Supplementary Material for Human Preference-Aligned Concept Customization Benchmark via Decomposed Evaluation

Reina Ishikawa[1], Ryo Fujii[1], Hideo Saito[1], Ryo Hachiuma[2]
[1]Keio University, [2]NVIDIA
{reina.ishikawa, ryo.fujii0112, hs}@keio.jp, rhachiuma@nvidia.com

A. Existing Benchmark for Concept Customization

Table 1 summarizes the comparison of the composition of existing benchmark datasets used for concept customization evaluation.

B. Dataset Details

B.1. Dataset License

We released our benchmark dataset, CC-AlignBench, and related materials under the CC BY-NC 4.0.

B.2. Text generation details

In Table 2, we list all the action types used in our dataset. The action types are consistent between the *Easy* and *Medium* levels. A portion of the action types used in the dataset are adapted from [18, 27].

The surroundings specified in the prompts are drawn from 20 distinct types, which are reused across different prompts.

B.3. Image generation details

We generated images using the text prompts listed in Table 3. Regarding the alignment between the text prompts and the resulting images, strict adherence was not prioritized. Instead, emphasis was placed on introducing variation, with visual inspection confirming that the subject's appearance remained consistent with the original image. Figure 1 presents ten samples generated for each concept.

Figure 1. **Sample images in CC-AlignBench**

C. Approach to Defining Evaluation Aspects

In determining the 18 evaluation aspects, we adopted a bottom-up categorization approach. This study aims to achieve human-aligned evaluation based on the premise that humans can comprehensively perceive all relevant factors. Accordingly, we first exhaustively identified potential perspectives for evaluating concept customization images. Considering the cost of using Multimodal Large Language Model (MLLM), we then minimized the number of aspects while maintaining appropriate evaluation granularity by removing redundant items and grouping related ones. The overall categorization was inspired by the evaluation aspects proposed in prior studies [8, 33].

D. Prompting of Aspect-Wise Evaluation

To obtain the aspect-wise scores, we provide the MLLM model (*i.e.*, the GPT model) with the generated image I_g, the text prompt $\mathcal{T}$, and the reference images $\mathcal{I}$ as inputs. Table 4 presents the prompting template for the MLLM model, while the prompts that vary depending on the evaluation aspect are detailed in Table 5.

It has been empirically demonstrated that by eliminating unnecessary inputs, MLLM can provide more accurate evaluations and reduce the associated costs. Therefore, we evaluated the necessity of the text prompt and reference images for each evaluation aspect. For instance, when evaluat-

Table 1. Comparison of existing benchmark datasets in terms of inclusion of essential components for concept customization evaluation and applicability to multi-concept customization tasks.

Dataset	Data Type		Multi-Concept Support	
	Images	text-prompt	Multiple Human	Mutual Interaction
MS COCO [17]	(group)	✓	✓	✓
Flickr30K [31]	(group)	✓	✓	✓
FFHQ [10]	✓	✓	✗	✗
UniDet [34]	✓	✗	✗	✗
LION-400M [25]	(group)	✓	✓	✗
DrawBench [24]	✗	✓	✗	✗
SR_{2D}Dataset VISOR [6]	✗	✓	✓	✗
HRS-Bench [1]	✗	✓	✓	✗
DreamBooth [23]	✓	✗	✗	✗
TIFA v1.0 [9]	✗	✓	✓	✗
DALLEval [5]	(group)	✓	✓	✗
CustomConcept101 [14]	✓	✓	✓	✗
ImagenHub [13]	✓	✓	✗	✗
DreamBench [22]	✓	✓	✗	✗

Table 2. Action types used in Human Preference-Aligned Concept Customization Benchmark (CC-AlignBench). *Easy* and *Medium* consist of the same 13 individual actions; however, in Easy, only a single person (A*) performs the actions, whereas in Medium, two persons (A* and B*) perform the same actions. *Hard* consists of 23 mutual actions.

