

UniEval: Unified Holistic Evaluation for Unified Multimodal Understanding and Generation

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, Xiaomeng Li✉
 The Hong Kong University of Science and Technology
 {ylini,xmli}@ust.hk
<https://xmed-lab.github.io/UniEval/>

Abstract

The emergence of unified multimodal understanding and generation models is rapidly attracting attention because of their ability to enhance instruction-following capabilities while minimizing model redundancy. However, there is a lack of a unified evaluation framework for these models, which would enable an elegant, simplified, and overall evaluation. Current models conduct evaluations on multiple task-specific benchmarks, but there are significant limitations, such as the lack of overall results, errors from extra evaluation models, reliance on extensive labeled images, benchmarks that lack diversity, and metrics with limited capacity for instruction-following evaluation. To tackle these challenges, we introduce UniEval, the first evaluation framework designed for unified multimodal models without extra models, images, or annotations. This facilitates a simplified and unified evaluation process. The **UniEval** framework contains a holistic benchmark, **UniBench** (supports both unified and visual generation models), along with the corresponding **UniScore** metric. UniBench includes 81 fine-grained tags contributing to high diversity. Experimental results indicate that UniBench is more challenging than existing benchmarks, and UniScore aligns closely with human evaluations, surpassing current metrics. Moreover, we extensively evaluated SoTA unified and visual generation models, uncovering new insights into UniEval’s unique values.

1 Introduction

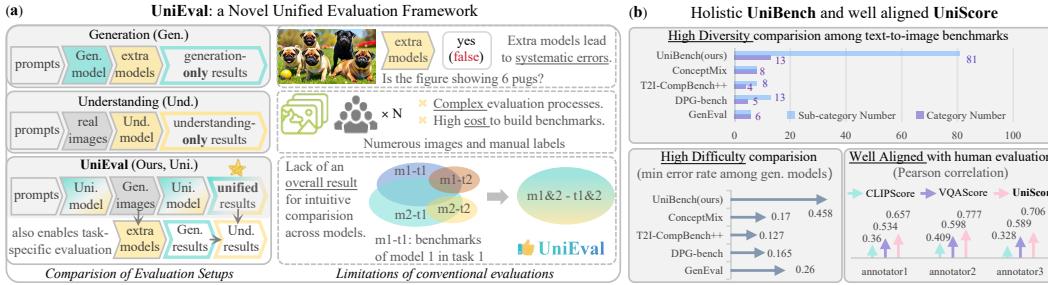


Figure 1: **Overview of UniEval.** (a). The proposed UniEval unifies the evaluation of both the multimodal understanding and generation, eliminating limitations due to extra models, labeled images, and the lack of overall results. (b). The proposed UniBench is a holistic and challenging benchmark, with the UniScore metric aligning well with humans.

The unified multimodal understanding and generation models [65, 62] are rapidly emerging. Many recent works [8, 64, 41, 30] have proved that unified models can enhance instruction-following capabilities in visual generation [21, 26] and reduce model redundancy. Although these models unify diverse tasks, their evaluations are still the same as task-specific models [50, 6, 7, 3, 10] relied on many conventional benchmarks [21, 26, 27, 71, 34, 39, 2, 40].

We cannot deny the value of task-specific evaluations via multiple benchmarks; however, current benchmarks [23, 21, 67, 11, 27, 71, 39, 2] have significant limitations for unified models: **1) Lack of overall results**: Averaging an overall result demands consistent benchmarks across models, which is usually not standardize in multitask settings, and many benchmarks raise the evaluation cost.**2) Systemic error from extra models**: Automatic evaluations of visual generative [23, 21, 67, 63] rely on extra models [1, 29, 70], and incorrect model predictions introduce unavoidable systemic errors (e.g., Wu et al [63] reported error rate from 9.14% to 25.8% even for common attribute). **3) High resource costs**: Evaluations of multiple understanding benchmarks [2, 34, 39, 71] demand massive labeled images, where datasets in dozens of GB size make evaluation complex, and manual labeling makes building new datasets difficult. **4) Limited benchmark diversity and difficulty**: Existing visual generation datasets are limited in diversity and difficulty, restricting the evaluation of more advanced models (see Fig. 1b). **5) Insufficient metrics for instruction-following**: Conventional metrics like FID [24], IS [52], and CLIPScore [23] are insufficient to evaluate instruction-following capabilities for complex prompts, which are crucial for unified models. So far, there is a lack of a unified benchmark for unified models, which would enable a simplified and overall evaluation.

To address these limitations, *our motivation is to leverage the dual capabilities of unified models to evaluate themselves*. Specifically, the understanding part is applied to evaluate its visual generation without extra models, where their systematic errors are cleverly converted into Und. performance merged in the overall result. Meanwhile, generated images from the visual generation part eliminate massive labeled images, simplifying the evaluation process. This solution also yields an overall result, making model comparisons more intuitive and standardized. To support this motivation, an informative benchmark is needed to evaluate both understanding and generation. While current text-to-image benchmarks [27, 63, 33, 51, 11] just focus on some basic concepts, which are inadequate to evaluate understanding. Thus, we aim to build a holistic, challenging, and fine-grained compositional benchmark to evaluate unified models, emphasizing the key instruction-following capability.

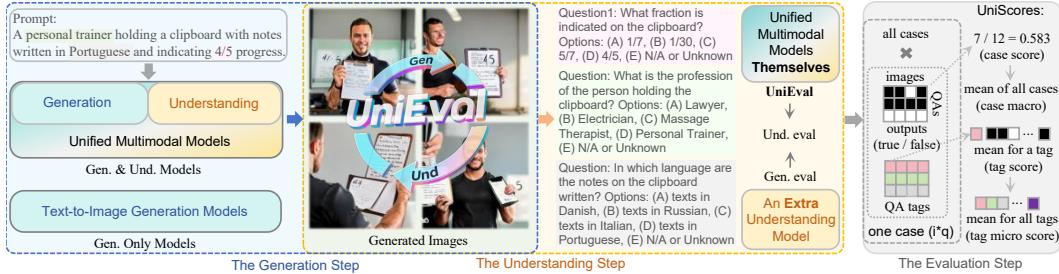


Figure 2: **Workflow of UniEval.** An example in UniBench processed by Janus-Pro-7B [8] to generate four images and outputs choices for each image and question (more examples in Appendix J). UniScores involves case-level accuracy in a case and tag-level accuracy from answers in the same tag. Our method is versatile, supporting generation evaluation with an extra model, and the understanding via the difference between unified and generation results (see Fig. 5 and Appendix G).

Based on the above motivation, we propose a unified evaluation framework called **UniEval** to evaluate unified multimodal models. It includes a holistic benchmark **UniBench** with the corresponding **UniScore** metric, which aligns well with human evaluations. We define 13 level-1 tags and 81 level-2 tags to ensure diversity, supporting evaluation in multiple aspects. Based on these tags, we extensively enumerate keywords, then generate 1,234 prompts and 4,231 question-answer (QA) pairs with manual quality control (see details in Fig. 3). In this benchmark, an evaluated model is required to generate four images for each prompt and output the correct options based on the given questions as shown in Fig. 2. We calculate UniScores in both case-level and tag-level to analyze model performance, with the average of the level-1 tags as the overall UniScore. Our method is versatile, also allowing for task-specific evaluations through an extra model for generation evaluation. This enables a comparison between unified and generated results to analyze the specific understanding performance.

UniEval is the first unified evaluation framework and outperforms existing benchmarks in multiple aspects. Specifically, the 81 fine-grained tags of UniBench significantly exceed existing benchmarks [11, 26] (≤ 13). Moreover, UniBench presents higher difficulty, whose min error rate is 0.458, compared to 0.26 of GenEval and 0.165 of DPG-Bench (see Tab. 1). Notably, UniScore is an effective metric. Our multiple-choice metric correlates better with human evaluation on UniBench, achieving a Pearson correlation [12] of 0.716 with three annotators, compared to CLIPScore [23] and VQAScore

[33], which scored 0.372 and 0.575, respectively. Moreover, UniEval has about twice the model discriminability beyond task-specific benchmarks measured by the coefficient of variation in Fig. 5, and it also supports separate analysis of Und. or Gen. From the leaderboards in Table 2 and Table 3, UniEval reveals new insights in Table 4, highlighting its unique value. Our contributions include:

- We propose UniEval, the first evaluation framework for unified multimodal models, eliminating reliance on extra models and labeled images, achieving a simplified unified evaluation.
- UniEval includes a holistic UniBench, which is currently the most challenging text-to-image benchmark with the highest number of fine-grained tags. The corresponding UniScore metric aligns well with human evaluations, surpassing existing metrics.
- We conducted extensive evaluations on SoTA unified models and visual generation models, highlighting that UniEval can provide new insights with its unique values.

Table 1: Benchmark Comparison. UniBench offers the most extensive tags and sub-tags in compositional text-to-image generation benchmarks, achieving high diversity. UniBench provides five related choices to minimize random error beyond binary options. UniBench has high difficulty, leading to a higher error rate of the SoTA model and more room for improvement. UniBench includes new features like generation evaluation, image-free, and annotation-free beyond Und. benchmarks.

| Compositional Gen. Benchmark | Diversity | | | Difficulty | | Avg. Rank | Representative Und. Benchmark | New Features | | |
|---------------------------------|-----------|-----------|--------------|------------|--------------|------------|----------------------------------|--------------|----------|-----------|
| | Tags | Sub-Tags | Prompts | Options. | SoTA Error | | | Gen. Eval | Img-Free | Anno-Free |
| T2I-CompBench++ [27] | 4 | 8 | 8,000 | 2 | 0.127 | 4 | VQA [2] | ✗ | ✗ | ✗ |
| GenAI-Bench [33] | 8 | 8 | 1,600 | 2 | 0.29 | 2.4 | GQA [28] | ✗ | ✗ | ✗ |
| DSG-1K [11] | 4 | 13 | 1,060 | 2 | 0.161 | 4 | SEED [34] | ✗ | ✗ | ✗ |
| ConceptMix [63] | 8 | 8 | 300 | 2 | 0.17 | 3.8 | MMBench [39] | ✗ | ✗ | ✗ |
| GenEval [21] | 6 | 6 | 553 | 2 | 0.26 | 4.4 | ScienceQA [40] | ✗ | ✗ | ✗ |
| DPG-Bench [26] | 5 | 13 | 1,000 | 2 | 0.165 | 3.8 | MMMU [71] | ✗ | ✗ | ✗ |
| UniBench (ours) | 13 | 81 | 1,234 | 5 | 0.458 | 1.4 | UniBench (ours) | ✓ | ✓ | ✓ |

2 Related Works

Unified Multimodal Understanding and Generation. Conventional generative models typically generate texts for understanding tasks via multimodal LLMs [38, 1, 3, 10], or generating images via diffusion models [50, 47, 6, 7, 31]. Some works [15, 18, 19, 20, 69] equip multimodal LLMs with pre-trained diffusion models for unified generation. More recent methods try to train end-to-end unified multimodal models [64, 30, 62, 8, 65, 66, 35, 61, 48, 73, 72] to reduce model redundancy and enhance instruction-following capabilities in visual generation. These unified models are rapidly emerging and attract much attention, such as DeepSeek Janus-Pro [8] and OpenAI GPT-4o [44] (native image generation, API not available yet). In this paper, we conducted extensive evaluations for both unified (Table 2) and generation methods (Table 3), except for some unavailable models.

Benchmarks. Evaluation of unified models typically involves multiple benchmarks for each tasks. For example, using benchmarks like ScienceQA [40], MMMU [71], etc [2, 28, 9, 54, 43, 42, 34, 39] to assess understanding capabilities, which rely on numerous images and labels. Our UniBench eliminates these dependencies, as a novel VQA benchmark for generated images. For the evaluation of generation models, image quality [24, 5, 67, 60] assessments on general image benchmarks [13, 36, 53] are widely used with other factors like alignment [23], fairness [32], style [46], etc [4]. Differently, unified models focus on instruction-following capabilities, making benchmarks like GenEval [21], DPG-Bench [26], T2I-CompBench++ [27], and other text-to-image evaluations [4, 63, 25, 33, 51, 17, 11] particularly relevant, with considered attributes like object, counting, colors, position, etc. Compared to these benchmarks, our UniBench evaluates many more aspects (see Table 1) to enhance the diversity, with greater difficulty and improvement potential. Most importantly, this is the first unified and elegant benchmark to evaluate both understanding and generation.

Evaluation Metrics. The accuracy for VQA base benchmarks [71, 2, 33, 21] is the most common metric to evaluate understanding models, with some NLG metrics [45, 14] for text generation tasks. Image generation quality assesment metrics like signal-to-noise ratio, FID [24], IS [5], ImageReward [67] are widely used in generation-only models [50, 47, 6, 7]. While unified models [8, 66, 64] focus more on the instruction-following capacity using the CLIPScore [23] or accuracy-based scores (e.g., VQAScore [21]) on compositional text-to-image (T2I) benchmarks [21, 27, 26, 11, 63]. Our UniScore is also an accuracy-based base score, while we provide multiple choices rather than a binary choice about keyword existence. This difference reduces the random error, and Fig. 4 suggests UniScore align well with human evaluations beyond other T2I metrics.

3 UniEval

UniEval is a novel unified evaluation framework featuring the holistic UniBench and the associated UniScore metrics, as shown in Fig. 2. We elaborate on this framework by detailing UniBench in Sec. 3.1 and describing the UniScore metrics in Sec. 3.2. Finally, we present a human study in Sec. 3.3, suggesting UniScore aligns well with human perception beyond other metrics.

3.1 UniBench

To achieve a unified evaluation, we need to construct a sufficiently diverse dataset that not only evaluates generation but also reflects various aspects of understanding ability. However, existing compositional benchmarks [21, 33, 27, 63] lack sufficient diversity; several attributes, such as color, objects, number, and position, are not enough to reflect a model’s understanding capabilities. Therefore, we have constructed a more holistic benchmark called UniBench via four steps.

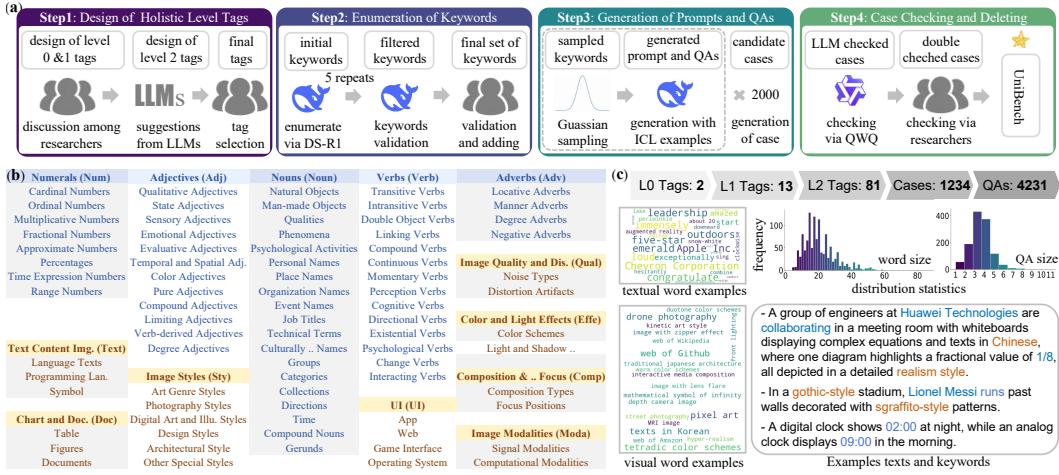


Figure 3: **UniBench.** (a). It is built by researchers with LLMs in four steps (details in Appendix C). (b). We designed holistic level-1 tags and level-2 tags with many novel attributes. (c). Details of UniBench, including data size, distribution of words and QAs, examples of keywords and prompts (see more keywords in Appendix B and prompts in Appendix C).

