

# Human Preference-Aligned Concept Customization Benchmark via Decomposed Evaluation

Reina Ishikawa<sup>1</sup>, Ryo Fujii<sup>1</sup>, Hideo Saito<sup>1</sup>, Ryo Hachiuma<sup>2</sup>

<sup>1</sup>Keio University, <sup>2</sup>NVIDIA

{reina.ishikawa, ryo.fujii0112, hs}@keio.jp, rhachiuma@nvidia.com

## Abstract

Evaluating concept customization is challenging, as it requires a comprehensive assessment of fidelity to generative prompts and concept images. Moreover, evaluating multiple concepts is considerably more difficult than evaluating a single concept, as it demands detailed assessment not only for each individual concept but also for the interactions among concepts. While humans can intuitively assess generated images, existing metrics often provide either overly narrow or overly generalized evaluations, resulting in misalignment with human preference. To address this, we propose **Decomposed GPT Score (D-GPTScore)**, a novel human-aligned evaluation method that decomposes evaluation criteria into finer aspects and incorporates aspect-wise assessments using Multimodal Large Language Model (MLLM). Additionally, we release **Human Preference-Aligned Concept Customization Benchmark (CC-AlignBench)**, a benchmark dataset containing both single- and multi-concept tasks, enabling stage-wise evaluation across a wide difficulty range—from individual actions to multi-person interactions. Our method significantly outperforms existing approaches on this benchmark, exhibiting higher correlation with human preferences. This work establishes a new standard for evaluating concept customization and highlights key challenges for future research. The benchmark and associated materials are available at <https://github.com/ReinaIshikawa/D-GPTScore>

## 1. Introduction

When evaluating AI-generated images from user-provided text prompts, what criteria do you employ to determine whether the generated images are of high quality? You may implicitly consider diverse evaluation aspects such as overall realism and fidelity to the text prompt. Although humans can intuitively and comprehensively evaluate generated images based on these implicit aspects, automatic methods that achieve comprehensive evaluation aligned with human

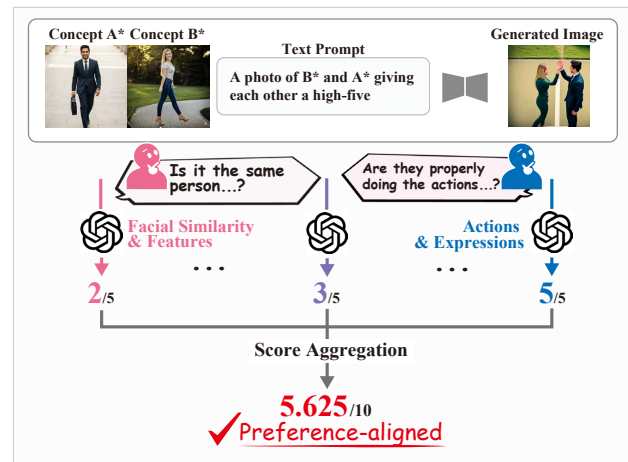


Figure 1. **D-GPTScore** evaluates images generated by concept customization through a two-step process: aspect-wise evaluation followed by aggregation. This approach achieves significantly higher correlation with human preference scores than existing methods, establishing a more reliable and human-aligned metric.

preferences remain underexplored in the research community.

In recent years, the rise of diffusion models [12, 28, 30, 35] has led to numerous text-to-image generation methods. Building upon these approaches, concept customization [31, 37, 46] aims to extend pretrained diffusion models to support single or multiple personalized concepts using only a few reference images per concept, combined with user-provided text prompts that condition the concepts (e.g., A\* is standing, where A\* denotes the concept).

To advance controllable image generation—hereafter referred to as *concept customization*—rigorous evaluation of generated images is essential to appropriately guide research progress. However, due to several critical issues in existing evaluation metrics and datasets, standardized benchmarks for the concept customization task remain underdeveloped.