Individual actions (*Easy / Medium*)	Mutual actions (*Hard*)
A* (and B*) standing	A* punching B*
A* (and B*) walking	A* kicking B*
A* (and B*) running	A* pushing B*
A* (and B*) waving one's hand	A* patting B* on the back
A* (and B*) clapping	A* pointing finger at B*
A* (and B*) putting one's hands in one's pockets	A* hugging B*
A* (and B*) jumping up	A* giving a book to B*
A* (and B*) checking the time on one's wristwatch	A* touching B*'s pocket
A* (and B*) crossing one's hands in front of one's chest	A* and B* shaking hands
A* (and B*) kneeling on the ground	A* hitting B* with a book
A* (and B*) squatting down	A* putting his arm around B*'s shoulder
A* (and B*) punching	A* knocking into B*
A* (and B*) shrugging one's shoulders	A* grabbing a book from B*
	A* stepping on B*'s foot
	A* and B* giving each other a high-five
	A* and B* clinking glasses
	A* and B* carrying a box together
	A* taking a picture of B* with a camera
	A* following B* down a street
	A* whispering into B*'s ear
	A* and B* exchanging a book
	A* supporting B* as they walk
	A* and B* playing rock-paper-scissors

ing aspects that only compare the consistency between the generated image and the text prompt, reference images are not provided to the MLLM. Table 5 indicates which input—text prompt, reference images, or neither—is used for each evaluation aspect, and Table 4 highlights the template differences between various input combinations.

E. Prompting of Vanilla-GPT

As an ablation study to evaluate the effect of decomposition, we conducted an experiment in which images were rated on a scale from 1 to 10 by directly prompting GPT-4o without applying decomposition. The prompting template used for GPT-4o in this experiment is provided in Table 6.

Table 3. Prompt templates used for image generation in our dataset

Prompting template for the base image generation	

A Woman	`Generate a high-quality, photo-realistic, full-body image (including the legs) of a woman as described below:`
	`- Hairstyle: long blonde hair`
	`- skin tone: white`
	`- Age: around 25 years old`
	`- Face: smiling at the camera`
	`- Outfit: casual wearing necklace`
	`- Output a square image`

A Man	`Generate a high-quality, photo-realistic, full-body image (including the legs) of a man as described below:`
	`- Hairstyle: black hair`
	`- skin tone: brown`
	`- Age: around 25 years old`
	`- Face: smiling at the camera`
	`- Outfit: in his suite, with a bag`
	`- Output a square image`

Prompting example for image replication	

`Generate close-up photo of this man / woman who looks lonely`
`Generate side-shot photo of this man / woman in the image`
`Generate close-up photo of him / her, who is thinking deeply with serious face`	

F. Experimental Details

F.1. Details of existing metrics

ArcFace utilizes face detection methods such as MTCNN [32] to localize faces and measures feature similarity within the embedding space [7, 12, 30]. In this study, faces detected with high confidence by MTCNN are used to extract embedding features for the specified number of concepts via Inception ResNet (V1) pretrained on VGGFace2 [2, 28]. Embedding features are similarly extracted from reference images, and their similarities are computed. For multi-concept cases, the highest similarity per embedding is selected, and the average across all embeddings forms the final score. If fewer faces are detected than concepts specified, missing face scores are set to zero.

CLIP T2T automatically generates captions from generated images using BLIP-2 [16], and then calculates the similarity of CLIP embedding features between the input text and the generated caption [14, 15, 20, 21]. In our investigation, this metric is currently the most popular evaluation method for concept customization.

CLIP T2I calculates the similarity of CLIP embedding features between the input text and the generated image [12, 14, 21]. This method is one of the most commonly used evaluation approaches, following CLIP T2T.

DINO Score is a method that inputs the generated image along with the text prompt or reference image into DINO's Vision Transformer encoder [3], and calculates the similarity of the resulting feature vectors [4, 15, 21, 23].

CLIP Aesthetic Score uses a CLIP model fine-tuned on an aesthetic evaluation dataset to calculate aesthetic scores from the embedding features [26]. In this study, we adopted the `LAION-Aesthetics Predictor V1`[1].