Step 1. Firstly, from the textual aspect, we selected five parts of speech suitable for image generation as level-1 tags. These include numerals, adjectives, nouns, verbs, and adverbs. From the visual perspective, we defined eight level-1 tags, including text content, chart and documents, image styles, image quality and distortion, color and lighting effects, composition and visual focus, image modality, and UI. Through collaboration between researchers and multiple large language models (LLMs), we further developed 81 hierarchical level-2 tags based on these 13 level-1 tags, as shown in Fig. 3b. These tags not only cover the attributes present in existing T2I benchmarks [21, 33, 27, 63] but also introduce many novel attributes, such as *time*, *emotions*, *celebrities*, *events*, *locations*, *actions*, *degrees*, *languages*, *symbols*, *programming*, *modalities*, *charts*, *figures*, *documents*, *UI*, *noise types*, *color schemes*, *lighting effects*, *composition*, *visual focus*, and other new attributes. Additionally, UniBench includes highly challenging reasoning tasks, like culturally specific names, which require reasoning about the region and personal appearances from names; and tasks requiring professional knowledge such as programming languages, operating systems, computational modalities, etc.

Step 2. We enumerated extensive keywords for each level-2 tag via Deepseek-R1-70B [22]. The enumeration is repeated 5 times, and unique keywords were dropped to ensure reliability. Then we used this LLM to select keywords that are drawable as Appendix C. Finally, we checked keywords manually with added ones to finalize the keyword set ($n = 3,285$). We show some keywords as shown in Fig. 3c and more example keywords of each level-2 tag in Appendix B.

Step 3. We bind keywords into prompts with corresponding questions and options in step 3. First, we randomly sample n level-2 tags using Gaussian sampling ($\text{std} & \text{mean} = 6$), along with four keywords in this tag as options. Then, we use Deepseek-R1-70B to select a suitable keyword from each option pair to generate suitable prompts with a corresponding question and given options. In this process, we set prompt generation and question generation as two separate tasks, providing detailed task

requirements, background, criteria, and several in-context examples. These requirements include ensuring the logical coherence of the prompts, ensuring that the keywords can be conveyed in the image, avoiding irrelevant content, and ensuring that questions can only be inferred from the image, etc. Please refer to more details and prompts to Appendix C.

Step 4. We filtered the 2,000 cases generated in Step 3 to ensure the quality of prompts and QAs (Q&A). First, we used another LLM, QWQ-32B [59], for validation involving prompt verification and QA verification. Each step included detailed backgrounds, requirements, criteria, and in-context examples. We instructed the model to ensure prompts are drawable without being overly complex. For questions, we required strict validation to ensure options were directly derived from the image and related to keywords. Detailed prompts are given in Appendix C. Ultimately, we manually validated to finalize 1,234 prompts and 4,231 QAs (see examples in Fig. 3c and Appendix C).

3.2 UniScore

This section focuses on the data flow of UniEval to introduce how we calculate the UniScore metrics. As shown in Fig. 2, we first use the unified model’s text-to-image capability to generate i images ($i = 4$) to reduce random error. Next, we utilize the model’s image-to-text understanding to answer q questions for each image, resulting in $i * q$ predictions. For each question, four options (A-D) belong to the same level-2 tag as the keyword, with an “N/A or Unknown” option for failed keyword generation. These five options help reduce random errors of binary options in other benchmarks [63, 11, 26, 21]. To enhance UniEval’s applicability, we support evaluating visual generation models with an extra Und. model in the same data flow as the unified model. Then, we analyze the understanding ability by the difference between unified and generation results, detailed in Appendix G.

After obtaining outputs of one case (true or false), we can calculate the accuracy of a case as a case-level UniScore. By averaging this across all 1,234 cases, we obtain the case macro UniScore. Additionally, we provide more case-level UniScores categorized by the number of words and QAs for analyzing the differences among few, middle, and many (middle words [15-23], middle QAs [3-4]). Once all questions ($n = 4,231$) have been processed, we get all outputs as \mathbf{o} ($n = 4 * 4, 231$). Then we aggregate outputs belonging to the same level-2 tag as \mathbf{o} and calculate its average accuracy to obtain the level-2 tag UniScore. By averaging the scores of level-2 tags under level-1, we derive the level-1 tag micro UniScore. Table 2 and Table 3 report the tag-level UniScores for all 13 level-1 tags, as well as their average as the final overall UniScore. We formulate the final UniScore as follows:

$$s = \frac{1}{n} \sum_{i=1}^n \mathbf{s}_i^1, \quad \mathbf{s}_i^1 = \frac{1}{m} \sum_{j=1}^m \mathbf{s}_j^2, \quad \mathbf{s}_j^2 = \frac{1}{k} \sum_{l=1}^k \mathbf{o}_l, \quad (1)$$

where s is the final UniScore, averaged from n level-1 tag UniScore \mathbf{s}^1 . Each level-1 score \mathbf{s}_i^1 is the mean from m under scores \mathbf{s}_j^2 of level-2 tags, averaged from k outputs \mathbf{o}_l (1 or 0) of a certain tag.

In addition to case-level and tag-level scores, we provide other analytical results, such as the distributions of response options to reflect model preferences and the invalid response rate out of A-E and corresponding keywords to assess format-following ability. Besides the averaged case scores, we also calculate the multiplied case scores to reflect the perfect cases UniScore, where all the image and questions are correct. All these metrics support holistic and in-depth analyses.

3.3 Human Evaluation

Setup. To prove the proposed metric is highly consistent with human evaluation in visual generation, we conducted a human study and compared UniScore with other metrics [23, 33] for subjective visual generation. Conversely, understanding is an objective task with certain QAs, thus, the human study is not conducted as common practice. To focus on analyzing generation without being affected by the weak understanding capabilities of the unified models, we introduced Qwen2.5-VL-7B [3] as the understanding model, which surpasses GPT-4v [1] on MMMU [71] (we also verified the 72B model in Appendix F). To ensure representativeness, we selected three models (Show-o [65], VILA-U [64], Janus-Pro-7B [8]) and randomly sampled 100 different cases from UniBench for each, totaling 300 cases. We recruited three annotators with different educational backgrounds (PhD-annotator 0, UG-annotator 1, master-annotator 2) to conduct independent labeling, with annotations covering 3 annotators * 300 random cases * 4 images in total. The annotators were asked to label whether the keywords were expressed in the generated images for each questions. Given the complexity of labels and the subjectivity in generation, we provided four labels: (1) generation failure, (2) between success

and failure, (3) successful generation, and (0) lacking knowledge to judge. Failures are scored as 0, successes as 1, while uncertain (score 2) and unknowing (score 4) labels are skipped. Both model results and human results associated with skipped labels are excluded from accuracy calculations to ensure the reliability of the evaluation (see human annotation examples in Appendix D).

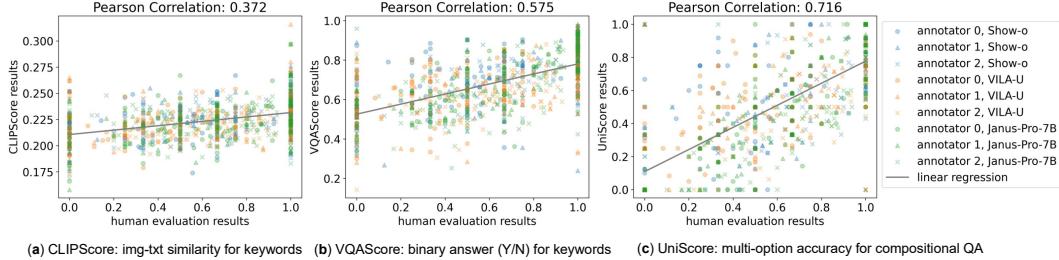


Figure 4: **Correlation with Human Evaluation.** x-axis indicates the accuracy for a case from humans, scores in the y-axes are from CLIPScore [23] (text-similarity), VQAScore [33] (binary option confidence), and the proposed UniScore (multiple options accuracy). Pearson correlation [12] measures the normalized covariance, and a higher value indicates closer alignment.

Analysis. For each case, we calculate its accuracy as the human score on the x-axis of Fig. 4. The y-axis includes comparison metrics, including the commonly used instruction-following metric, CLIPScore [23], and the recent VQAScore. These two metrics represent auto-evaluation metrics based on text similarity and keyword-based binary options (existence or not), respectively. In contrast, we employ an accuracy-based metric with multiple options, which has smaller random errors with non-template questions (providing more information to avoid ambiguity). We use Pearson correlation [12] to assess the alignment between human and auto evaluations. Since Pearson correlation is a normalized covariance, it is independent of score distribution without any threshold. The results indicate that the proposed UniScore has an average correlation of 0.716 with human annotators, much higher than CLIPScore’s 0.372 and VQAScore’s 0.575, demonstrating a strong alignment with human judgment. We also provide more analyses in Appendix E in aspects of annotators and models.

4 Experiments

4.1 Implementation

Evaluated Models. For the unified setting, we have implemented ten open-source models via pytorch from their official codes and weights, including VARGPT [73], TokenFlow [48], Show-o-Turbo [66], Show-o [65], Janus-Pro-1B [8], Janus-1.3B [62], VILA-U [64], UniToken-stageII [30], JanusFlow-1.3B [41], Janus-Pro-7B [8]. Some methods are not implemented due to unavailable models [61] or APIs [44], non-pytorch environment [37], or dependence on third-party generation models [15, 18, 19, 56, 20, 69, 57]. We also released implementations of ten visual generation models in our codebase, including SDv1.5 [50], SDv2.1 [50], PixArt- α [7], SDXL [47], FLUX.1-dev [31], SDv3.5-Medium [55], SDv3-Medium [16], FLUX.1-schnell [31], DALL-E3 [6], DALL-E2 [49].

UniEval. The implementation of UniBench has been described in Sec. 3.1, including used LLMs (Deepseek-R1-70B [22], QWQ-32B [59]), Gaussian sampling hyperparameters (std & mean = 6), with benchmark information (1,234 prompts and 4,231 QAs). The “LLMs” in step 1 involve POE, Gemini-1.5-Pro [58], Deepseek-R1-70B, Qwen2.5-72B [68] on webs. More detailed prompts are shown in Appendix C with example cases in Appendix J. UniEval generates 4 images for each prompt, which are combined with each question to calculate the UniScore as described in Sec. 3.2. When parsing responses, we prioritize matching letters A-E from a predefined format. If no match, use the last keyword found in the options. If neither exists, mark as invalid. Questions related to failed image generation also receive 0 scores. For details of human evaluation, like samples, models [66, 64, 8, 3], annotators, and compared metrics [23, 42], are given in Sec. 3.3 with annotated cases in Appendix D.

4.2 Unified Multimodal Understanding and Generation

Benchmarking Unified Models: We implement most of the open-source unified models [73, 48, 65, 66, 8, 62, 64, 30, 41] and report their results on UniEval in Table 2 as a leaderboard. We sort them by final UniScore and report the specific level-1 tag micro UniScores (see Appendix A for more detailed level-2 tag UniScores). Overall, the results range from 0.204 to 0.572, reflecting

sufficient differences and difficulty. Compared to the random accuracy of about 0.5 from the binary choice benchmark [21, 11, 27], our benchmark significantly reduces the random error (expected 0.2). Tags such as adjectives, nouns, and styles perform well, while numerals, text, documents, and UI meet greater challenges. For specific models, VARGPT [73] performs worst, it only outputs texts instead of the required images for many prompts, resulting in 52% invalid responses. The best performer is Janus-Pro-7B [8] at 0.572, whereas Janus-Pro-1B [8] performed worse than the earlier JanusFlow-1.3B [41]. This is due to Janus-Pro-1B’s poor format-following, often failing to output required formats, leading to 16.7% invalid responses. We discuss other anomalous results in detail in Sect. 4.4 with corresponding insights.

Table 2: UniEval Results. Fine-grained and overall UniScores on unified understanding and generation models. See tag names in Fig. 3a, level-2 scores in Appendix A, and analysis in Table 4.

| Model | Num | Adj | Noun | Verb | Adv | Text | Doc | Sty | Moda | Qual | Effe | Comp | UI | UniScore↑ |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| VARGPT [73] | 0.097 | 0.326 | 0.284 | 0.288 | 0.210 | 0.049 | 0.104 | 0.227 | 0.227 | 0.335 | 0.241 | 0.155 | 0.109 | 0.204 |
| TokenFlow [48] | 0.093 | 0.522 | 0.388 | 0.330 | 0.275 | 0.157 | 0.223 | 0.600 | 0.352 | 0.163 | 0.535 | 0.517 | 0.163 | 0.332 |
| Show-o-Turbo [66] | 0.250 | 0.302 | 0.353 | 0.274 | 0.256 | 0.381 | 0.331 | 0.386 | 0.331 | 0.546 | 0.394 | 0.360 | 0.398 | 0.351 |
| Show-o [65] | 0.250 | 0.362 | 0.422 | 0.316 | 0.285 | 0.381 | 0.358 | 0.390 | 0.346 | 0.472 | 0.432 | 0.360 | 0.398 | 0.367 |
| Janus-Pro-1B [8] | 0.186 | 0.504 | 0.443 | 0.413 | 0.370 | 0.174 | 0.233 | 0.536 | 0.503 | 0.396 | 0.397 | 0.350 | 0.301 | 0.370 |
| Janus-1.3B [62] | 0.202 | 0.484 | 0.497 | 0.384 | 0.284 | 0.246 | 0.319 | 0.641 | 0.381 | 0.408 | 0.476 | 0.423 | 0.449 | 0.400 |
| VILA-U [64] | 0.231 | 0.604 | 0.558 | 0.549 | 0.397 | 0.254 | 0.376 | 0.704 | 0.567 | 0.362 | 0.592 | 0.453 | 0.285 | 0.456 |
| UniToken-II [30] | 0.349 | 0.637 | 0.624 | 0.565 | 0.386 | 0.277 | 0.430 | 0.669 | 0.593 | 0.329 | 0.568 | 0.589 | 0.380 | 0.492 |
| JanusFlow-1.3B [41] | 0.324 | 0.608 | 0.588 | 0.528 | 0.423 | 0.400 | 0.354 | 0.706 | 0.645 | 0.521 | 0.585 | 0.496 | 0.426 | 0.508 |
| Janus-Pro-7B [8] | 0.356 | 0.716 | 0.666 | 0.621 | 0.509 | 0.456 | 0.477 | 0.777 | 0.672 | 0.542 | 0.655 | 0.527 | 0.459 | 0.572 |

UniEval for Task-specific Evaluations:

First, we validated that UniEval has a relatively strong consistency (Pearson correlation of 0.529) with the combination of task-specific evaluations (representative MMMU [71] + GenEval [21]) in Fig. 5a, demonstrating its rationality. We further calculated the coefficient of variation among models, showing that UniEval is twice that of task-specific evaluations, indicating a stronger discriminability (0.194 vs. 0.099). Analyzing outliers can reveal new insights from UniEval, such as the fact that Janus-Pro-1B [62] performs well independently, but its invalid response rate leads to poor unified results (see Table 4). Second, UniEval also supports task-specific evaluations. As shown in Fig. 5b, we introduce QWen2.5-VL-7B [3] to obtain generation-only results. Then, use the difference between unified and generation-only results to analyze understanding capabilities. A higher score on the y-axis indicates that the unified model’s understanding surpasses the extra understanding model. Among the models, JanusFlow-1.3B [41] demonstrates the best understanding of generated images, while many models struggled in understanding generated images, such as TokenFlow [48], Show-o [65], and Show-o-Turbo [66]. These results indicate that UniEval not only excels in unified evaluation but also supports task-specific evaluation, enabling detailed model analysis for further improvements.