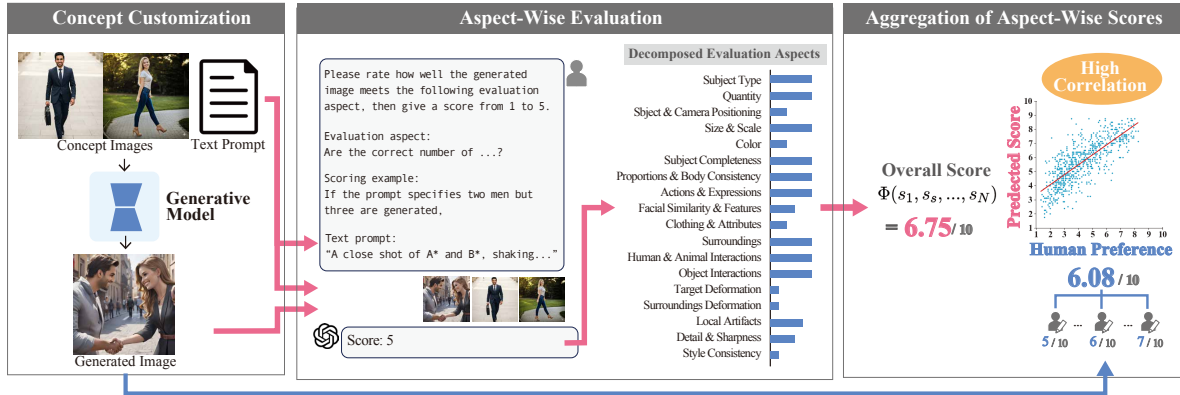


Figure 2. The pipeline of D-GPTScore. This metric comprises two phases: (1) **Aspect-wise evaluation with MLLM**, and (2) **Aggregation of aspect-wise scores**.

First, **existing evaluation metrics are not aligned with human preferences**. Humans can comprehensively consider various evaluation criteria and implicitly weigh important aspects to form an overall judgment of generated image quality. In contrast, existing evaluation metrics for concept customization [6, 11, 26] either assess only partial aspects or provide overly abstracted evaluations of generated images, resulting in low correlation with human preference scores (see Section 5.2 for experimental results).

Second, current benchmarks [18] only partially evaluate concept customization models, and **the evaluation settings are limited to relatively simple cases**. Most studies employ text-to-image benchmarks [1, 7, 13, 32] for concept customization tasks; however, these datasets contain incomplete inputs and evaluate only partial aspects of the task. Although CustomConcept101 [18] is a widely used and comprehensive benchmark, the target concepts and text prompts are relatively simple for customization models. For example, the target concepts predominantly consist of objects that do not interact with scenes or other concepts; the reference images are close-up shots of the concepts; and the text prompts for concepts describe non-interactive scenarios (e.g.,  $A^*$  is walking).

In this paper, we propose CC-AlignBench, which includes a human-aligned evaluation metric and a comprehensive evaluation dataset for the concept customization task. Inspired by LLM-as-a-Judge [47], we leverage the capabilities of state-of-the-art MLLMs, particularly GPT-4o [24], to develop a human-aligned evaluation metric. These MLLMs demonstrate excellent multi-modal language understanding by processing combined text and visual inputs, enabling tasks such as image captioning and visual question answering. This multi-modal understanding makes them well-suited for evaluating concept customization, which requires consideration of both text prompts and reference images.

However, if GPT-4o directly assesses the entire gener-

ated image at once, it may only provide a coarse overall score. To address this, we propose to *decompose* the evaluation criteria into multiple predefined aspects and evaluate each aspect individually, thereby preventing the omission of detailed factors and ensuring better alignment with human preferences. The individual aspect scores are then aggregated to produce the final evaluation metric.

Furthermore, we construct a dataset consisting of pairs of text prompts and reference images, focusing specifically on humans exhibiting more complex and diverse interactions with scenes and/or other humans. The dataset contains 980 text prompts encompassing three levels of human actions (*i.e.*, a single person’s action, two persons’ independent actions, and two persons’ mutual actions), and five conditioning types (*i.e.*, five different combinations of action, layout, expressions, and surroundings) (see Section 4 for details). By evaluating according to the levels of human actions or conditioning types, this dataset enables a systematic assessment of a model’s concept customization capabilities.

Our main contributions can be summarized as follows:

- We propose the **CC-AlignBench** for comprehensive evaluation of both single- and multi-concept customization, offering varying difficulty levels based on character count and scene complexity, thus allowing gradual assessment of model capabilities.
- We introduce an evaluation metric, **D-GPTScore**, leveraging MLLM that enables comprehensive and automated evaluation of concept customization. Our metric demonstrates better alignment with human preferences than existing methods and supports multi-concept evaluation.
- We conduct extensive ablation studies using the CC-AlignBench dataset to verify the effectiveness of our proposed metric. Experimental results confirm that our metric correlates more strongly with human preference scores than existing evaluation metrics.

We believe this benchmark will advance the concept cus-

tomization research community toward more realistic and human preference-aligned evaluation.

## 2. Related Work

### 2.1. Metrics for concept customization

Evaluating the quality of generated images using appropriate metrics and obtaining feedback is crucial for model improvement. However, existing methods fail to comprehensively assess all relevant aspects of concept customization, resulting in evaluations that diverge from human preferences.

Metrics such as Counting [26], SOA-I Score [11], Compositions [26], and YOLO Score [19] can recognize the number of subjects and layout in multi-subject scenarios, but their assessments are limited and do not evaluate verbs in the text prompt (*e.g.*, actions). Conversely, methods like the CLIP series [29] and DINO Score [4] provide overall evaluations by comparing images with text or texts with each other; however, their assessments are general and insufficient to capture fine-grained differences. QS [8], CLIP Aesthetic [34], and IS [8] demonstrate high correlation with human ratings but are restricted to assessing image quality. ImageReward [43] and VQA Score [21] enable integrated evaluation of a text prompt and a generated image through a preference-based model and Question Answering, respectively. However, both models are primarily designed for text-to-image tasks. Although GPT-4V Score [48] can evaluate both composition quality and image quality using MLLM, it does not account for multi-concept scenarios. In contrast, our proposed method enables a comprehensive evaluation of generated images across all necessary aspects, including fidelity to mutual interactions.