F.2. Implementation details of generative models

CustomDiffusion For each concept tuning, 20 images were used, and the model was trained for 3000 steps with a batch size of 2 and a learning rate of 5e-6, after which it was adapted for multi-concept use. During both training and sampling, the cross-attention parameters were frozen.

OMG+LoRA To train the character LoRA, we utilized the Kohya_ss [2] with a learning rate of 0.003, the Adafactor optimizer, a rank of 256, a batch size of 4, and 5 epochs, following the experimental setup of the original paper. For visual comprehension, we employed the GroundingDINO [19] combined with SAM [11]. The negative prompt used was: "noisy, blurry, soft, deformed, ugly."

OMG+InstantID Similar to OMG+LoRA, we employed GroundingDINO [19] combined with SAM [11] for visual comprehension, using the same negative prompt: "noisy, blurry, soft, deformed, ugly." Since this method requires

[1] https://github.com/LAION-AI/aesthetic-predictor

[2] https://github.com/bmaltais/kohya_ss

Table 4. **Prompting template for different types of inputs to GPT**. For each evaluation aspect, we selectively determined whether to provide the text prompt $\mathcal{T}$, the reference images $\mathcal{I}$, or neither with the generated image I_g, and accordingly made slight adjustments to the prompting templates.

I_g	$\mathcal{T}$	$\mathcal{I}$	**Prompting Template**
✓	✓		Task: I will provide a text prompt, followed by a generated image. Please rate how well the generated image meets the following evaluation aspect, then give a score from 1 to 5. DO NOT check whether the generated image matches the entire text prompt. Instead, rate it solely based on the following evaluation aspect. Evaluation aspect: <*Evaluation aspect*> Scoring example: <*Scoring example*> The text prompt: "<*Text prompt*>" <*Generated image*> Score:
✓		✓	Task: I will provide a generated image, followed by one or two reference images. Please observe carefully how well the generated image meets the following evaluation aspect, then give a score from 1 to 5. Evaluation aspect: <*Evaluation aspect*> Scoring example: <*Scoring example*> <*Generated image, reference images*> Score:
✓			Task: I will present a set of images cropped at different locations of the same generated image. Please observe carefully how well the generated image meets the following evaluation aspect, then give a score from 1 to 5 based on the worst image. Evaluation aspect: <*Evaluation aspect*> Scoring example: <*Scoring example*> <*generated image*> Score:

only a single image per concept for sampling, one full-body image was selected for each concept.

FastComposer For each concept, a single full-body image was selected. The parameters guidance_scale, inference_steps, and start_merge_step were set to their default values of 5, 50, and 10, respectively. No additional retraining or fine-tuning was performed.

Mix-of-Show For tuning the embedding-decomposed LoRA (ED-LoRA) per concept, all 20 images for each concept were used, and SAM was employed to generate the required mask images for training. In multi-concept settings, both unet_alpha and text_encoder_alpha were set to 1.0 as the default for concept fusion. Although keypose images or sketches can be used in the multi-concept sampling process, these inputs were disabled to ensure consistency with other methods. Since layout was the only mandatory aspect, the input consisted solely of a split layout, dividing the screen into left and right sections. The negative prompt used was the following, as specified in the original method's proposal: "longbody, lowres, bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality."

DreamBooth The implementation of the proposed model adopts Diffuser's Multi-Subject DreamBooth [29]. All parameters were kept at their default settings, and for each concept, training was conducted using all 20 images in our dataset, with a batch size of 1, 1500 training steps, and a learning rate of 1e-6.

G. Annotation Details

Detailed instructions for the annotation process are provided in Figure 2. All annotators were informed about the purpose of this dataset and provided consent before participation. The purpose of this annotation was to measure the correlation between human intuitive evaluations and the results obtained by the proposed method. To avoid bias introduced by detailed evaluation criteria, only the minimum necessary information for evaluating concept customization was provided. Detailed evaluation aspects were intentionally omitted to encourage annotators to rely on their own judgment criteria.

