images for *character action design*



(b) Generated images for *character expression design*

Figure 11. **Generations for selected cases in the text-to-multi-image category.**

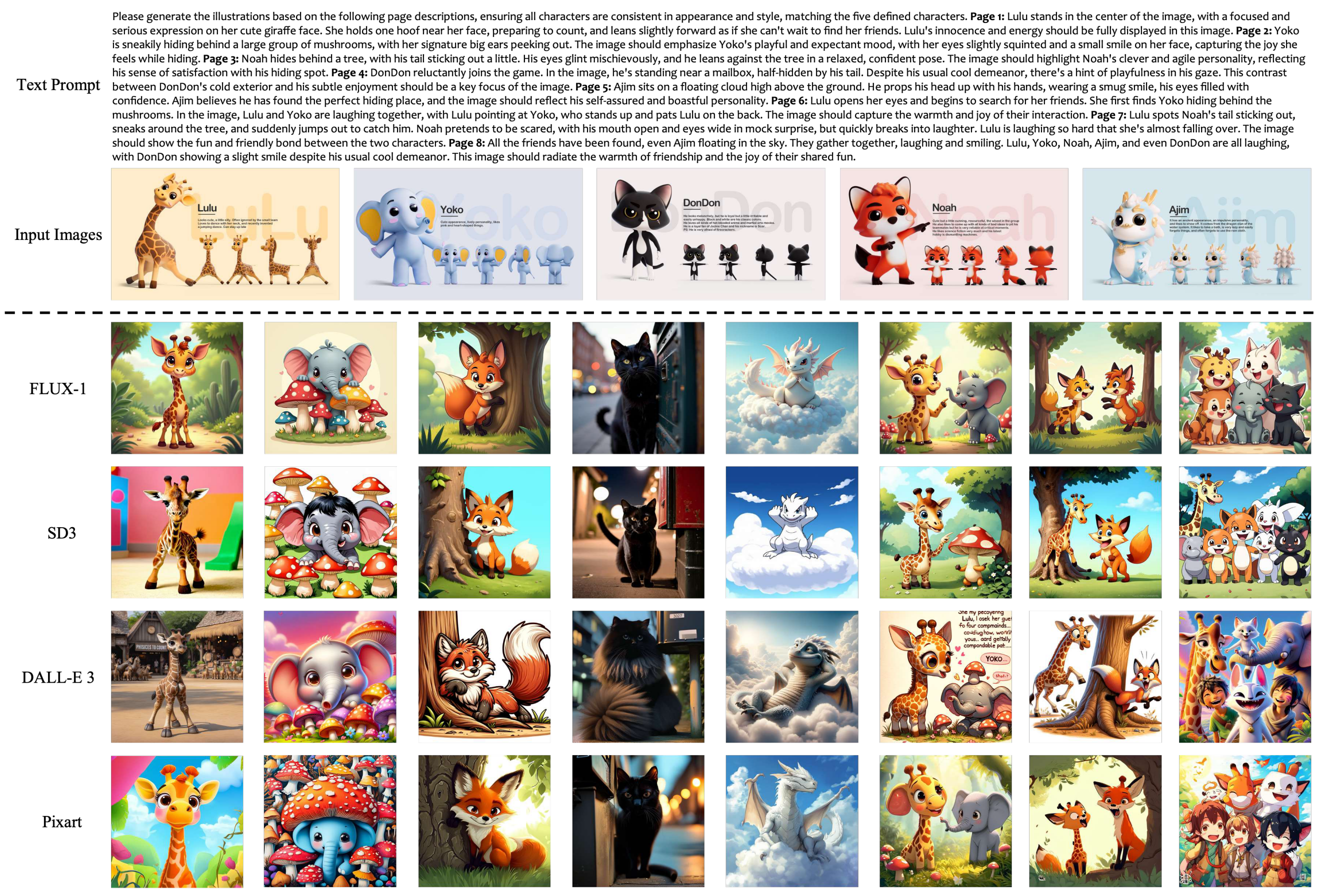


Figure 12. **Generations for the task of *children's storybook generation*.** The dashed line above represents the model's input text prompts and role definition images, while the dashed line below illustrates the prompt generation results after rephrasing by GPT-4o [33] for four models.

You are a professional image designer, and you are now required to conduct a strict evaluation of the following design work. The work consists of only one image as the response provided by a student. The task objective is to generate a poster based on the text requirements.

The text requirement is:
This poster is designed in a vintage racing theme featuring vibrant colors and bold graphics. At the top, the words "PISTON CUP" are written in large, black, retro-styled letters, with a trophy icon replacing the letter "O" to emphasize the racing theme. Below it, in smaller italicized letters, reads "RACING SERIES." The center of the poster highlights three racing cars in dynamic motion. The red car, "Lightning McQueen" with the number "95" prominently displayed on its side, takes the forefront, showcasing a smile and intense speed lines that enhance its fast movement. A turquoise car labeled "DINOC" races alongside, while a black car marked "The King" trails slightly behind. The background has stylized palm trees and an orange sky, adding a California racing vibe. On the left side, a "Goodyear" blimp floats above a checkered flag, reinforcing the racing atmosphere. At the bottom, a red flame background displays the name "LIGHTNING MCQUEEN" in bold yellow letters, along with an illustration of additional characters from the "Cars" series. The "CARS" logo appears in bold letters at the very bottom, along with Disney and Pixar logos.

Your review question is:
Are the text elements (e.g., title, tagline, body text) in the poster clear, readable, and appropriately placed to convey the intended message?
- 0 points: The text elements are unclear, difficult to read, or poorly positioned, affecting the communication of the message.
- 1 point: The text elements are clear, readable, and well-placed, effectively conveying the intended message.