### 2.2. Benchmarks for concept customization

Proper evaluation of images generated through concept customization requires the consideration of both text prompts and reference images alongside the outputs. However, many existing datasets lack either prompts or reference images [1, 7, 13, 31, 32, 49]. Even when both are available, reference images often depict multiple individuals [5, 20, 27, 33, 44] or focus solely on faces or upper bodies [14, 18, 25], limiting their suitability for representing individual concepts. ImagenHub [17] is composed of pairs of concept images and corresponding text prompts, but excludes human subjects. As a result, no existing dataset adequately supports the evaluation of dynamic movements in concept customization. Accordingly, current methods often rely on combinations of such datasets and their associated metrics. Furthermore, datasets featuring complex actions, such as mutual interactions, remain scarce, and no benchmark explicitly addresses them.

To bridge these gaps, we introduce a dataset that includes

both prompts and reference images for single- and multi-concept scenarios. It also provides prompts of varying difficulty, enabling step-wise evaluation of generation performance.

### 2.3. MLLM-based human-preference scoring

Recent research has actively explored scoring methods aligned with human preferences using MLLMs, particularly for video evaluation [2, 3, 22, 38]. To improve alignment, several methods assess generative models by defining various aspects and scoring each individually [10, 41]. Some works aggregate aspect-wise scores based on human preferences, such as MMHE [23], which employs harmonic weighting, and MetaMetrics [40], which trains a regression model to combine the outputs of existing metrics.

For MLLM-based metrics in concept customization tasks, VIEScore [16] and CIGEval [39] proposed evaluation metrics by decomposing the task, but neither accounts for the action fidelity of interactions between living subjects. DreamBench++ [25] improves alignment with human scores on single-concept customization tasks by performing decomposition and outputting the reasoning process for task understanding. However, it does not explicitly generate scores for each evaluation aspect, which may reduce the reliability of the final score due to implicit reasoning [45].

Our method is unique in that it achieves human-aligned evaluation for concept customization—supporting both single- and multi-concept tasks—by performing unified explicit aspect-wise evaluation and integration using a single MLLM.

## 3. D-GPTScore Metric

Our concept customization evaluation metric is straightforward and consists of two phases: (1) **Aspect-wise evaluation using MLLM**, and (2) **Aggregation of aspect-wise scores**. As a preliminary step, we provide a brief overview of concept customization as follows:

**Concept Customization.** An image  $I_g$  is generated from a text prompt  $\mathcal{T} = \{t_1, t_2, \dots, t_k\}$ , which comprises several short text elements concatenated with commas, and a set of reference images  $\mathcal{I} = \{i_1, i_2, \dots, i_l\}$ , using the generative model  $\theta$ :

$$I_g = \theta(\mathcal{T}, \mathcal{I}). \quad (1)$$

### 3.1. Aspect-wise evaluation with MLLM

In the aspect-wise evaluation phase, we first decompose the evaluation criteria for concept customization into  $N$  predefined aspects  $\mathcal{A} = \{a_1, a_2, \dots, a_N\}$ , considering not only single- but also multi-concept scenarios (see Section 3.3 for details). To obtain the  $n$ -th aspect-wise score  $s_n \in \mathbb{R}$  of the generated image  $I_g$ , we independently feed the  $n$ -th evaluation aspect  $a_n$  into a MLLM  $\phi$ , along with the text prompt

Table 1. **Composition of the CC-AlignBench dataset**, divided into three difficulty levels based on the number of persons and interaction types.

Difficulty	#Total	#Base	#Persons per prompt	#Different Actions	Action Type
easy	260	52	1	13	non-interaction
medium	260	52	1	13	non-interaction
hard	460	92	2	23	mutual interaction

$\mathcal{T}$  and reference images  $\mathcal{I}$ :

$$s_n = \phi(a_n, I_g, \mathcal{T}, \mathcal{I}), \quad (2)$$

### 3.2. Aggregation of aspect-wise scores

Once the scores  $s_1, s_2, \dots, s_N$  for multiple aspects are obtained, the aggregation model  $\Phi$  is applied to compute the overall score  $s_{overall}$  on a scale of [1, 10] as follows:

$$s_{overall} = \Phi(s_1, s_2, \dots, s_N). \quad (3)$$

An ablation study on the choice of aggregation model is provided in Section 5.5.

### 3.3. Identification of evaluation aspects

We predefine the evaluation aspects into two main categories: (1) Concept Fidelity and (2) Quality Assessment.