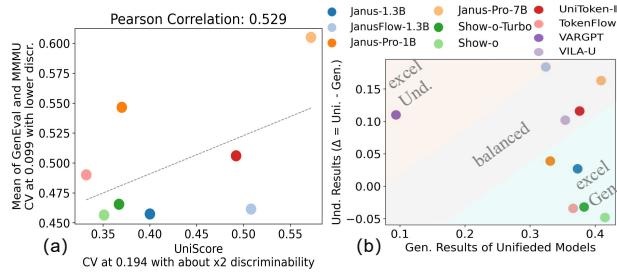


Figure 5: Task-specific Evaluation. (a): UniEval aligns with the average of MMMU [71] and GenEval [21], exhibiting twice the discriminability measured by the coefficient of variation (CV). (b): UniEval supports task-specific evaluations. “Gen. Results” are evaluated with QWen2.5-VL-7B [3]. “Und. Results” are from the difference between Uni. and Gen. results (see Appendix G), indicating preference on generation (blue region) or Und. in yellow.

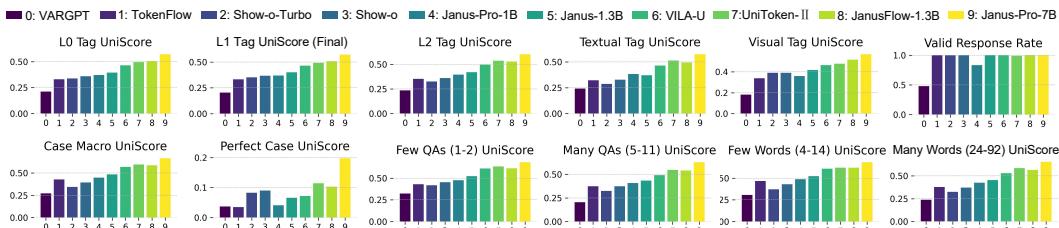


Figure 6: Comparison with Detailed Metrics. We illustrate results in more detailed metrics, including 5 tag-level scores on top with valid response rate, and 6 case-level scores on the bottom. Perfect indicates the all correct case ratio. Few and many are counted in varied QA and word sizes.

Comparison in More Aspects: We present various metrics in Fig. 6 for detailed analysis. The case macro UniScore and L1 tag UniScore align closely, but TokenFlow [48] works better at the case level over the tag level, owing to weak abilities in some fine-grained tags like Num. and UI. The perfect case UniScore (all questions correct for 4 images in a case) is challenging, peaking at about 0.2, while the second-tier models are around 0.1. The valid response rate indicates generating expected texts or images as required, with obvious errors in VARGPT [73] and Janus-Pro-1B [8]. Model rankings for various word and QA sizes correspond with case scores, showing reduced results for larger sizes.

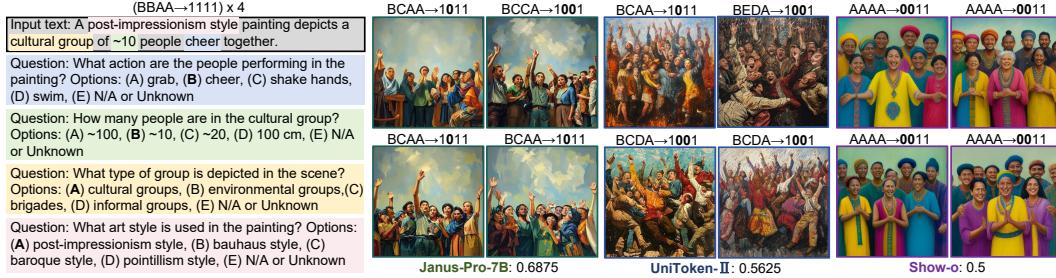


Figure 7: **Visual Comparision.** An example with the responses of three unified models [8, 30, 65]. "BCAA→1011" indicates answers are BCAA of 4 questions, with the second one being incorrect.

Case Studies: In addition to the examples in Fig. 2, we conducted a visual comparison of models in Fig. 7. For a case with four questions, we displayed the images generated by Janus-Pro-7B [8], UniToken-II [30], and Show-o [65], along with their corresponding answers and correctness results. Janus-Pro-7B [8] achieved the highest case-level UniScore of 0.6875, while Show-o [65] performed poorly due to biased responses (all A). Although UniToken-II produced images more similar to real images, it made obvious errors in visual generation regarding quantity and understanding of group types. This indicates that evaluation of instruction-following differs from image quality assessment, highlighting the unique value of UniEval. We also provide more case studies in Appendix J.

4.3 Text-to-Image Generation

Benchmarking T2I Generation Models:

The proposed UniBench not only evaluates unified models but also supports visual generation models. Since generation-only models lack understanding capabilities, we introduced Qwen2.5-VL-7B [3] for automatic evaluation. This model outperforms GPT-4v [1] on MMMU [71] and aligns better with humans than the 72B model in Appendix F. We conducted extensive evaluations on 10 popular models in Table 3, including Stable Diffusion series [50, 47, 55, 16], PixArt- α [7], FLUX series [31], and DALL-E series [49, 6]. Results show that the earlier SDv1.5 [50] had the lowest UniScore of 0.33, while recent models like DALL-E [6], FLUX [31], and SDv3 [16] performed well around 0.5. Instruction-following and image quality are not always aligned; for example, DALL-E3 is slightly lower than DALL-E2, because the prompt augmentation of DALL-E3 makes the prompt more complex with strict safety control. Similarly, FLUX.1-dev and SDv3.5-Medium sacrificed some instruction-following ability on this challenging benchmark. These insights are analyzed in Table 4. In Appendix G, we compared UniScore for unified models using the same understanding model, showing that higher resolution benefits SoTA visual generation models in complex scenes, highlighting the need for higher resolution in unified models. Higher UniScores in Table 2 indicate that unified models have a stronger understanding of generated images.

Table 3: **UniBench for Visaul Generation.** UniBench is versatile and capable of evaluating text-to-image models. Note that measured instruction-following is different from image quality assessment, providing different insights. Bold and underline indicate the best and second, respectively.

| Model | Num | Adj | Noun | Verb | Adv | Text | Doc | Sty | Moda | Qual | Effe | Comp | UI | UniScore↑ |
|----------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| SDv1.5 [50] | 0.133 | 0.453 | 0.376 | 0.274 | 0.223 | 0.065 | 0.236 | 0.673 | 0.430 | 0.238 | 0.504 | 0.505 | 0.176 | 0.330 |
| SDv2.1 [50] | 0.139 | 0.479 | 0.427 | 0.305 | 0.253 | 0.101 | 0.261 | 0.678 | 0.422 | 0.266 | 0.533 | 0.519 | 0.247 | 0.356 |
| PixArt- α [7] | 0.166 | 0.550 | 0.442 | 0.348 | 0.268 | 0.065 | 0.271 | 0.729 | 0.456 | 0.233 | 0.624 | 0.616 | 0.154 | 0.379 |
| SDXL [47] | 0.149 | 0.562 | 0.461 | 0.365 | 0.311 | 0.106 | 0.294 | <u>0.752</u> | 0.512 | 0.354 | 0.626 | 0.547 | 0.210 | 0.404 |
| FLUX.1-dev [31] | 0.270 | 0.591 | 0.470 | 0.410 | 0.321 | 0.260 | 0.459 | 0.625 | 0.453 | 0.135 | 0.589 | 0.621 | 0.307 | 0.424 |
| SDv3.5-Medium [55] | 0.261 | 0.609 | 0.529 | 0.421 | 0.295 | 0.318 | 0.534 | 0.718 | 0.520 | 0.346 | 0.631 | 0.547 | 0.522 | 0.481 |
| SDv3-Medium [16] | 0.289 | 0.581 | 0.539 | 0.461 | 0.331 | 0.378 | 0.596 | 0.670 | 0.529 | 0.314 | 0.622 | 0.568 | <u>0.555</u> | 0.495 |
| FLUX.1-schnell [31] | <u>0.345</u> | 0.642 | <u>0.562</u> | 0.451 | 0.364 | 0.305 | <u>0.624</u> | 0.717 | 0.529 | 0.190 | 0.644 | 0.606 | 0.644 | 0.509 |
| DALL-E3 [6] | 0.312 | 0.650 | 0.545 | 0.489 | 0.376 | <u>0.375</u> | 0.627 | 0.734 | 0.616 | 0.444 | 0.680 | <u>0.632</u> | 0.499 | <u>0.537</u> |
| DALL-E2 [49] | 0.369 | 0.624 | 0.605 | 0.474 | 0.360 | 0.406 | 0.610 | 0.762 | 0.587 | 0.362 | 0.668 | 0.690 | 0.527 | 0.542 |

Comparison with Other Benchmarks: Our UniBench is not only the first unified benchmark, but it also outperforms existing benchmarks in evaluating text-to-image (T2I) models. In Fig. 8, we compare T2I benchmarks from three perspectives: difficulty (error rate of the best model), discriminability (coefficient of variation), and diversity (min-max normalized number of attributes). For ConceptMix [63], we report its K=1 scores, as K>1 (multiplied accuracy) differs from others. Since T2I-CompBench++ uses multiple models to evaluate attributes and lacks an overall value, we report the complexity results based on GPT-4v [1] (see specific models and data in Appendix H). Experimental results show that UniBench significantly surpasses existing benchmarks in terms of difficulty and diversity, ranking second in discriminability (due to higher difficulty), with an overall score of 0.961, significantly exceeding the second, GenEval of 0.466.

4.4 Insights and Analysis

Studied from Table 2, Table 3, Fig. 5, Fig. 6, and Appendix H, we conclude several interesting insights as shown in Table 4 with cues and analyses. Key insights include: 1. Models like VARGPT [73] may fail to generate images (52% invalid responses); 2. Models like Janus-Pro-1B [8] sometimes fail to follow the formats, with 16.7% of responses outside A-E; 3. Show-o [65] performs well with additional understanding models (Appendix H), but its own understanding model outputs very biased responses, with 89.2% yielding A, leading to low unified results; 4. Models often struggle in some attributes, such as Num. (quantity and time); 5. UniBench introduced many new attributes, such as visual series labels; 6. Janus-Pro-7B [8] demonstrates good self-consistency and understanding of generated images, thus achieving a high UniScore; 7. Recent visual generation models yield higher UniScores beyond the unified model using the same understanding model because of higher resolution (1024 vs. 224-512), emphasizing the importance of resolution in complex prompts. 8. The extra models [3] tend to be stricter, generally scoring lower than unified models. 9. There is a trade-off between image quality and instruction-following; some earlier models like DALL-E2 [49] may outperform newer models like DALL-E3 [6], owing to the prompt augmentation in DALL-E3 (introducing extra content). Overall, UniEval provides valuable insights reflecting its unique values.

Table 4: **Insights.** UniEval provides valuable insights when evaluating unified and visual generation models, with the cues and corresponding analysis.

| Insights | Model | Analysis | Cues |
|--------------------|-------------------|------------------------------------------------------|----------------------------------------------|
| Failure generation | VARGPT [73] | Often output texts when generating images | 52% invalid response |
| Weak formatting | Janus-Pro-1B [8] | Often output response deviated from formats | 16.7% responses out of ABCDE |
| Biased response | Show-o [65] | Ranks 1 in Appendix G but only 7th in Table 2 | 89.2% responses in A e.g., quantity (Num) |
| Weak abilities | Uni. & Gen. | Some tags are not generated well | e.g., challenging web, language, and table |
| Visual aspects | Uni. & Gen. | UniBench provides extensive visual tags | Only 7% “N/A or Unknown” |
| Self-consistency | Janus-Pro-7B [8] | Good understanding for self-generated images | Table 3 vs. Appendix G |
| Crucial resolution | Uni. vs. Gen. | Higer resolutions bring higher UniScore | More “N/A or Unknown” |
| Strict criteria | Qwen-2.5VL-7B [3] | Extra model is strict and may reduce UniScore | e.g., DALL-E2 vs. DALL-E3 |
| Trade-off | Gen. models | Better quality may not enhance instruction-following | |

5 Conclusion

In conclusion, we proposed UniEval, the first evaluation framework for unified multimodal understanding and generation models. By addressing the limitations of existing task-specific benchmarks, UniEval eliminates the reliance on extra models and images, enabling an elegant, simplified, and overall evaluation. The involved benchmark, UniBench, provides the most holistic fine-grained attributes beyond existing generation benchmarks, with the UniScore metric aligned well with human evaluations. Through extensive evaluations on SoTA unified and visual generation models, UniEval offers many valuable insights reflecting its unique values. As the field continues to evolve, UniEval stands out as a pioneering tool that can foster further advancements in multimodal understanding and generation. As the first unified evaluation framework, UniEval sets a new standard for unified multimodal evaluation and well supports visual generation fields, fostering continued progress in generative AI. Besides the significances of UniEval, we also discussed the limitations in Appendix I.

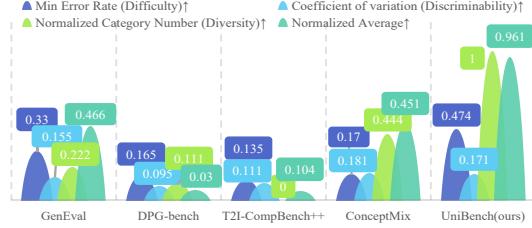


Figure 8: **Advantages Over T2I Benchmarks.** The proposed UniBench shows better difficulty and diversity beyond existing T2I benchmarks [21, 26, 27, 63]. See specific data in Appendix H.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [3] Shuai Bai, Kebin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.
- [5] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.
- [6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf*, 2(3):8, 2023.
- [7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- [8] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [10] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.
- [11] Jaemin Cho, Yushi Hu, Jason M Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *ICLR*, 2024.
- [12] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [14] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.
- [15] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.
- [16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [17] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*.
- [18] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023.