H. Scatter Plot of Predicted Scores

Figure 3 presents scatter plots of the human preference scores versus our predicted scores, shown both for all models collectively and for each model individually. The red line represents the regression line. From the figure, it can be observed that, for each method, the human preference scores and predicted scores exhibit a strong correlation.

I. Ranking Comparison

Table 7 presents the average ranking of each metric's scores across models. Existing methods exhibit considerable variation in their deviation from human preference based on average rankings. In contrast, our proposed method consistently achieves a ranking difference of less than 1 from human preference across all generative models, demonstrating stable alignment with human preference.

J. Case Study

Figure 4 shows some examples of our aspect-wise scores and their aggregated scores.

Table 5. **Details of the input to GPT for each evaluation aspect**. Each evaluation aspect is fed into the model along with the generated image I_g, and either the text prompt $\mathcal{T}$ and reference images $\mathcal{I}$, or neither.

Evaluation Aspect	I_g	$\mathcal{T}$	$\mathcal{I}$	Prompt

Subject Type	✓	✓		`Evaluation aspect:`
				`Do the generated objects and people match the specified types (e.g., 'a man' should not be misrepresented as 'a woman')? Focus ONLY on the subject type accuracy.`
				
				`Scoring example:`
				`If the genders are swapped, subtract 4 points.`

Quantity	✓	✓		`Evaluation aspect:`
				`Are the correct number of objects and persons generated according to the prompt? Focus ONLY on quantity accuracy.`
				
				`Scoring example:`
				`If the prompt specifies two men but three are generated, subtract 4 points.`

Subject & Camera Positioning	✓	✓		`Evaluation aspect:` `Are objects and people positioned correctly and arranged logically within the scene, preserving appropriate spatial relationships, depth, and occlusion according to the specified layout? Focus ONLY on the subject and camera positioning. If there is no relevant part in the text prompt, ignore the prompt.`  `Scoring example:` `If a 'long shot' is required but a close-up is generated, subtract 3 points.`

Size & Scale	✓	✓		`Evaluation aspect:`
				`Are the absolute and relative sizes of objects and people appropriate for the scene? Focus ONLY on the size and scale.`
				
				`Scoring example:`
				`If the man in the image appears too small relative to the surrounding objects, subtract 4 points.`

Color	✓		✓	`Evaluation aspect:`
				`Are the colors applied appropriately according to the reference images? Focus ONLY on the color accuracy.`
				
				`Scoring example:`
				`If the skin tone or hair color is different from the reference image, subtract 3 points.`

Subject Completeness	✓			`Evaluation aspect:`
				`Is the object or person fully generated with no missing or extra parts? Focus ONLY on the subject completeness.`
				`*Pay special attention to where the two individuals are in contact*`
				
				`Scoring example:`
				`If the hands touching the other person are semi-transparent or unclear, subtract 3 points.`

Evaluation Aspect	I_g	$\mathcal{T}$	$\mathcal{I}$	Prompt

Proportions & Body Consistency	✓		✓	`Evaluation aspect:`
				`Are body proportions and limb positioning natural and consistent with the given text prompt or reference image? Focus ONLY on the proportions and body consistency.`
				
				`Scoring example:`
				`If the limb or arm is unnatural, subtract 4 points; if the body proportions are off, subtract 3 points.`

Actions & Expressions	✓	✓		`Evaluation aspect:`
				`Are specified actions, poses, gaze direction, and facial expressions correctly depicted, reflecting the intended motion and emotion from the text prompt? Focus ONLY on the actions and expressions.`
				
				`Scoring example:`
				`If the man is instructed to laugh but isn't, subtract 4 points.`

Clothing & Attributes	✓		✓	`Evaluation aspect:`
				`Are clothing, accessories, and key features consistent with the reference images? Focus ONLY on the clothing and attributes.`
				
				`Scoring example:`
				`If the person is missing accessories, subtract 1 point; if the clothing differs completely from the reference, subtract 2 points.`

Facial Similarity & Features	✓		✓	`Evaluation aspect:`
				`Does the generated face resemble the reference image, preserving key characteristics like shape, expression, and symmetry? Focus ONLY on the facial similarity and features.`
				