“Concept Fidelity” assesses the adherence of the generated image to the text prompt and reference images and is further divided into the following four subcategories, totaling 13 aspects.

**1) Object Existence&Accuracy** assesses whether the specified objects or people are generated according to the prompt, in appropriate quantities, based on the *Subject Type* and *Quantity*.

**2) Layout & Composition Fidelity** assesses whether the *Subject & Camera Positioning* (i.e., relative positions, including camera angles) and the relative *Size & Scale* of the generated characters and objects align with the instructions in the prompt.

**3) Object-Level Fidelity** assesses whether the generated objects, people, and surroundings are accurately depicted, maintaining consistency with both the text and reference images in terms of *Color, Proportions & Body Consistency, Actions & Expressions, Facial Similarity & Features, Clothing & Attributes*, and *Surroundings*.

**4) Multi-Concept Interaction Consistency** assesses whether interactions between humans (i.e., *Human & Animal Interactions*) or between objects (i.e., *Object Interactions*) are generated in a natural and coherent manner, as specified in the text prompt.

“Quality Assessment” evaluates image quality, consisting of 5 aspects: *Subject Deformation, Surroundings Deformation, Local Artifacts, Detail&Sharpness*, and *Style Consistency*.

## 4. Dataset

### 4.1. Text preparation

As shown in Table 1, CC-AlignBench consists of 196 prompts divided into three difficulty levels: 52 *Easy*, 52 *Medium*, and 92 *Hard* prompts (referred to as **difficulty levels**). *Easy* prompts describe a single person acting; *Medium* prompts involve two individuals performing independent actions such as “standing” or “walking”; and *Hard* prompts depict interactions between two individuals, such as “hugging” or “whispering.”

Each of the 196 base prompts comprises four elements: *action*, *layout*, *expression*, and *surroundings*. To evaluate the model’s expressive capabilities, five variations are generated from each base prompt: *action-only* prompt  $t_{act}$ , *action+layout* prompt  $t_{act}, t_{lt}$ , *action+expression* prompt  $t_{act}, t_{exp}$ , *action+surroundings* prompt  $t_{act}, t_{surr}$ , and *all* prompt  $t_{act}, t_{lt}, t_{exp}, t_{surr}$ . Since each of the 196 prompts is expanded into these five variations, the total number of prompt variations in CC-AlignBench amounts to  $196 \times 5 = 980$ . An example text prompt is shown in Table 2.

### 4.2. Image preparation

To address concerns related to portrait rights, we used generative AI to create images without relying on photographs of real individuals. First, we generated one image each of a man and a woman using *Gemini 2.5 Flash Preview* [36]. Then, using *Gemini 2.0 Flash Preview Image Generation*, we instructed the model to produce various patterns (e.g., close-up, long shot, frontal, and profile views) of the same man and woman. We added 19 additional images per person, resulting in 20 images per individual and a total of 40 images. For models that require a single input image, we ensured the same full-body image was consistently used.

## 5. Experiments

### 5.1. Experimental setup

**Baselines of evaluation metrics** To verify whether the proposed method aligns with human preferences, we compared it with five existing evaluation metrics commonly used for concept customization: identity preservation score by ArcFace [6], CLIP T2T score [29], CLIP T2I score [29], CLIP Aesthetic Score [34], and DINO score [4]. See our supplementary material for the details.

Table 2. **Example text prompt in CC-AlignBench.** Each prompt consists of the four base prompts: action, layout, expression, and surroundings.

type	Example Text Prompt
action-only	A photo of a woman putting her arm around a man’s shoulder, Ultra HD quality.
action +layout	A high angle shot of a woman putting her arm around a man’s shoulder, standing close, Ultra HD quality.
action +expression	A photo of a woman putting her arm around a man’s shoulder, both looking amused, Ultra HD quality.
action +surroundings	A photo of a woman putting her arm around a man’s shoulder, in an open green park, Ultra HD quality.
all	A high angle shot of a woman putting her arm around a man’s shoulder, standing close, both looking amused, in an open green park, Ultra HD quality.

Table 3. **Comparison of correlation with human preference on CC-AlignBench.** The left shows Pearson’s correlation coefficient, and the right shows Spearman’s rank correlation (higher is better). A correlation of 0.5 or higher is generally considered strong, and the proposed metric significantly exceeds this threshold for the overall score and most individual models.