- [19] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.
- [20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.
- [21] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.
- [22] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, 2021.
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [25] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.
- [26] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.
- [27] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [28] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [29] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [30] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. *arXiv preprint arXiv:2504.04423*, 2025.
- [31] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [32] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.
- [33] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301, 2024.
- [34] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [35] Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. *arXiv preprint arXiv:2501.00289*, 2024.
- [36] 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 *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [37] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv e-prints*, pages arXiv–2402, 2024.

- [38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [40] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [41] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.
- [42] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [43] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.
- [44] OpenAI. Gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2025.
- [45] Kishore Papineni et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [46] 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. *arXiv preprint arXiv:2406.16855*, 2024.
- [47] 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.
- [48] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.
- [49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.
- [54] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.
- [55] stability.ai. Sdv3.5. <https://stability.ai/news/introducing-stable-diffusion-3-5>, 2025.
- [56] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*.

- [57] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [58] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [59] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025.
- [60] Yu Tian, Yue Liu, Shiqi Wang, and Sam Kwong. Quality assessment for text-to-image generation: A survey. *IEEE MultiMedia*, 2025.
- [61] Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *arXiv preprint arXiv:2412.06673*, 2024.
- [62] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.
- [63] Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *arXiv preprint arXiv:2408.14339*, 2024.
- [64] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024.
- [65] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.
- [66] Chenkai Xu, Xu Wang, Zhenyi Liao, Yishun Li, Tianqi Hou, and Zhijie Deng. Show-o turbo: Towards accelerated unified multimodal understanding and generation. *arXiv preprint arXiv:2502.05415*, 2025.
- [67] 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. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.
- [68] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [69] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024.
- [70] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [71] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [72] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- [73] Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*, 2025.

Appendix

Contents

| | | |
|----------|-----------------------------------------------------------|-----------|
| 1 | Introduction | 1 |
| 2 | Related Works | 3 |
| 3 | UniEval | 4 |
| 3.1 | UniBench | 4 |
| 3.2 | UniScore | 5 |
| 3.3 | Human Evaluation | 5 |
| 4 | Experiments | 6 |
| 4.1 | Implementation | 6 |
| 4.2 | Unified Multimodal Understanding and Generation | 6 |
| 4.3 | Text-to-Image Generation | 8 |
| 4.4 | Insights and Analysis | 9 |
| 5 | Conclusion | 9 |
| A | Results Comparision Among Level-2 Tags | 15 |
| B | Examples of Keywords | 17 |
| C | Details of UniBench Construction | 20 |
| D | Human Evaluation Cases | 23 |
| E | Human Evaluation Analysis | 24 |
| F | Comparision with Larger Extra Model | 25 |
| G | UniEval for Task-specific Evaluations. | 26 |
| H | Comparison with T2I Benchmarks. | 27 |
| I | Limitation and Broader Impacts | 28 |
| J | Case Study | 28 |

A Results Comparison Among Level-2 Tags

Besides level-1 results in Table 2 and Table 3, we present more detailed results for level-2 tags in this section. Fig. 9 compares the unified multimodal models [73, 48, 66, 65, 8, 62, 64, 30, 41], where results with yellow backgrounds indicate better-performing attributes. Overall, adjectives, nouns, and style-related level-2 tags perform well, such as natural objects, man-made objects, and compound nouns, achieving many results above 0.9. Numbers, texts, UI, and other related level-2 tags are more challenging, with many models performing below 0.3, and TokenFlow [48] even scoring 0 in the programming language. From the model view, Janus-Pro-7B [8] performs the best, while VARGPT [73] performs the worst.

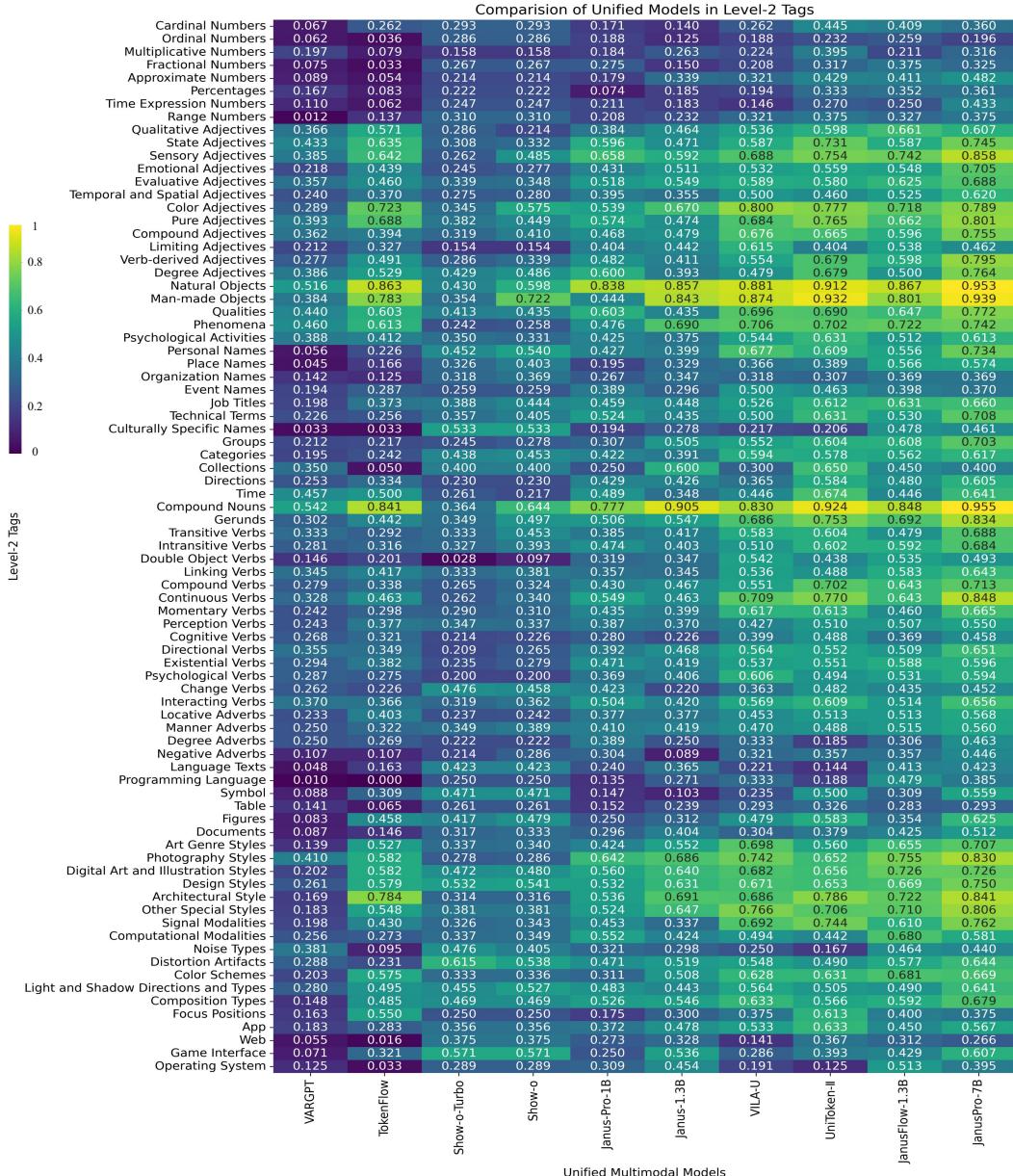


Figure 9: Results of Unified Models Evaluated on Level-2 Tags.

Fig. 10 compares visual generation models [50, 7, 47, 31, 55, 16, 6, 49], and the overall performances are similar to those of the unified models, but the differences between high and low performances are more pronounced. For example, natural objects have a minimum score of 0.869, significantly higher than the unified models' lowest score of 0.43. In contrast, for culturally specific names that require reasoning, the highest score for generation-only models is only 0.156, whereas the unified models reach a maximum of 0.533. This indicates that visual generation models perform better for common attributes (benefiting from higher resolution), while unified models excel in instruction-following under complex conditions.

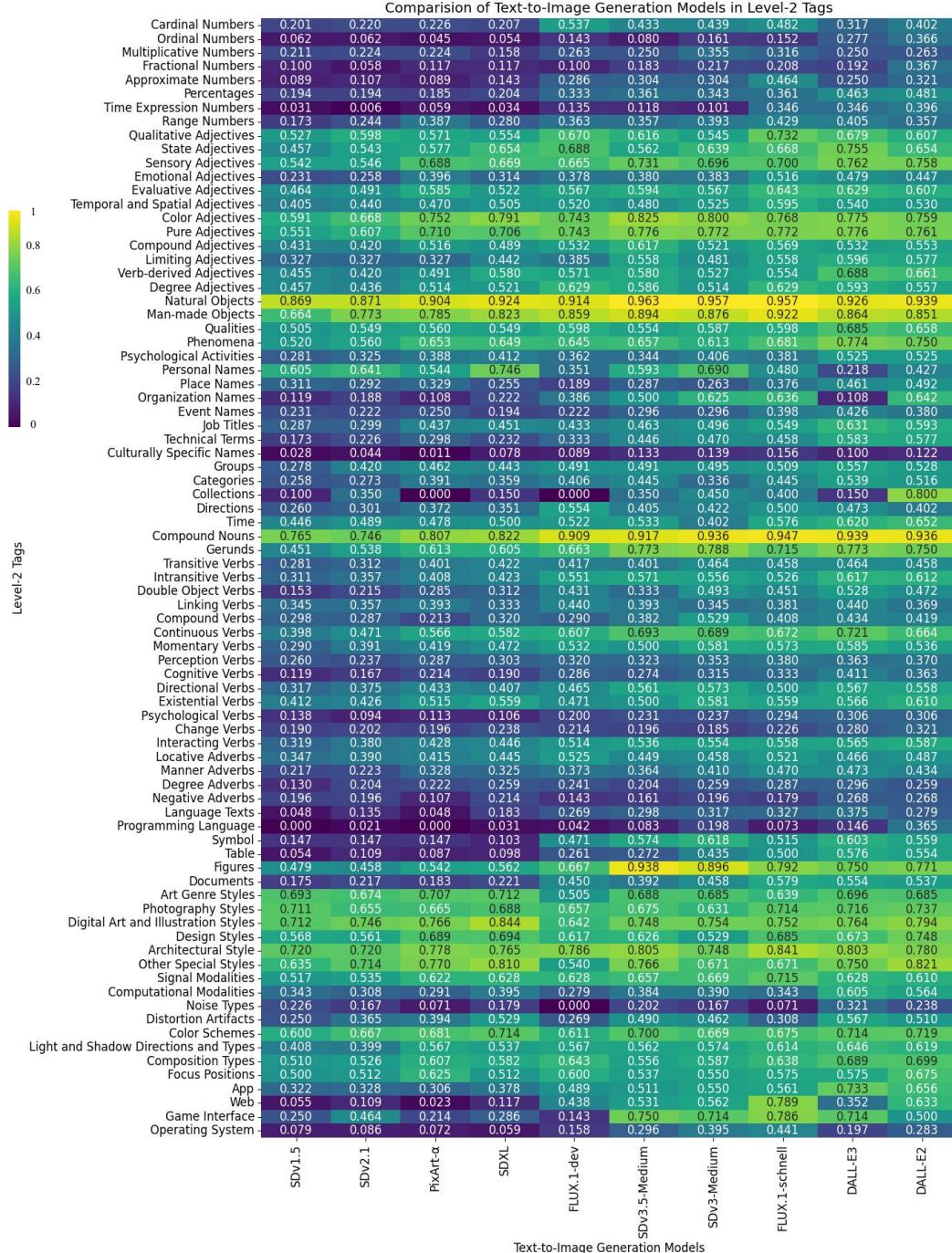


Figure 10: Results of Visual Generation Models Evaluated on Level-2 Tags.

B Examples of Keywords

We have provided keyword examples for each level-2 tag as shown in Fig. 11, Fig. 12, and Fig. 13. These keywords are used to bind prompts and generate questions. UniBench includes a wealth of detailed attributes to ensure diversity in evaluation. These tags include many novel attributes, such as time, emotions, celebrities, events, locations, actions, degrees, languages, symbols, programming, modalities, charts, figures, documents, UI, noise types, color schemes, lighting effects, composition, visual focus, and other new attributes.

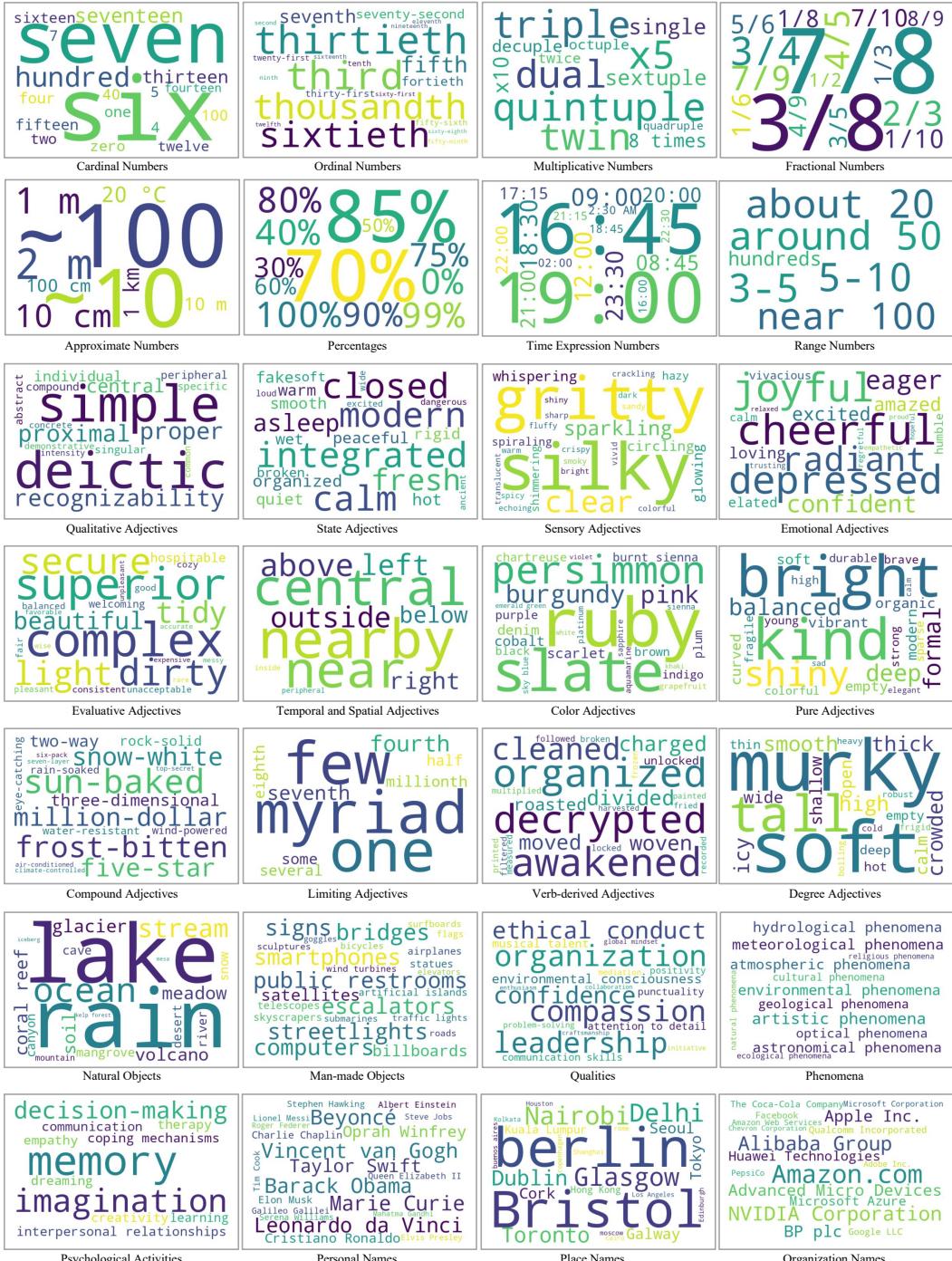


Figure 11: Example of keywords. For each level-2 tag, we displayed up to 15 keywords through the word cloud, with random sampling and text size.