				`Scoring example:`
				`If the face differs from the reference but keeps key features like hairstyle, subtract 3 points.`

Surroundings	✓	✓		`Evaluation aspect:`
				`Is the surrounding environment accurately depicted according to the provided text prompt? Focus ONLY on the surroundings.`
				
				`Scoring example:`
				`If a cafe is specified but a photo of a park is generated, assign 1 point; if there is no relevant part in the text prompt, ignore the prompt.`

Human & Animal Interactions	✓	✓		`Evaluation aspect:`
				`Are persons and animals interacting naturally with objects and each other as specified in the text prompt? Focus ONLY on the human and animal interactions.`
				
				`Scoring example:`
				`If the prompt specifies hugging but the image shows handshaking, subtract 4 points.`

Evaluation Aspect	I_g	$\mathcal{T}$	$\mathcal{I}$	Prompt

Object Interactions	✓	✓		`Evaluation aspect:`
				`Are objects interacting logically within the scene as specified in the text prompt? Focus ONLY on the object interactions.`
				
				`Scoring example:`
				`If the book in the prompt sinks into the table, subtract 4 points.`

Subject Deformation	✓			`Evaluation aspect:`
				`Are the people in the image (especially the faces and where the two individuals are in contact) rendered without deformations? Focus ONLY on the subject's deformation.`
				
				`Scoring example:`
				`If the person's face has any deformation or unrecognizable, subtract 4 points.`

Surroundings Deformation	✓			`Evaluation aspect:`
				`Are the surroundings in the image rendered naturally, without deformations such as crooked lines or unnatural parts? Focus ONLY on the surroundings' deformation.`
				
				`Scoring example:`
				`If the surroundings have deformation, subtract 4 points.`

Local Artifacts	✓			`Evaluation aspect:`
				`Are the image rendered without unwanted noise, strange patterns, or incomplete renderings? Focus ONLY on the local artifacts.`
				
				`Scoring example:`
				`If there is an unwanted watermark on the generated image, subtract 3 points.`

Detail & Sharpness	✓			`Evaluation aspect:`
				`Are facial features, hands, and intricate details well-defined? Focus ONLY on the detail and sharpness.`
				
				`Scoring example:`
				`If the entire image lacks detail, subtract 4 points; if a person is missing detail in any part (e.g., hands, legs, arms, face), subtract 2 points.`

Style Consistency	✓	✓		`Evaluation aspect:`
				`Does the generated image adhere to the artistic or visual style specified in the text prompt? Focus ONLY on the style consistency.`
				
				`Scoring example:`
				`If the prompt requires a realistic image but the style is anime-like, subtract 4 points.`

Table 6. Prompting template of Vanilla-GPT

I_g	$\mathcal{T}$	$\mathcal{I}$	**Prompting Template**
✓	✓	✓	`Task:` `I will provide a text prompt, followed by a generated image and one or two reference images. Please evaluate the generated image and assign a score on a scale from 1 to 10. Pay attention to whether the characteristics of the individuals in the reference images (including clothing, etc.) are preserved and whether the generated image follows the text prompt.`  `The text prompt` "*<Text prompt>*"  *<Generated image, reference images>*  `Score:`

Reference Images	A man  A woman 
Instruction	We present images generated by AI. Each image was created based on reference images and a text prompt (i.e., an instruction to generate the image). Please evaluate each image and assign a score on a scale from 1 to 10 based on intuitive preference, paying attention to whether the characteristics of the individuals in the reference images (including clothing, etc.) are preserved and whether the generated image follows the text prompt.