Metric	CustomDiffusion	OMG +LoRA	OMG +InstantID	FastComposer	Mix-of-Show	DreamBooth	Overall
ArcFace	0.34 / 0.26	0.05 / 0.01	-0.01 / 0.06	0.21 / 0.27	0.50 / 0.33	0.21 / 0.15	0.23 / 0.04
CLIP T2I	0.20 / 0.27	-0.01 / 0.13	0.11 / 0.08	0.09 / 0.36	0.21 / 0.38	0.20 / 0.33	0.29 / 0.42
CLIP T2T	0.01 / 0.27	-0.04 / -0.07	-0.04 / -0.09	0.16 / 0.20	0.19 / 0.23	0.16 / 0.21	0.14 / 0.21
CLIP Aes.	0.46 / 0.09	0.14 / 0.04	0.20 / 0.00	0.26 / 0.15	0.29 / 0.02	0.44 / 0.26	0.51 / 0.49
DINO	0.28 / 0.10	0.04 / 0.14	0.30 / 0.10	-0.17 / 0.09	-0.07 / -0.20	0.26 / -0.02	0.10 / 0.04
Ours	<b>0.80 / 0.54</b>	<b>0.66 / 0.34</b>	<b>0.70 / 0.47</b>	<b>0.64 / 0.46</b>	<b>0.65 / 0.44</b>	<b>0.64 / 0.38</b>	<b>0.78 / 0.69</b>

**Target generative models** We evaluated five widely used and state-of-the-art image generation models that support multi-concept customization: CustomDiffusion [18], OMG [15] with LoRA or InstantID, FastComposer [42], Mix-of-Show [9], and DreamBooth [31], using our proposed benchmark.

**Aggregation model selection** For the aggregation model  $\Phi$ , we adopted the average function:

$$\Phi(s_1, s_2, \dots, s_N) = \frac{1}{N} \sum_{n=1}^N s_n, \quad (4)$$

meaning that each aspect is treated equally to compute the final score. The choice of aggregation model is further discussed in Section 5.5.

**MLLM setting** We employed gpt-4o-2024-08-06 [24] as the MLLM for aspect-wise evaluation. All generated and reference images fed to GPT-4o were resized to 512×512 pixels. For evaluation aspects where only generated images are provided (e.g., deformations or local artifacts), resolution can affect accuracy. Therefore, we additionally provided GPT with two square-cropped images, each covering 50% of the original size, centered vertically and placed on the left and right halves, along with the original image.

**Preference score annotation** To evaluate human alignment, we measured the correlation between predicted scores and human preference scores. The human scores were obtained from 12 expert annotators (5 females, 7 males), all members of our laboratory. We randomly selected 40 prompts from each of the three difficulty levels (i.e., *easy*, *medium*, *hard*) and generated images using six different models for each prompt. The resulting 720 images were evaluated by the annotators, who scored them from 1 to 10 with reference to both the text prompts and the reference images. We then averaged the scores across annotators to obtain a single human preference score per image. The annotation took approximately two hours per annotator. Each annotator was compensated \$48 according to the current exchange rate.

**Experimental environment** All experiments were conducted on a system equipped with a single NVIDIA RTX A6000 (48GB), an Intel(R) Core(TM) i9-9940X CPU, and the OpenAI API.

## 5.2. Correlation with human preference

For each evaluation metric, we computed the Pearson’s correlation coefficient and Spearman’s rank correlation between the human preference scores and the predicted scores. The results are summarized in Table 3. Our metric achieves an overall Pearson’s correlation of 0.78 and an overall

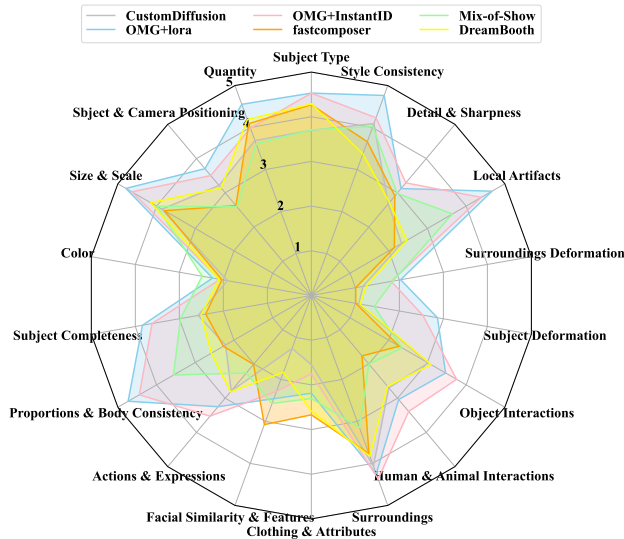


Figure 3. **Benchmark results of aspect-wise evaluation.** Each aspect is scored on a scale from 1 to 5, with higher values indicating better performance.

Spearman’s rank correlation of 0.69 with human preferences, significantly surpassing all existing metrics. These results demonstrate that the proposed method provides human-aligned evaluation for concept customization and outperforms existing metrics. Furthermore, since omission of any key aspect may reduce correlation, the strong correlation exceeding 0.7 suggests that the proposed decomposed aspects are necessary and sufficient.

### 5.3. Benchmark results

Table 4 presents the benchmark scores of each image generation model on our dataset, and Figure 3 shows the 18 aspect-wise scores.