Figure 12: **Example of keywords.** For each level-2 tag, we displayed up to 15 keywords through the word cloud, with random sampling and text size.



Figure 13: **Example of keywords.** For each level-2 tag, we displayed up to 15 keywords through the word cloud, with random sampling and text size.

C Details of UniBench Construction

We have detailed the construction of UniBench in Sec. 3.1. In this section, we highlight the prompts used for invoking LLMs and some details on processing the samples. In step 1 (Fig. 14), we employed four LLMs to construct level-2 tags, including POE, Gemini-1.5-Pro [58], Deepseek-R1-70B [22], and Qwen2.5-72B [68] on webs. The prompts involved include “What are the subcategories of [a specific Level-1 Tag]?” and “Thoroughly and systematically classify [a specific Level-1 Tag]:”. After collecting outputs from the LLMs, tags, and prompts, we performed deduplication and then manually selected appropriate level-2 tags. Our considerations included “Can this tag be generated by the image model?”, “Is this label reasonable and not duplicated?”, and “What other reasonable attributes are applicable?”.

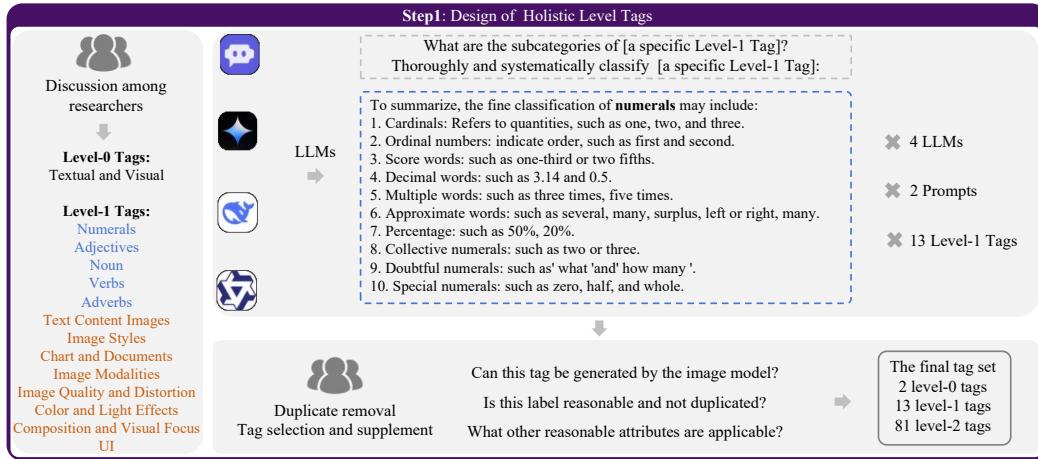


Figure 14: **Step 1 of UniBench Construction.** The gray box indicates the used prompts with answers marked by the blue box.

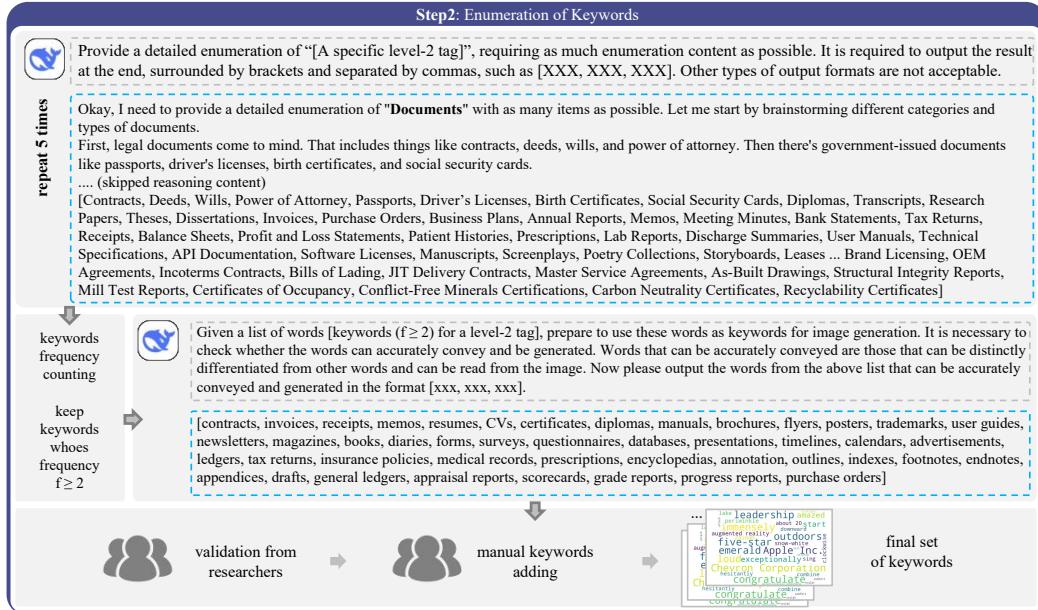


Figure 15: **Step 2 of UniBench Construction.** The gray box indicates the used prompts with answers marked by the blue box. The used LLM is Deepseek-R1-70B [22].

As shown in Fig. 15, in the second step, we used Deepseek-R1-70B [22] to enumerate keywords for level-2 tags. The prompt used was “Provide a detailed enumeration of “[A specific level-2 tag]”,

requiring as much enumeration content as possible. It is required to output the result at the end, surrounded by brackets and separated by commas, such as [xxx, xxx, xxx]. Other types of output formats are not acceptable.”. We required the model to run five times, retaining only keywords that appeared at least twice to minimize random errors. Then, we used the prompt to verify the keywords: “Given a list of words [keywords ($f \geq 2$) for a level-2 tag], prepare to use these words as keywords for image generation. It is necessary to check whether the words can accurately convey and be generated. Words that can be accurately conveyed are those that can be distinctly differentiated from other words and can be read from the image. Now please output the words from the above list that can be accurately conveyed and generated in the format [xxx, xxx, xxx].”, ensuring they were suitable for image generation. Finally, we conducted a manual review, adding appropriate keywords, and ultimately determined the final set of keywords.

In the third step, we used a Gaussian sampling to randomly take N tags with four keywords from each tag as options, constructing the prompt input for the upper part of Fig. 16. Combining this with the prompts for the lower part, we instructed Deepseek-R1-70B to choose keywords and generate sentences and questions. We defined two tasks in the prompts, clearly outlining the criteria. After the model produced structured outputs, we parsed them and obtained 2,000 initial cases.

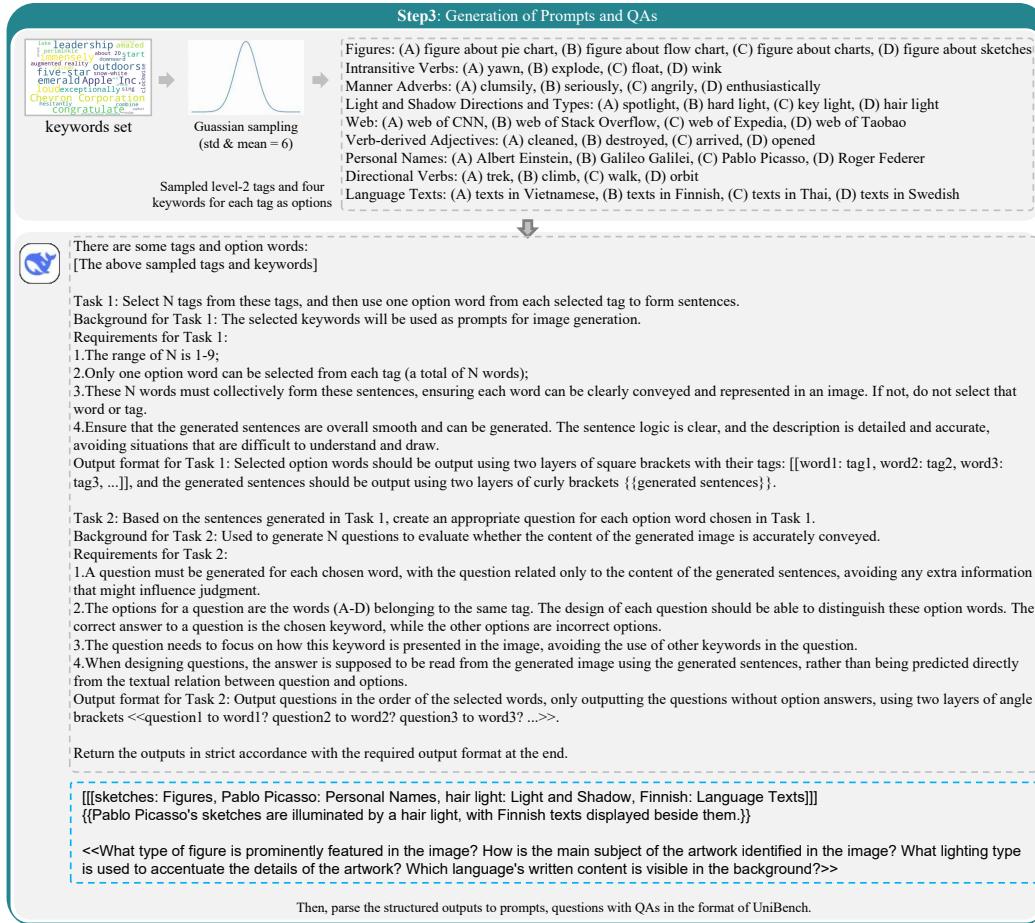


Figure 16: Step 3 of UniBench Construction. The gray box indicates the used prompts with answers marked by the blue box.

To further ensure the quality of the benchmark, we conducted validation as shown in Fig. 17. First, we introduced another LLM, QWQ-32B [59], to verify the initial prompts generated by Deepseek-R1-70B [22]. Our requirement was “Suitable text meets the following criteria: a) the content can be illustrated and accurately conveyed, b) the text is clear with no logical or linguistic errors, c) avoid very complex scenes and excessive references, d) avoid contradictory and hard-to-understand word combinations, e) allow combinations of unrelated objects or scenes as long as they can be accurately

conveyed. f) need to conduct a strict selection, only very certain text is regarded as suitable.”. We also provided five positive and five negative examples as in-context examples to assist the LLM in making judgments. Prompts for visual generation that did not meet the requirements were discarded, while those that did proceeded to the next step of question validation. We also used QWQ-32B to check questions with requirements: “Now, you are required to check whether the design of this question is reasonable. A reasonable question meets the following criteria: a) The answer can only be inferred from the generated image and cannot be directly chosen from the question. b) The question set must relate to the options and the given text. c) The question should not involve too much irrelevant text. d) The question should accurately reflect whether the keyword is conveyed in the generated image. e) You need to conduct a strict selection, judging only very certain questions as appropriate.”. We also supplied three positive and three negative examples as in-context examples. Finally, we calculated the ratio r of failed questions in each case. If r was less than $1/3$, we deleted the unsuitable questions; otherwise, we considered the effective keywords too few and skipped the entire case. Subsequently, we conducted a manual review using the same criteria to ensure the quality of prompts and questions. Ultimately, we confirmed 1,234 prompts and 4,231 QAs.

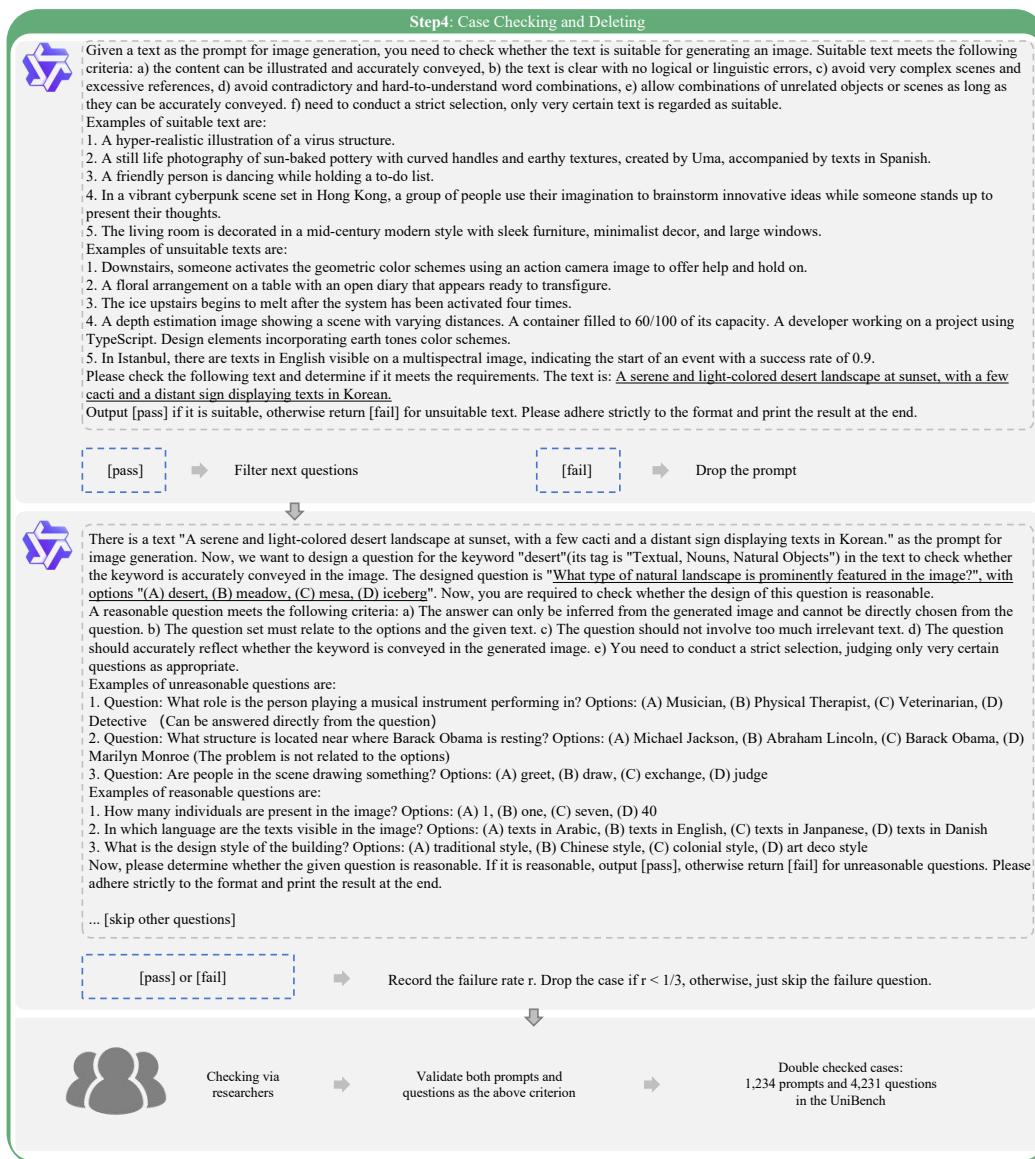


Figure 17: **Step 4 of UniBench Construction.** The gray box indicates the used prompts with answers marked by the blue box. The used LLM is QWQ-32b [59].