Figure 2. Instruction of annotation

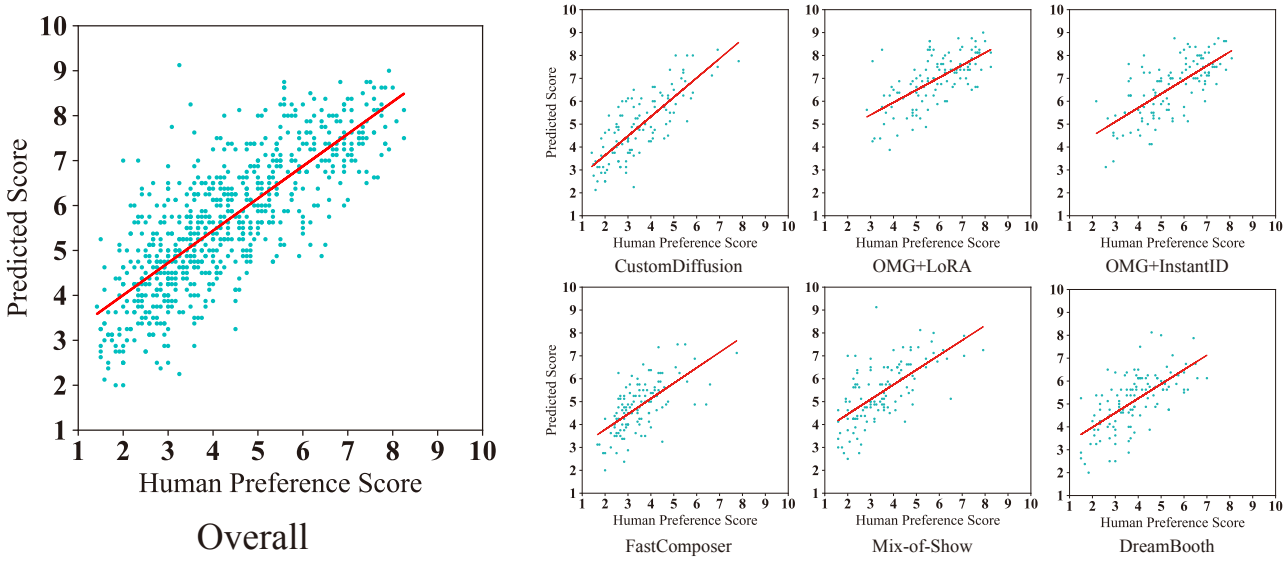

Figure 3. Scatter plots of the human preference scores versus our predicted scores.

Table 7. **Average rank of scores across models**. The metric closest to human preference is highlighted in **bold**, and the second closest metric is indicated by an underline.

Metric	CustomDiffusion	OMG +LoRA	OMG +InstantID	FastComposer	Mix-of-Show	DreamBooth
ArcFace	4.50 (-0.14)	3.41 (-1.83)	3.75 (-1.73)	1.88 (+2.55)	2.64 (+1.88)	4.56 (-0.68)
CLIP T2I	3.15 (+1.21)	2.62 (-1.04)	2.61 (-0.59)	4.58 (-0.15)	**4.53** (-0.01)	3.51 (+0.37)
CLIP T2T	3.49 (+0.87)	3.52 (-1.94)	2.63 (-0.61)	3.89 (+0.54)	4.05 (+0.47)	3.37 (+0.51)
CLIP Aes.	5.14 (-0.78)	2.39 (-0.81)	1.72 (+0.30)	3.86 (+0.57)	4.60 (-0.08)	3.29 (+0.59)
DINO	**4.31** (+0.05)	3.40 (-1.82)	3.49 (-1.47)	3.27 (+1.16)	3.11 (+1.41)	3.42 (+0.46)
Vanilla-GPT	4.23 (+0.13)	**1.66** (-0.08)	**1.99** (+0.03)	3.29 (+1.14)	3.58 (+0.94)	3.14 (+0.74)
Ours	4.30 (+0.06)	1.68 (-0.10)	2.16 (-0.14)	**4.56** (-0.13)	3.79 (+0.73)	**4.13** (-0.25)
Human Preference	**4.36**	**1.58**	**2.02**	**4.43**	**4.52**	**3.88**