As shown in Table 4, scores generally decrease as the action difficulty increases from *Easy* to *Hard*. OMG+LoRA and OMG+InstantID achieved the highest scores at the *Easy* level, with 7.28 and 7.13, respectively; however, OMG+InstantID’s score dropped notably at the *Hard* level. In contrast, DreamBooth showed the smallest decrease, indicating more stable performance.

While the primary objective of our paper is to achieve human preference-aligned evaluation, the aspect-wise scores obtained as intermediate results provide detailed feedback, as illustrated in Figure 3. All models exhibited difficulty with complex actions, leading to low scores in object fidelity aspects such as *Color*, *Facial Similarity & Features*, and *Clothing & Attributes*, as well as in *Actions & Expressions* and *Human & Animal Interactions*. However, notable differences emerged in *Proportions & Body Consistency*, *Object Interactions*, and *Local Artifacts*, where the OMG-based methods consistently scored high across diffi-

culty levels, while others struggled.

### 5.4. Qualitative results

Table 5 presents a scoring example on CC-AlignBench. The table shows that existing metrics produce rankings that differ considerably from human preference scores. In contrast, although minor fluctuations are observed, the proposed method generally aligns well with human preferences.

### 5.5. Ablation study

This section presents the results of ablation studies on three components of the proposed metric: the effectiveness of decomposition, the MLLM used for aspect-wise scoring, and the aggregation model. All results are summarized in Table 6.

**Decomposition ablation** To examine the necessity of decomposition, we compared our method with a metric that directly outputs a score on a 1–10 scale using the GPT-4o model without decomposition (referred to as Vanilla-GPT). As shown in Table 6, the proposed decomposed metric significantly outperforms Vanilla-GPT in both individual model scores and overall correlation with human preferences. This suggests that directly using GPT-4o without decomposition fails to capture fine-grained evaluation aspects, resulting in poor alignment with human preferences.

**MLLM ablation** We adopted GPT-4o as the MLLM for aspect-wise evaluation due to its superior multi-modal understanding capabilities. However, considering API costs, we also evaluated with GPT-4o mini as a substitute. GPT-4o mini yielded lower overall Pearson’s and Spearman’s correlations, both by 0.10, compared to GPT-4o. Nevertheless, GPT-4o mini still outperformed existing metrics significantly (see Table 3). Thus, while GPT-4o provides more accurate evaluation, GPT-4o mini offers a viable alternative depending on computational resources. Notably, the average token counts per generated image were 12,091 (input) and 96 (output) for GPT-4o, versus 299,672 (input) and 104 (output) for GPT-4o mini.

**Aggregation model ablation** Although this study employs averaging of aspect-wise scores to compute the final score, learnable models such as linear regression can also be used to better align with human preference scores. Results using linear regression are reported in Table 6.

With linear regression, the overall Pearson correlation coefficient and Spearman rank correlation coefficient are slightly lower, at 0.75 and 0.62, than those obtained with averaging. This may be due to distribution shifts among generation models, causing slight overfitting. We expect that as concept customization research progresses and more

Table 4. Average benchmark scores for each difficulty level on CC-AlignBench, predicted using D-GPTScore. Each score ranges from 1 to 10, with higher values indicating better image generation performance. When applying linear regression, the leave-one-out approach is employed to ensure that the generated images from the model under evaluation are excluded from the training data.

Difficulty Level	CustomDiffusion	OMG+LoRA	OMG+InstantID	FastComposer	Mix-of-Show	DreamBooth
Easy	5.64	<b>7.28</b>	7.13	5.21	5.75	5.53
Medium	4.62	<b>7.05</b>	6.64	4.72	5.27	5.01
Hard	4.30	<b>6.41</b>	6.10	4.45	4.82	4.87
Overall	4.74	<b>6.81</b>	6.52	4.72	5.19	5.09

Table 5. D-GPTScore scoring examples on the *Hard* level subset of CC-AlignBench. For each model, the top three scores are highlighted in red, blue, and yellow, respectively, to illustrate their correspondence with the human preference rankings.

Input	Reference Images		Text Prompts	Outputs					
				CustomDiffusion	OMG+LoRA	OMG+InstantID	FastComposer	Mix-of-Show	DreamBooth
			A long shot of A* and B* exchanging a book, A* is standing slightly in front of B*, both looking engaged, in a European town square, Ultra HD quality.						
	ArcFace	0.396	0.940	0.952	0.971	0.957	0.476		
	CLIP T2I	0.319	0.337	0.361	0.304	0.238	0.315		
	CLIP T2T	0.814	0.855	0.856	0.688	0.584	0.750		
	CLIP Aesthetic	5.478	6.107	7.017	5.630	5.718	6.320		
	DINO	0.717	0.773	0.757	0.787	0.779	0.785		
	<b>Ours</b>	<b>3.375</b>	<b>8.250</b>	<b>7.000</b>	<b>5.250</b>	<b>3.375</b>	<b>5.000</b>		
	Human Preference	2.583	1 <sup>st</sup> 5.833	2 <sup>nd</sup> 5.417	3 <sup>rd</sup> 4.250	2.417	3.167		

generative models become available, regression-based aggregation may become more effective.