D Human Evaluation Cases

In Sec. 3.3, we have introduced the criterion of human evaluation: “The annotators were asked to label whether the keywords were expressed in the generated images for each question. Given the complexity of labels and the subjectivity in generation, we provided four labels: (1) generation failure, (2) between success and failure, (3) successful generation, and (0) lacking knowledge to judge”. In this section, we provide some visualized annotation results to help readers understand our annotation process. Fig. 18 shows annotation examples from annotator 1 (Undergraduate background) on Janus-Pro-7B [8]. We list the prompt and questions on the left for reference with corresponding generated images on the right of this figure. The annotations are colored the same as the image borders, including failure (1), uncertain (2), success (3), and unknown (0). These two examples are complex, where annotator 1 annotated four types of labels in the first case, and marked more unknown labels for the second one.

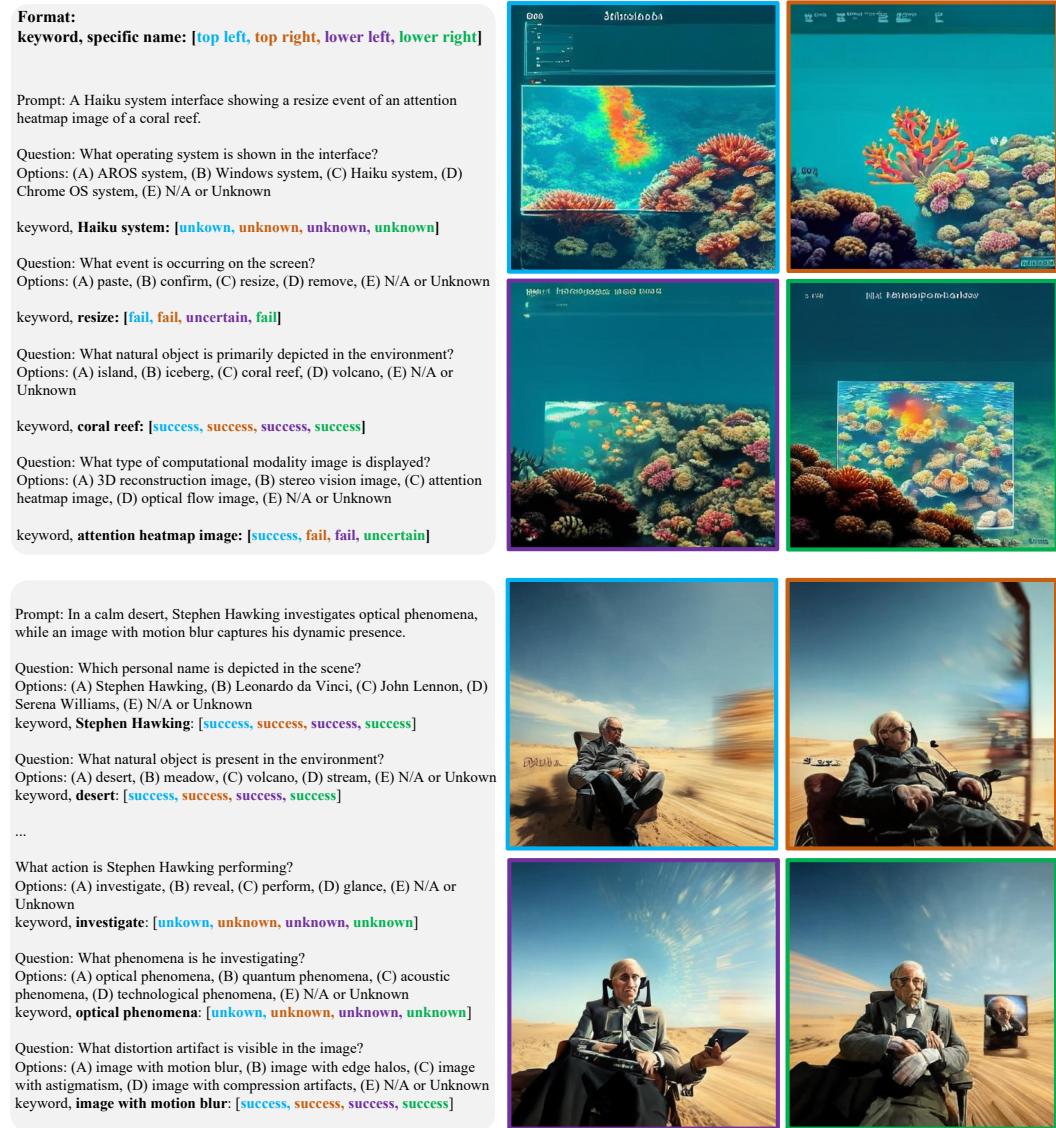


Figure 18: **Visualization of Annotations.** The prompts, questions, and labeled results are listed on the left, with generated images on the right. The label colors correspond to the image with the same color board. The uncertain label is related to ambiguous generation, while the unknown means unable to judge. These two labels are excluded from the evaluation to ensure reliability. These results are from annotator 1 (undergraduate background) on Janus-Pro-7B [8].

E Human Evaluation Analysis

In addition to the overall human study in Fig. 4, we conducted human studies in the aspect of varied annotators (Fig. 19) and different models (Fig. 20) using the same 300 random cases from UniBench. Fig. 19 shows the correlation results from different annotators. The first row presents the results from annotator 0 (PhD background), with the lowest overall scores, indicating a stricter evaluation criterion and a lower tendency to label responses as unknown. The second row comes from annotator 1 (undergraduate background), who performed the best overall, with a correlation of 0.777 between UniScore and human evaluations. The results from annotator 2 (master background) are between other annotators. From the three different annotators, we found that UniScore is consistently higher than the recent VQAScore [33] and significantly exceeds the commonly used instruction-following metric, CLIPScore [23].

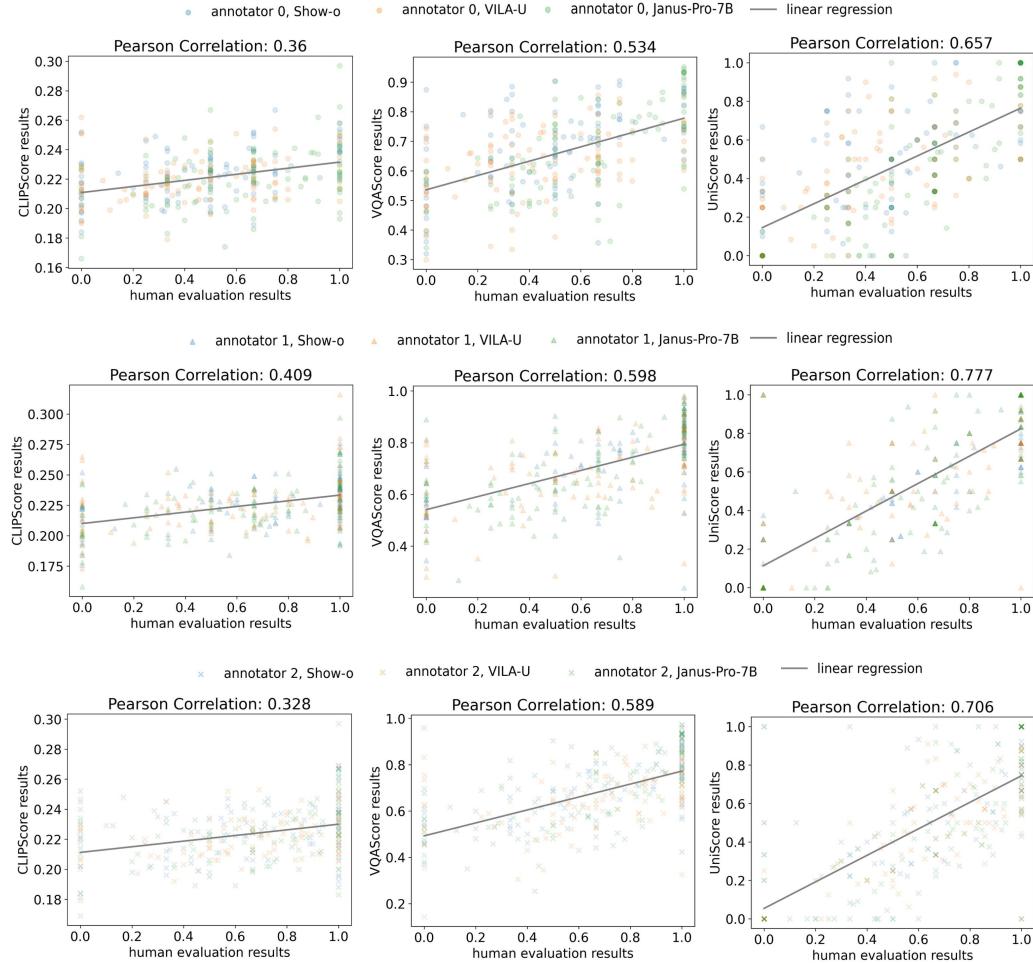


Figure 19: Human Study in the Annotator Aspect. The first, second, and third rows indicate human studies from annotator 0 (PhD), 1 (undergraduate), and 2 (master), respectively. The Pearson Correlation [12] is a normalized covariance to measure the alignment between auto evaluation metrics (CLIPScore [23], VQAScore [33], our UniScore) and human evaluations.

Fig. 20 compares the correlation of various metrics with human evaluations across different models. The results in the first row come from Show-o [65], marked in blue; the second row is from VILA-U [64], in orange; and the third row presents the results from Janus-Pro-7B [8]. Consistent with the overall results and the annotator aspect, the proposed UniScore shows superior alignment compared to existing metrics across different models, surpassing VQAScore [33] by 0.134, 0.21, and 0.095, respectively. This is attributed to UniScore providing more options and relevant prompts, which can reduce random errors beyond binary options while offering correlational information from the prompt and options related to keywords, thereby avoiding ambiguity. In contrast, CLIPScore is based solely

on keyword similarity, while VQAScore relies on prompt templates. The experiments indicate that UniScore is a robust instruction-following metric that aligns well with human perception.

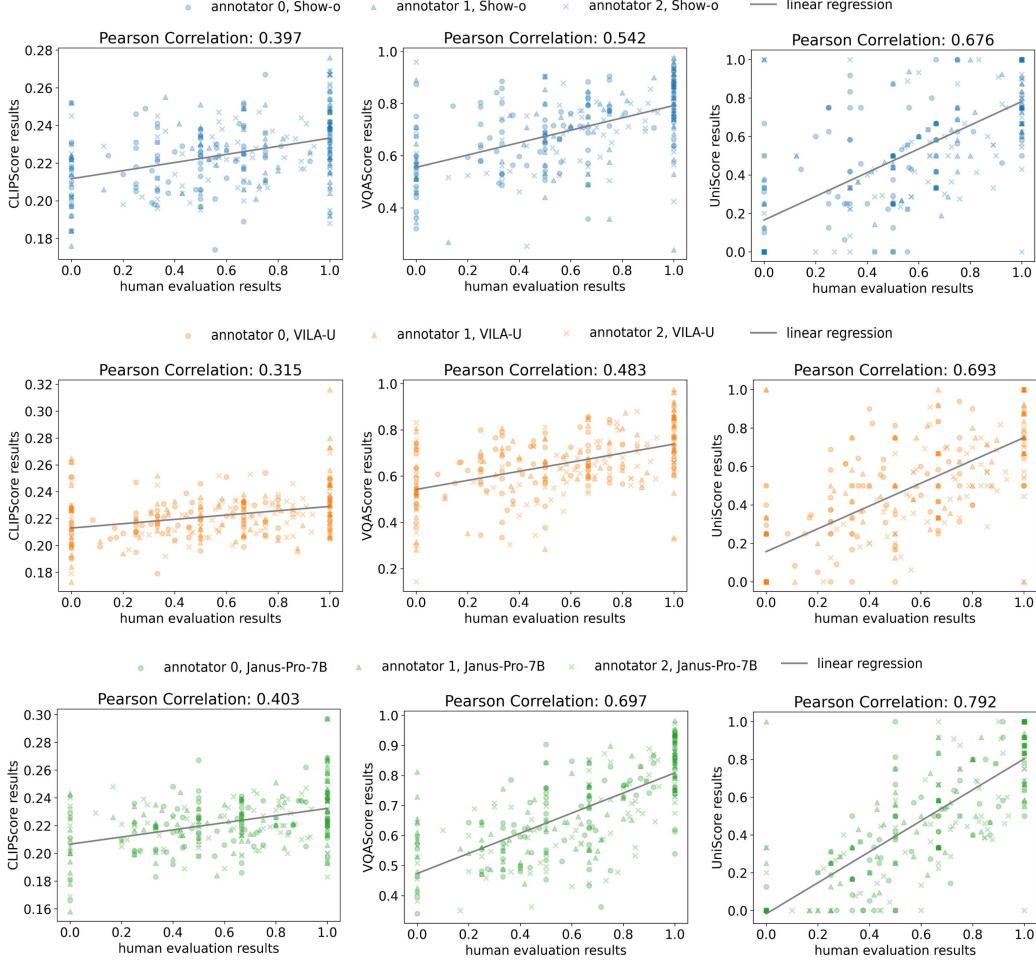


Figure 20: Human Study in the Model Aspect. The first, second, and third rows indicate human studies conducted on Show-o [65], VILA-U [64], and Janus-Pro-7B [8], respectively. The Pearson Correlation [12] is a normalized covariance to measure the alignment between auto evaluation metrics (CLIPScore [23], VQAScore [33], our UniScore) and human evaluations.

F Comparision with Larger Extra Model

In human study and visual generation evaluation, we introduced an extra model, Qwen2.5-VL-7B [3]. In this analysis, we also compared it to the larger Qwen2.5-VL-72B [3] to explore the relationship between model scale and human alignment. It should be noted that both models outperformed the closed-source GPT-4v on MMMU [71], while the 72B model’s performance is closer to GPT-4o. We avoided using closed-source models to mitigate the usage costs for followers, and Qwen2.5-VL-7B is currently the strongest 7B model on MMMU.