Use this JSON schema:
Evaluation = {'score': int, 'reason': str} Return: Evaluation

```
{
  "score": 0,
  "reason": "While the "RACING CUP" title is large and readable, it lacks the specified retro styling and trophy icon replacing the 'O'. "RACING SERIES" is present but not italicized. The text placements and fonts are not as described, and key elements like "Lightning McQueen", the "Cars" logo, flame background, checkered flag, blimp, and Disney/Pixar logos are missing. The misinterpretations significantly impact the communication of the intended message. "
}
```

Figure 13. **Automated evaluation of *poster generation*.** The red font represents content that changes with each task or evaluation question.

You are a professional image designer, and you are now required to conduct a strict evaluation of the following design work. The work consists of two images, with the left image as the reference picture for the design task and the right image as the response provided by a student. The task objective is to generate a realistic 3D rendering based on the provided design sketch and text requirements.

The text requirement is:
Please generate a 3D rendering based on the provided interior design sketch. The task objective is to accurately convert the room layout, furniture, and decorative elements in the image into a three-dimensional rendering. The model should infer the appropriate spatial structure of the room, the proportions and placement of the furniture, and render realistic lighting, shadows, and material effects. Ensure that the generated 3D rendering aligns with the design details in the sketch, showcasing the overall spatial feel and visual impact of the bedroom, ultimately producing a high-quality, realistic 3D rendering.

Your review question is:
Does the generated 3D rendering maintain the overall structure and proportions of the line drawing, ensuring consistency between the line drawing and the generated image?
· 0 points: The object's structure in the 3D rendering has been noticeably altered, with unbalanced proportions.
· 1 point: The structure and proportions of the object in the 3D rendering are consistent with the line drawing and are well-balanced.

Use this JSON schema:
Evaluation = {'score': int, 'reason': str} Return: Evaluation

{
"score": 0,
"reason": "While the 3D rendering captures the general layout and furniture types from the sketch, there are noticeable deviations in the specific shapes and outlines. For example: **Bed:** The headboard in the 3D model is a simple rectangular shape, whereas the sketch shows a more stylized design. **Bench:** The bench at the foot of the bed in the 3D model has angled legs, unlike the straight legs in the sketch. Also the bench in the 3D render is shorter and thicker. **Chair:** The chair in the 3D rendering is a different style compared to the foldable sling chair in the sketch. **Closet:** The closet in the 3D model has solid doors, while the sketch depicts sliding doors. **Light Fixture:** The hanging lamp in the sketch is not present in the 3D rendering. A simple table lamp is used instead. **Rug:** The rug's shape and texture differ significantly between the sketch and the 3D model. **Wall Decor/Decorative Elements:** The decorative branch depicted in the sketch is missing from the 3D render. The nightstand is a different design, and the alarm clock and small vase are absent. The curtains are also missing. These discrepancies demonstrate that the 3D rendering doesn't accurately retain every detail of the shapes and outlines from the original line drawing. Though the overall 'feel' is similar, the task was to be a faithful recreation of the sketch's design. "
}

Figure 14. **Automated evaluation of *3D rendering generation*.** The red font represents content that changes with each task or evaluation question.

You are a professional image designer, and you are now required to conduct a strict evaluation of the following design work. This is the third and fourth output images of the response provided by a student. The task objective is to generate different application scenario images of a specific product.

The text requirement is:
Please generate a series of images based on the following text prompt, showcasing the use of this speaker in different application scenarios. A total of four images need to be generated, each representing a different scene. The first image shows the speaker on a modern-style office desk indoors, with a computer in the background, sunlight streaming through the window illuminating the desk, and the black speaker contrasting with the white and wooden elements of the desk. The second image shows the speaker on the armrest of a soft sofa, with fine fabric textures on the sofa, a warm table lamp nearby, creating a cozy and relaxed atmosphere. The third image shows the speaker outdoors on a tree stump, with sunlight filtering through the leaves in the background, a backpack placed next to the speaker, creating a natural and lively atmosphere. The fourth image shows the speaker placed on an outdoor wooden table, with a smartphone next to it playing music, autumn leaves scattered on the table, blending nature with technology.

Your review question is:
Do the third and fourth output images convey a clear sense of the speaker being used in suitable environments (e.g., outdoor and nature-friendly setups), as intended in the prompt?
· 0 points: The context does not convincingly represent suitable use environments, making the scenes feel out of place.
· 1 point: The scenes appropriately match the outdoor and natural settings, making the speaker's use feel relevant and purposeful in the context.

Use this JSON schema:
Evaluation = {'score': int, 'reason': str} Return: Evaluation

{
"score": 1,
"reason": "The third image effectively showcases the speaker on a tree stump, creating a plausible outdoor setting. The inclusion of a few scattered leaves strengthens the natural ambiance and suggests a hiking or camping scenario where such a speaker might be used. The fourth image, however, is slightly less convincing, although acceptable. While the wooden table and scattered autumn leaves evoke an outdoor setting, the scene feels somewhat staged and lacks the depth and context of a genuine outdoor use case. It doesn't clearly depict an activity or purpose that would naturally involve a speaker in this location. For instance, showing a picnic blanket, a book, or other outdoor items would enhance the scene's realism and purpose. Despite this minor weakness in the fourth image, both images generally succeed in portraying suitable natural and outdoor environments for the speaker."
}

Figure 15. **Automated evaluation of *product usage scenario generation*.** The red font represents content that changes with each task or evaluation question.