Importantly, achieving high correlation with human preferences even without regression suggests that the 18 evaluation aspects were appropriately and sufficiently defined.

## 6. Discussion

In this section, we explore an extension of the proposed method by combining it with existing metrics to achieve a more comprehensive and accurate evaluation. Specifically, after obtaining 18 scores through aspect-wise evaluation, aggregation is performed by combining these scores with those from existing metrics. Table 7 presents the results of combining the 18 aspect-wise scores (Ours++) with scores from six existing metrics used for comparison (*i.e.*, ArcFace, CLIP T2T score, CLIP T2I score, CLIP Aesthetic Score, DINO, and Vanilla GPT). Both the aspect-wise and existing scores were min-max normalized and scaled to a 1–10 range, resulting in a total of 24 aspects. The table

shows that combining the proposed method with existing metrics yields a slight increase in correlation compared to using the proposed method alone.

## 7. Limitations

**Dataset** This paper’s dataset excludes stationary objects and animals, focusing on humans due to the greater challenges posed by their complex structures and movements. However, since the evaluation metric also applies to scenes with objects and animals, future work will expand the dataset to include these, enabling more comprehensive evaluations.

The study focuses on generation difficulties arising from variations in human actions and emphasizes text prompt diversity. While only male and female humans were used as concept images in this study, future work will broaden the range of subjects.

Table 6. **Ablation studies on three components of our metric:** decomposition (Decomposition), the MLLM used for aspect-wise scoring (MLLM), and the aggregation model (Aggregation). The left value in each cell denotes Pearson’s correlation coefficient, and the right value denotes Spearman’s rank correlation.

Decomposition	MLLM	Aggregation	Custom Diffusion	OMG+ LoRA	OMG+ InstantID	Fast Composer	Mix-of-Show	Dream Booth	Overall
✗	GPT-4o	average	0.77 / 0.42	0.42 / 0.25	0.50 / 0.41	0.51 / 0.43	<b>0.67 / 0.47</b>	0.58 / 0.27	0.67 / 0.58
✓	GPT-4o mini	average	0.75 / <b>0.56</b>	0.49 / <b>0.42</b>	0.63 / 0.44	0.47 / 0.28	0.65 / 0.37	0.56 / 0.29	0.68 / 0.59
✓	GPT-4o	linear regression	0.80 / 0.50	0.65 / 0.32	0.69 / 0.36	0.64 / 0.40	0.65 / 0.36	<b>0.67 / 0.39</b>	0.74 / 0.62
✓	GPT-4o	average	<b>0.80</b> / 0.54	<b>0.66</b> / 0.34	<b>0.70 / 0.47</b>	<b>0.64 / 0.46</b>	0.65 / 0.44	0.64 / 0.38	<b>0.78 / 0.69</b>

Table 7. **Exploration of extending our metric by combining it with existing metrics.** Ours++ denotes the combination with existing metrics. The left value in each cell represents Pearson’s correlation coefficient, and the right value represents Spearman’s rank correlation.

Model	Custom Diffusion	OMG+ LoRA	OMG+ InstantID	Fast Composer	Mix-of-Show	Dream Booth	Overall
Ours (average)	0.80 / <b>0.54</b>	0.66 / 0.34	<b>0.70 / 0.47</b>	<b>0.64</b> / 0.46	0.65 / 0.44	0.64 / 0.38	0.78 / 0.69
Ours++ (average)	0.82 / 0.51	<b>0.67</b> / 0.34	0.69 / 0.45	0.63 / 0.44	0.67 / <b>0.40</b>	<b>0.70</b> / 0.40	<b>0.80 / 0.72</b>
Ours (linear regression)	0.80 / 0.50	0.64 / <b>0.38</b>	0.69 / 0.36	0.64 / 0.41	0.65 / 0.36	0.67 / 0.41	0.75 / 0.62
Ours++ (linear regression)	<b>0.83</b> / 0.48	0.66 / 0.36	0.66 / 0.37	0.64 / <b>0.50</b>	<b>0.70</b> / 0.38	0.68 / <b>0.45</b>	0.78 / 0.67

**Supported input sets** Some image generation methods accept optional inputs, such as pose or sketch images. However, our evaluation metric assumes that generative models take only text and reference images as inputs and excludes additional inputs from the evaluation. Considering the input cost to the MLLM, providing all additional inputs and expanding the evaluation aspects accordingly is impractical.

Moreover, accounting for these additional inputs would increase the variability of inputs used for evaluation across models, thereby complicating accurate model comparisons. Therefore, it is preferable to treat such additional inputs as optional during image generation, while the proposed evaluation metric focuses on the most fundamental and common inputs: text and reference images.