In Fig. 21, we compared the alignment of Qwen2.5-VL-7B and Qwen2.5-VL-72B with humans on UniBench (using the same models and sampled cases as human evaluations). The results show that the performance of both models is similar, with the 7B model having slightly better alignment. This is mainly because the criteria for the 72B model are stricter; for example, in the evaluation of Janus-Pro-7B, the UniScore for the 72B model is 0.388, while for the 7B model, it is 0.402. Compared to the 72B model, human expectations for the success of the generation are not as strict, which leads to better alignment for Qwen2.5-VL-7B. Additionally, the 7B model has lower memory usage, significantly reducing the hardware requirements for the visual generation part. Note that the

Unified model does not require additional understanding of the model; most models can run on a single 24GB memory GPU, while visual generation model evaluations generally require two 24GB GPUs with an extra understanding model.

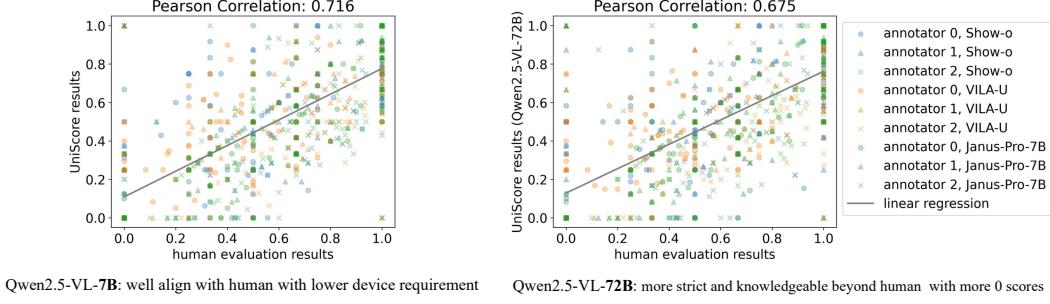


Figure 21: Comparision with Larger Model in Alignment. The alignment degrees in the human study (measured by Pearson correlation) are close between the 7B and 72B models. Since Qwen2.5-VL-72B is stricter (lower UniScore when evaluating the same model) than the 7B model, its alignment degree is slightly lower than the 7B model. Qwen2.5-VL-7B [3] is the best 7B model on MMMU [3] currently, and saves API cost compared with close-set models, as well as GPU memory compared with larger models.

G UniEval for Task-specific Evaluations.

In this section, we independently evaluate the generation and understanding abilities of unified models. This helps researchers to analyze the model strengths and weaknesses on specific tasks, providing more insights besides the overall results. It also proves the wide applicability of UniEval.

Table 5: Evaluation for the Visual Generation Part of Unified Models. We apply the Qwen2.5-VL-7B [3] to evaluate the task-specific visual generation performances. By comparing unified models with visual generation models using the same model, we find new insights discussed in Sec. 4.4.

| Model | Num | Adj | Noun | Verb | Adv | Text | Doc | Sty | Moda | Qual | Effe | Comp | UI | UniScore↑ |
|---------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----------|
| VARGPT [73] | 0.036 | 0.168 | 0.114 | 0.074 | 0.066 | 0.010 | 0.035 | 0.114 | 0.116 | 0.120 | 0.169 | 0.164 | 0.032 | 0.094 |
| JanusFlow-1.3B [41] | 0.141 | 0.483 | 0.354 | 0.267 | 0.234 | 0.104 | 0.186 | 0.608 | 0.328 | 0.340 | 0.539 | 0.480 | 0.146 | 0.324 |
| Janus-Pro-1B [8] | 0.126 | 0.478 | 0.334 | 0.243 | 0.237 | 0.135 | 0.213 | 0.611 | 0.442 | 0.340 | 0.575 | 0.486 | 0.085 | 0.331 |
| VILA-U [64] | 0.148 | 0.548 | 0.367 | 0.337 | 0.258 | 0.036 | 0.208 | 0.660 | 0.392 | 0.272 | 0.609 | 0.608 | 0.159 | 0.354 |
| TokenFlow [48] | 0.178 | 0.529 | 0.405 | 0.292 | 0.284 | 0.136 | 0.249 | 0.660 | 0.401 | 0.242 | 0.574 | 0.589 | 0.224 | 0.366 |
| Janus-1.3B [62] | 0.174 | 0.523 | 0.387 | 0.312 | 0.306 | 0.120 | 0.262 | 0.650 | 0.387 | 0.418 | 0.618 | 0.541 | 0.148 | 0.373 |
| UniToken-II [30] | 0.172 | 0.538 | 0.434 | 0.361 | 0.266 | 0.109 | 0.301 | 0.601 | 0.366 | 0.348 | 0.626 | 0.552 | 0.215 | 0.376 |
| Show-o-Turbo [66] | 0.180 | 0.569 | 0.452 | 0.344 | 0.268 | 0.103 | 0.271 | 0.660 | 0.390 | 0.351 | 0.662 | 0.529 | 0.197 | 0.383 |
| Janus-Pro-7B [8] | 0.194 | 0.584 | 0.464 | 0.358 | 0.267 | 0.167 | 0.391 | 0.682 | 0.462 | 0.322 | 0.637 | 0.562 | 0.230 | 0.409 |
| Show-o [65] | 0.202 | 0.611 | 0.483 | 0.397 | 0.290 | 0.126 | 0.419 | 0.659 | 0.453 | 0.355 | 0.653 | 0.469 | 0.280 | 0.415 |

For the evaluation of visual generation, we introduced the same understanding model Qwen2.5-VL-7B [3] as visual generation models in Table 3, focusing on comparing its visual generation part in a fair setting. It can be seen that Show-o [65] achieves the highest pure generation capability at 0.415, while the unified model scores 0.367 in Tab. 2, indicating good generation quality but limited in understanding. Show-o ranks first in Table 5 mainly because of a relatively high resolution (512), while other unified models typically use resolutions of 224 or 336. However, compared to the pure generation models in Table 3, the understanding models overall demonstrate weaker generation capabilities owing to lower resolution. For instance, among ten models, only two unified models have a UniScore exceeding 0.4 under understanding models, whereas seven pure generation models exceed 0.4. The ones that do not exceed 0.4 are primarily those with a resolution of 512 (models with results higher than 0.4 are all 1024 in resolution). This indicates that high resolution is very important in complex generation scenarios, and the unified model is lacking in this regard. Another reason is that the unified model must balance generation and understanding, which may lead to shortcomings in generation performance. Through the visual comparisons in Appendix J, we find that there is still room for improvement in its generation quality. In another aspect, we believe unified models are promising. At the same resolution, 512, the unified Show-o still outperforms generation-only PixArt- α [7], suggesting the potential of the unified model.

For evaluating the model’s understanding ability, using an extra generation model is not appropriate because existing models perform inadequately. Our approach is to compare the overall results of the

Table 6: Evaluation for the Understanding Part of Unified Models. We evaluate the understanding ability of the unified model itself by the difference between the overall results (Table 2) and the understanding results in Table 5. The results are $\Delta = Uni. - Gen.$, measuring the understanding ability difference between the unified model and the extra model. A positive value indicates that the unified model outperforms the extra model (Qwen2.5-VL-7B [3]) in understanding generated images, highlighting understanding as its strength. A negative value indicates that the understanding of the unified model is weaker than that of the extra model, highlighting understanding as its weakness. The best understanding result is marked in green, while the lowest value is marked in red.

| Model | Num | Adj | Noun | Verb | Adv | Text | Doc | Sty | Moda | Qual | Effe | Comp | UI | UniScore↑ |
|---------------------|---------------|---------------|---------------|---------------|---------------|--------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| VARGPT [73] | 0.061 | 0.158 | 0.17 | 0.214 | 0.144 | 0.039 | 0.069 | 0.113 | 0.111 | 0.215 | 0.072 | -0.009 | 0.077 | 0.11 |
| TokenFlow [48] | -0.085 | -0.007 | -0.017 | 0.038 | -0.009 | 0.021 | -0.026 | -0.06 | -0.049 | -0.079 | -0.039 | -0.072 | -0.061 | -0.034 |
| Show-o-Turbo [66] | 0.07 | -0.267 | -0.099 | -0.07 | -0.012 | 0.278 | 0.06 | -0.274 | -0.059 | 0.195 | -0.268 | -0.169 | 0.201 | -0.032 |
| Show-o [65] | 0.048 | -0.249 | -0.061 | -0.081 | -0.005 | 0.255 | -0.061 | -0.269 | -0.107 | 0.117 | -0.221 | -0.109 | 0.118 | -0.048 |
| Janus-Pro-1B [8] | 0.06 | 0.026 | 0.109 | 0.17 | 0.133 | 0.039 | 0.02 | -0.075 | 0.061 | 0.056 | -0.178 | -0.136 | 0.216 | 0.039 |
| Janus-1.3B [62] | 0.028 | -0.039 | 0.11 | 0.072 | -0.022 | 0.126 | 0.057 | -0.009 | -0.006 | -0.01 | -0.142 | -0.118 | 0.301 | 0.027 |
| VILA-U [64] | 0.083 | 0.056 | 0.191 | 0.212 | 0.139 | 0.218 | 0.168 | 0.044 | 0.175 | 0.09 | -0.017 | -0.155 | 0.126 | 0.102 |
| UniToken-II [30] | 0.177 | 0.099 | 0.19 | 0.204 | 0.12 | 0.168 | 0.129 | 0.068 | 0.227 | -0.019 | -0.058 | 0.037 | 0.165 | 0.116 |
| JanusFlow-1.3B [41] | 0.183 | 0.125 | 0.234 | 0.261 | 0.189 | 0.296 | 0.168 | 0.098 | 0.317 | 0.181 | 0.046 | 0.016 | 0.28 | 0.184 |
| Janus-Pro-7B [8] | 0.162 | 0.132 | 0.202 | 0.263 | 0.242 | 0.289 | 0.086 | 0.095 | 0.21 | 0.22 | 0.018 | -0.035 | 0.229 | 0.163 |

unified model with those of the visual generation model. This ensures that the generation model is the same, allowing a fair and reasonable comparison between the understanding ability of the unified model and that of the extra understanding model. As shown in Table 6, we compare the unified overall results with the results based on the extra model, calculating the difference $\Delta = Uni. - Gen.$, which measures the difference in understanding ability between the unified model and the extra model. A positive value indicates that the unified model outperforms the extra model (Qwen2.5-VL-7B [3]) in understanding generated images, highlighting understanding as its strength. A negative value indicates that the unified model’s understanding is weaker than that of the extra model, highlighting understanding as its weakness. In the table, the best results for understanding are marked in green, while the worst results are marked in red. We observe that Janus-Flow-1.3B [41] and Janus-Pro-7B [8] have the best understanding of generated images. In Sec. 4.4, we refer to this ability as self-consistency, meaning the model’s capability to accurately understand the images it generates itself. The models with the largest understanding bias are Show-o [65], Show-o-Turbo [66], and TokenFlow [48]. The reasons for this are also analyzed in Table 4, mainly attributed to bias in the understanding model’s responses. This reflects that the understanding ability of these models needs improvement. Additionally, although VARGPT [73] has three green marks, this model often refuses to generate images, resulting in an overall low score with limited reference value. Overall, the scores of unified models are higher than those of the extra models, with most values being positive. This indicates that unified models generally have better understanding abilities for generated images than understanding-only models, demonstrating the unique value of unified models.

H Comparison with T2I Benchmarks.

We compared the difficulty, discriminability, and diversity of across text-to-image benchmarks in Fig. 8. In this section, we reports the specific data involved in Table 7. Due to differet evaluated models reported by different benchmarks, we selected five commonly used models for comparison, including SDV1 [50], SDV2 [50], PixArt- α [7], SDXL [47], and DALL-E3 [6]. Among these, only T2I-CompBench++ [27] uses SDv1.4 and SDv2.0; the others are SDv1.5 and SDv1.4. We applied min-max normalization to the quantity for a normalized overall result, which was counted at the level-1 tags, as some benchmarks do not contain fine-grained labels. Among them, T2I-CompBench++ is evaluated by multiple models and does not have an overall metric; we report its complexity metric evaluated by GPT-4v [1]. Based on this data, we quantified the comparisons in various aspects. Fig. 8 indicates that our UniBench significantly outperforms existing benchmarks, with an overall value of 0.961, notably exceeding the second-best GenEval [21] of 0.466.

Table 7: **Specific Data in Benchmark Comparison.** Different benchmarks use diverse models, thus, we pick these five common models for fair comparison. The category number is min-max normalized to count the normalized average.

| Models / Aspects | GenEval [21] | DPG-bench [26] | T2I-CompBench++ [27] | ConceptMix [63] | UniBench (ours) |
|---------------------------------------------|--------------|----------------|----------------------|-----------------|-----------------|
| SDv1 [50] | 0.43 | 0.6318 | 0.6453 | 0.52 | 0.33 |
| SDv2 [50] | 0.5 | 0.6809 | 0.6483 | 0.52 | 0.355 |
| PixArt- α [7] | 0.48 | 0.7111 | 0.7223 | 0.66 | 0.379 |
| SDXL [47] | 0.55 | 0.7465 | 0.717 | 0.69 | 0.404 |
| DALL-E3 [6] | 0.67 | 0.835 | 0.8653 | 0.83 | 0.526 |
| Min Error Rate (Difficulty) | 0.33 | 0.165 | 0.135 | 0.17 | 0.474 |
| Coefficient of variation (Discriminability) | 0.155 | 0.095 | 0.111 | 0.181 | 0.171 |
| Category Number (Diversity) | 6 | 5 | 6 | 7 | 13 |
| Normalized Average (Overall) | 0.466 | 0.03 | 0.104 | 0.451 | 0.961 |

I Limitation and Broader Impacts

Although UniEval, as the first evaluation framework designed for unified models, addresses many limitations of existing task-specific benchmarks, limitations still objectively exist. First, our evaluation framework emphasizes instruction-following and does not include image quality assessment following most text-to-image benchmarks [27, 63, 21, 26]. If we add evaluate metrics like FID [24], additional images and models would need to be introduced, which goes against our motivation. Second, UniEval emphasizes overall evaluation, only partially achieving individual assessment. Although our UniBench, combined with extra models [3], shows significant advantages in evaluation of visual generation compared to conventional text-to-image benchmarks [27, 63, 21, 26], the current approach to analyzing understanding ability is limited to comparing overall results and visual generation results as discussed in Appendix G. We attempted to fix the visual generation model to generate a fixed dataset for evaluating understanding ability, but due to the limitations of existing model generation capabilities, we cannot directly evaluate understanding ability without human efforts to select the correct generated images. Third, ensuring the quality of the benchmark still incurs human effort. Although we do not need annotators to label ground truth, the LLMs sometimes generate overly complex and unsuitable prompts. The quality of questions can occasionally be poor, such as when answers from options appear in the questions. This necessitates the introduction of a certain level of manual checking costs to ensure quality (still far less than directly annotating answers).