Model	Mix-of-Show	OMG+InstantID	CustomDIffusion	OMG+LoRA
Output				
Text Prompt	A low angle shot of B* jumping up, Ultra HD quality.	A photo of A* punching, smiling playfully, Ultra HD quality.	A photo of B* and A* checking the time on their wristwatch, in a rustic café with wooden walls, Ultra HD quality.	A close shot of A* and B* putting their hands in their pockets, standing side by side, A* looking serious while B* looks relaxed, in a modern living room, Ultra HD quality.
Subject Type	5	5	1	5
Quantity	5	5	1	5
Sbject & Camera Positioning	5	3	1	5
Size & Scale	5	5	1	5
Color	2	2	1	2
Subject Completeness	2	5	5	5
Proportions & Body Consistency	2	5	1	5
Actions & Expressions	5	5	1	5
Facial Similarity & Features	3	4	1	3
Clothing & Attributes	3	1	1	2
Surroundings	3	4	5	5
Human & Animal Interactions	5	1	1	5
Object Interactions	5	5	1	5
Target Deformation	1	5	5	5
Surroundings Deformation	1	5	1	5
Local Artifacts	5	5	2	5
Detail & Sharpness	3	5	3	5
Style Consistency	5	5	1	5
Total	**6.50**	**8.13**	**2.88**	**9.00**

Model	FastComposer	DreamBooth	OMG+InstantID	OMG+LoRA
Output				
Text Prompt	A low angle shot of B* and A* giving each other a high-five, B* and A* are mid-air, Ultra HD quality.	A photo of A* putting his arm around B*'s shoulder, both smiling warmly, Ultra HD quality.	A close shot of A* and B* shaking hands, standing side by side, facing the camera, both smiling politely, on a quiet city street, Ultra HD quality.	A high angle shot of B* and A* playing rock-paper-scissors, crouching slightly, looking amused, in a cozy living room, Ultra HD quality.
Subject Type	5	5	5	5
Quantity	3	5	5	3
Sbject & Camera Positioning	2	5	2	5
Size & Scale	1	5	5	5
Color	2	2	2	2
Subject Completeness	2	2	2	2
Proportions & Body Consistency	2	2	5	5
Actions & Expressions	1	4	5	5
Facial Similarity & Features	4	2	2	2
Clothing & Attributes	2	3	2	2
Surroundings	4	3	5	5
Human & Animal Interactions	1	5	5	5
Object Interactions	1	5	5	5
Target Deformation	1	1	1	1
Surroundings Deformation	1	1	5	1
Local Artifacts	2	2	5	5
Detail & Sharpness	3	3	3	3
Style Consistency	1	2	5	5
Total	**3.50**	**5.88**	**7.38**	**7.00**

Figure 4. Case study

References

[1] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. In *ICCV*, 2023. 2

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 3

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021. 3

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot Object-level Image Customization. In *CVPR*, 2024. 3

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *ICCV*, 2023. 2

[6] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 2

[7] Junjie He, Yifeng Geng, and Liefeng Bo. UniPortrait: A Unified Framework for Identity-Preserving Single- and Multi-Human Image Personalization. *arXiv preprint arXiv:2408.05939*, 2024. 3

[8] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *EMNLP*, 2024. 1

[9] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. In *ICCV*, 2023. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, 2023. 3

[12] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. OMG: Occlusion-Friendly Personalized Multi-concept Generation in Diffusion Models. In *ECCV*, 2024. 3

[13] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *ICLR*, 2024. 2

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*, 2023. 2, 3

[15] Dongxu Li, Junnan Li, and Steven C.H. Hoi. BLIP-diffusion: pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 3

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2

[18] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE TPAMI*, 42(10):2684–2701, 2020. 1

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3

[20] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: customizable image synthesis with multiple subjects. In *NeurIPS*, 2023. 3

[21] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. In *SIGGRAPH*, 2024. 3

[22] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *ICLR*, 2025. 2

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2, 3

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[27] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *CVPR*, 2016. 1

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 3

[29] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. *GitHub repository*, 2022. 5

[30] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-Free Multi-subject Image Generation with Localized Attention. *IJCV*, 133(3):1175–1194, 2024. 3

[31] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 2

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23:1499–1503, 2016. 3

[33] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-LoRA Composition for Image Generation. *arXiv preprint arXiv:2402.16843*, 2024. 1

[34] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 2