**Generative models** In this paper, we selected six generative models that support multi-concept customization as evaluation targets and obtained benchmark results. However, as discussed in Section 5.5, due to large score variance among these models, the use of trainable aggregation methods, such as regression models, led to overfitting during training. Consequently, the correlation achieved using linear regression did not improve significantly compared to averaging. In the future, as multi-concept customization research advances and more models become available, this overfitting issue is expected to be mitigated, rendering regression a more effective aggregation method and enabling scoring with higher correlation to human preferences.

## 8. Ethical Considerations

In creating our benchmark dataset, we carefully removed harmful, offensive, or abusive expressions. Since the dataset is intended solely for image evaluation, it poses

no direct social risks. However, it is important to note that terms such as “hit” and “kick” are included to introduce variations in mutual interactions. As generated images heavily depend on the training data, controlling the output is challenging. To respect portrait rights, only AI-generated reference images are used, eliminating concerns about harm to individuals. Additionally, in our experiments, visual inspections were conducted to ensure the images do not contain harmful or offensive content. Nevertheless, toxic images may still be generated depending on the model. Users should be aware of the potential risks and refrain from using this dataset to generate offensive or inappropriate images, especially of real individuals. Caution is advised when releasing these images.

## 9. Conclusion

This paper proposes D-GPTScore, a novel evaluation metric for concept customization. Since existing metrics have yielded results misaligned with human preferences, we propose to decompose evaluation criteria for single and multiple concepts into 18 aspects, perform aspect-wise evaluation using MLLM, and aggregate the results to achieve a comprehensive and human-aligned assessment. Furthermore, we introduced CC-AlignBench, a benchmark dataset supporting both single- and multi-concept evaluations. Extensive experiments demonstrate that our metric correlates significantly better with human preferences compared to prior metrics, and ablation studies confirm the effectiveness of the decomposition strategy.

## Acknowledgment

This work was supported by JST BOOST, Japan Grant Number JPMJBS2409.

## 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, 3
- [2] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *ICRL*, pages 102075–102121, 2025. 3
- [3] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. *arXiv preprint arXiv:2503.06800*, 2025. 3
- [4] 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
- [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. 3
- [6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE TPAMI*, 44(10\_Part\_1):5962–5979, 2022. 2, 4
- [7] 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, 3
- [8] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. GIQA: Generated Image Quality Assessment. In *ECCV*, 2020. 3
- [9] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, 2023. 5
- [10] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhui Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *EMNLP*, 2024. 3
- [11] Tobias Hinze, Stefan Heinrich, and Stefan Wermter. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *IEEE TPAMI*, 44(3):1552–1565, 2022. 2, 3
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 1
- [13] 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, 3
- [14] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019. 3
- [15] 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. 5
- [16] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *ACL*, pages 12268–12290. *ACL*, 2024. 3
- [17] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhui Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *ICLR*, 2024. 3
- [18] 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, 5
- [19] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image Synthesis From Layout With Locality-Aware Mask Adaption. In *ICCV*, 2021. 3
- [20] 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. 3
- [21] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, page 366–384, 2024. 3
- [22] Nema, Preksha and Khapra, Mitesh M. Towards a Better Metric for Evaluating Question Generation Systems. In *EMNLP*, 2018. 3
- [23] Masanari Ohi, Masahiro Kaneko, Naoaki Okazaki, and Nakamasa Inoue. Multi-modal, Multi-task, Multi-criteria Automatic Evaluation with Vision Language Models. *arXiv preprint arXiv:2412.14613*, 2025. 3
- [24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024. 2, 5
- [25] 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. 3
- [26] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded Text-to-Image Synthesis with Attention Refocusing. In *CVPR*, 2024. 2, 3
- [27] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *ICCV*, 2015. 3
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 3, 4
- [30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022. 1
- [31] 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. 1, 3, 5
- [32] 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, 3
- [33] 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. 3
- [34] 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, 4
- [35] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1
- [36] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, and David Silver et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2025. 4
- [37] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*, 2023. 1
- [38] Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. AIGV-Assessor: Benchmarking and Evaluating the Perceptual Quality of Text-to-Video Generation with LMM. *arXiv preprint arXiv:2411.17221*, 2024. 3
- [39] Jifang Wang, Xue Yang, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. A unified agentic framework for evaluating conditional image generation. *arXiv preprint arXiv:2504.07046*, 2025. 3
- [40] Genta Indra Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Tanti Wijaya. MetaMetrics: Calibrating Metrics For Generation Tasks Using Human Preferences. *arXiv preprint arXiv:2410.02381*, 2025. 3
- [41] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *ICML*, 2024. 3
- [42] 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. 5
- [43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, pages 15903–15935, 2023. 3
- [44] 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. 3
- [45] Yijiong Yu. Do llms really think step-by-step in implicit reasoning? *arXiv preprint arXiv:2411.15862*, 2025. 3
- [46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 1
- [47] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*, 2023. 2
- [48] 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. 3
- [49] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 3