Opportunities and challenges coexist. As the first unified evaluation framework, there is still much room for improvement. For example, it is possible to incorporate more diverse content, such as image quality assessments, with minimal additional resources. Alternatively, followers could select specific generation models and include manual screening to create a benchmark for generated images, accommodating both overall and task-specific evaluations. Additionally, enhancing the quality and quantity of synthesis could enable more detailed evaluations, such as requiring specific textual content beyond just language type. We believe that unified evaluations will be a great pathway to achieve simplified, comprehensive, convenient, and high-quality evaluations. This approach also has strong potential to generalize to future unified models encompassing video, audio, and other capabilities, establishing a new standard for multimodal model evaluation, as well as inspiring the development of more powerful models and applications.

J Case Study

We visualized example cases of UniEval in Fig. 22 and Fig. 23 with analysis of insights. On the left side of the image, we showcase the visual generation prompts, multimodal understanding questions, options, and inference prompts. In the bottom, we present sample answers and evaluation results for each question. Finally, we emphasize the insights corresponding to this case. On the right side, there are four images generated by the model; the top of the images corresponds to the model’s outputs, and whether they are correct. At the bottom right, we provide specific model names and case-level UniScore.

From Fig. 22, we find some visual tags are very challenging, e.g., the programming language. Besides, a model with good self-consistency can achieve high scores. The second case in this figure shows that some models are biased in response (almost answer A in Show-o [65]). Thus, hurt the overall results. Besides, tags like number require accurate both visual generation and understanding, which is challenging. The case also tells us that UniBench requires the visual reasoning ability, such as the culturally specific name (Quan). Moreover, UniEval enables task-specific evaluation to analyze

each part of unified models, where the generation ability of Show-o is good, but the understanding is wrong. From Fig. 23, we find some models like Janus-Pro-1B[8] may not follow the format and output some invalid responses (the “?” on the image top). The second case shows that visual generation using a complex prompt is challenging. These examples provide templates for the case study, which help researchers to investigate the models and foster further improvements.

Prompt:
Galileo Galilei is programming in JavaScript next to a spacecraft.

QAs:
Question: Who is the person in the image?
Options: (A) Galileo Galilei, (B) Stephen Hawking, (C) Mahatma Gandhi, (D) Winston Churchill, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What programming language is being used?
Options: (A) programming in Apex, (B) programming in COBOL, (C) programming in JavaScript, (D) programming in Rust, (E) N/A or Unknown...

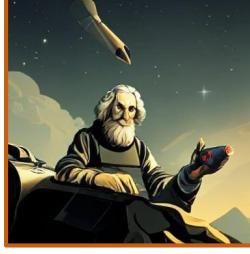
Answer :
ACC (QA1,QA2,QA3) →111 (ture / false)

Insight:

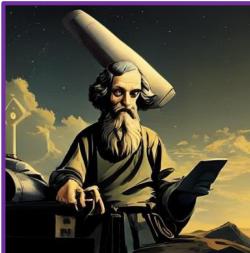
- Some visual tags are very challenging, e.g., the programming language.
- Model with good self-consistency can achieve high scores



AEC→101



AEC→101



AEC→101



AEC→101

Prompt:
In the sumi-e style painting, around 50 birds fly gracefully in the background while Quan sits meditatively by a pond.

QAs:
Question: What artistic style is used in the painting?
Options: (A) art deco style, (B) hyperrealism style, (C) sumi-e style, (D) screen printing style, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What is the name of the person sitting by the pond?
Options: (A) Ahmad, (B) Mateo, (C) Irena, (D) Quan, (E) N/A or Unknown...

Question: Approximately how many birds are flying in the background?
Options: (A) 1-2, (B) around 50, (C) 5-10, (D) hundreds, (E) N/A or Unknown ...

Answer :
CDB (QA1,QA2,QA3) →111 (ture / false)

Insight:

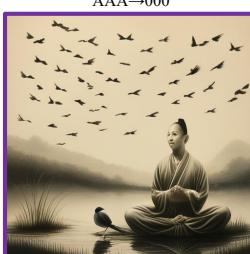
- Some models are biased in response (almost answer A in Show-o), thus hurt the overall results.
- Some tags are challenging, e.g., the number requires both accurate visual generation and understanding.
- UniBench also requires the visual reasoning ability, such as the culturally specific name (Quan).
- UniEval enables task-specific evaluation to analyze each part of unified models.



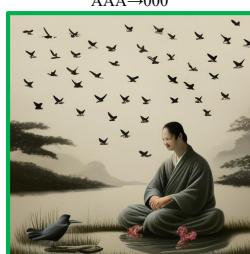
AAA→000



AAA→000



AAA→000



AAA→000

Show-o: 0
Show-o (Gen. only): 0.667 (CEB→101 *4)
Show-o (Und. Δ): -0.667 (Und. is weak.)

Figure 22: . Insight Analysis From Cases of Unified Models. Some visual tags are very challenging, e.g., the programming language, some attributes also require the visual reasoning ability, such as the culturally specific name (Quan). We find model with good self-consistency can achieve high scores. While some models are biased in response (almost answer A in Show-o), thus, hurt the overall results. Besides, UniEval enables task-specific evaluation to analyze each part of unified models as in the case of Show-o [65]. The case information is shown on the left, including prompt, questions, options, answers, and insights. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case. The Gen. only score and Und. score is calculated as introduced in Appendix G.

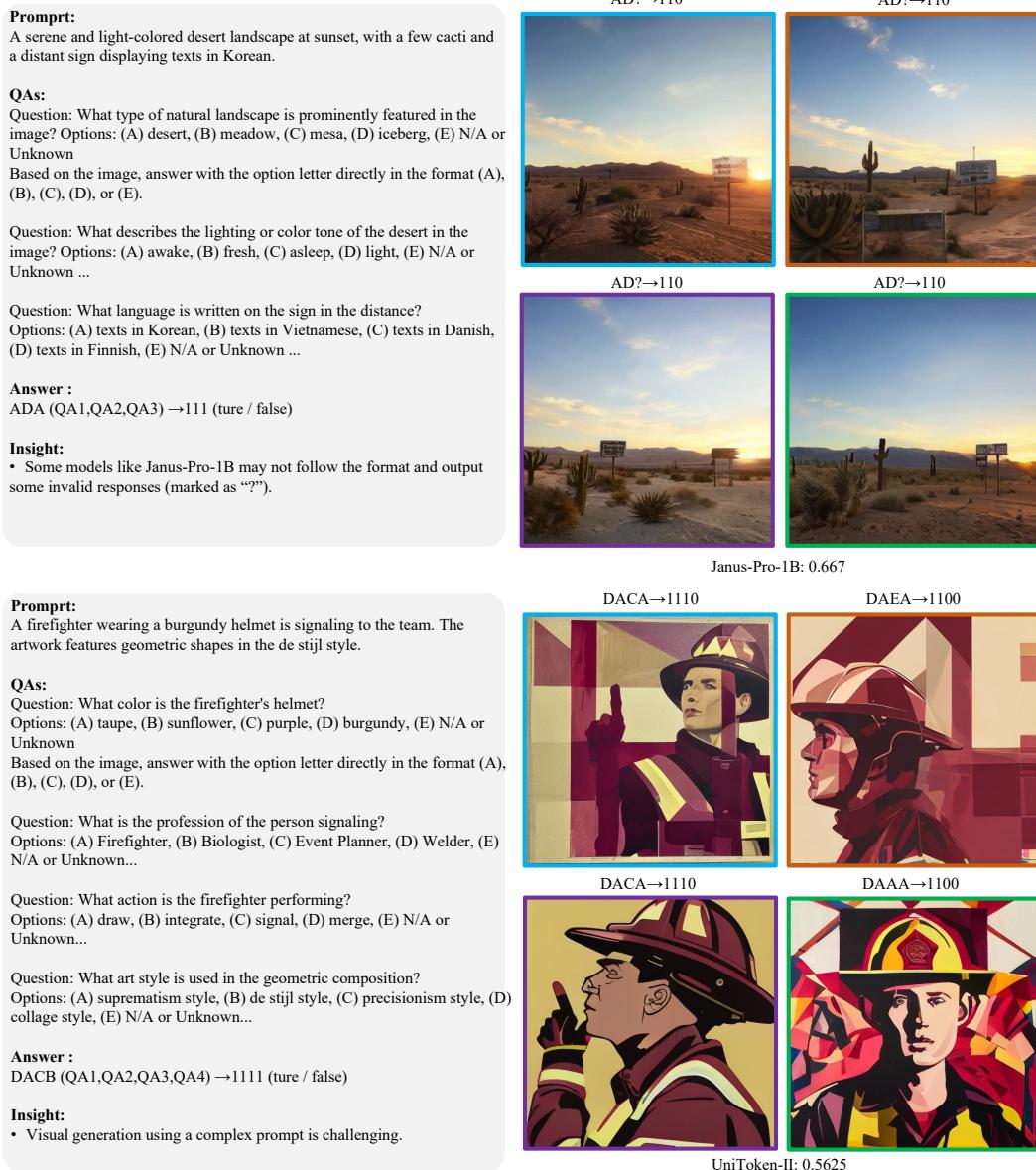


Figure 23: . **Insight Analysis From Cases of Unified Models.** Some models like Janus-Pro-1B [8] may not follow the format and output some invalid responses (marked as “?”). Besides, visual generation using a complex prompt is challenging. The case information is shown on the left, including prompt, questions, options, answers, and insights. The generated images are depicted on the right, with corresponding model predictions on the image top (“?” indicates an invalid response out of A-E). The model name is listed on the bottom right, with the UniScore of this case.

In Fig. 24, we also analyze the cases of the visual generation models. We can see that the generation-only models are of high quality, especially with fewer flaws in the details. However, we find that there is a trade-off between image quality and instruction following. Better quality in visual generation models may not enhance instruction-following. For example, elements like digital, 02:00, and 09:00 are not accurately fulfilled, even though the image quality is good with fine details. These findings suggest that instruction-following is not a simple task, and further improvements are necessary, as measured by this challenging and diverse benchmark.

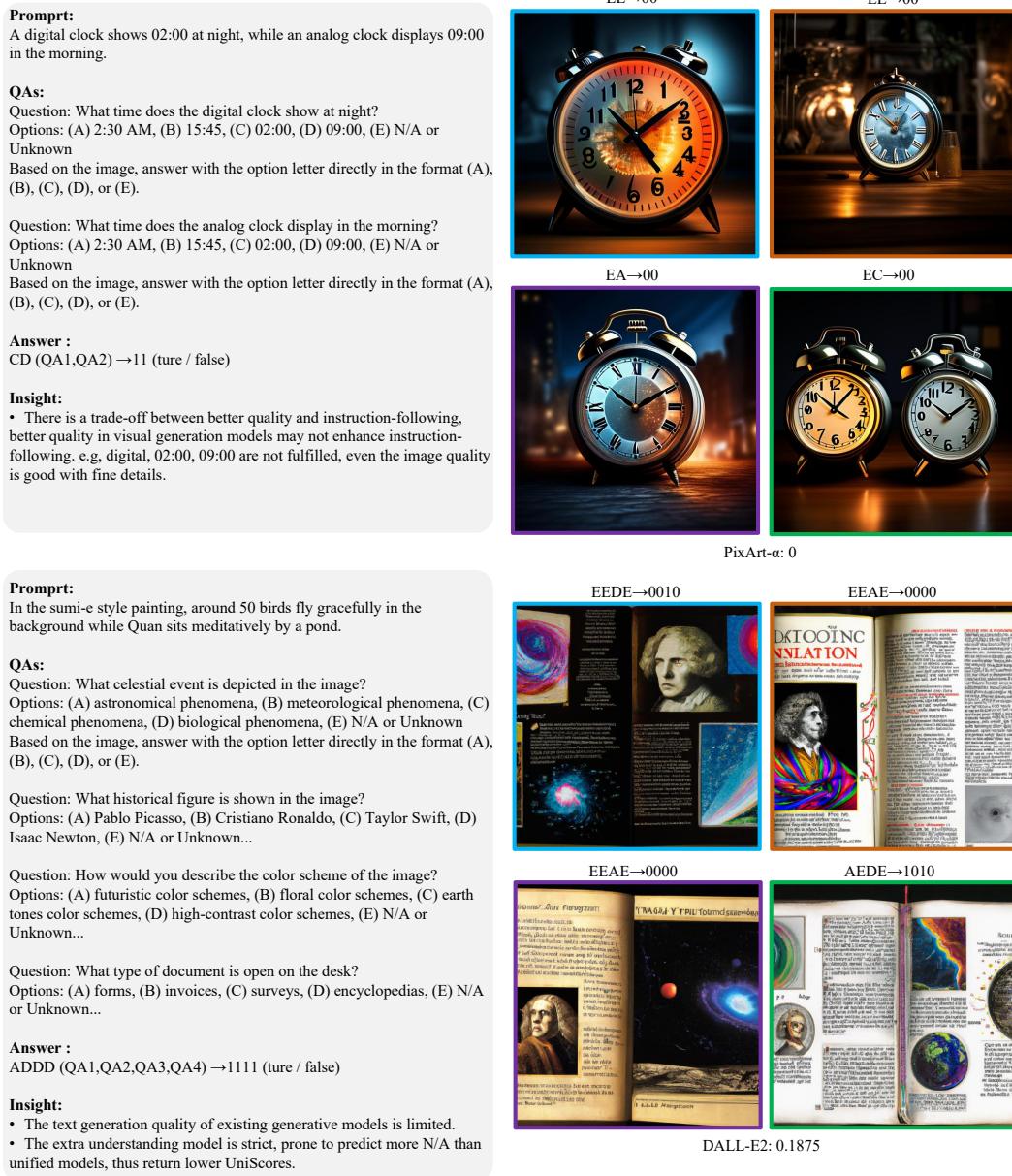


Figure 24: . Insight Analysis From Cases of Visual Generation Models. There is a trade-off between better quality and instruction-following; better quality in visual generation models may not enhance instruction-following. Besides, the text generation quality of existing generative models is limited. The extra understanding model is strict; prone to predict more N/A than unified models. The case information is shown on the left, including prompt, questions, options, answers, and insights. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case.

We further showcased the challenging attributes and failure cases. As shown in Fig. 25, the instruction-following capability poses significant challenges for some models like VARGPT [73], leading to difficulties in generating controllable images. We found that certain tags, such as emotion, programming language, and numbers are particularly challenging. Additionally, we observed that unified models tend to guess answers, resulting in a higher UniScore compared to strict extra evaluation models. However, under the same additional models, their overall performance still lags behind visual generation models (see Appendix G). Examples in Fig. 26 also support this

claim. In cases where visual generation models produce better quality than unified models while the understanding models tend to output N/A. Furthermore, some labels, like technical terms, remain challenging for state-of-the-art visual generation models. We also found that the safety checks of DALLE-3 [6] can reduce model performance, while these safety checks exhibit inconsistent judgments for the same prompt (sometimes it is valid, sometimes prone to refuse generation).

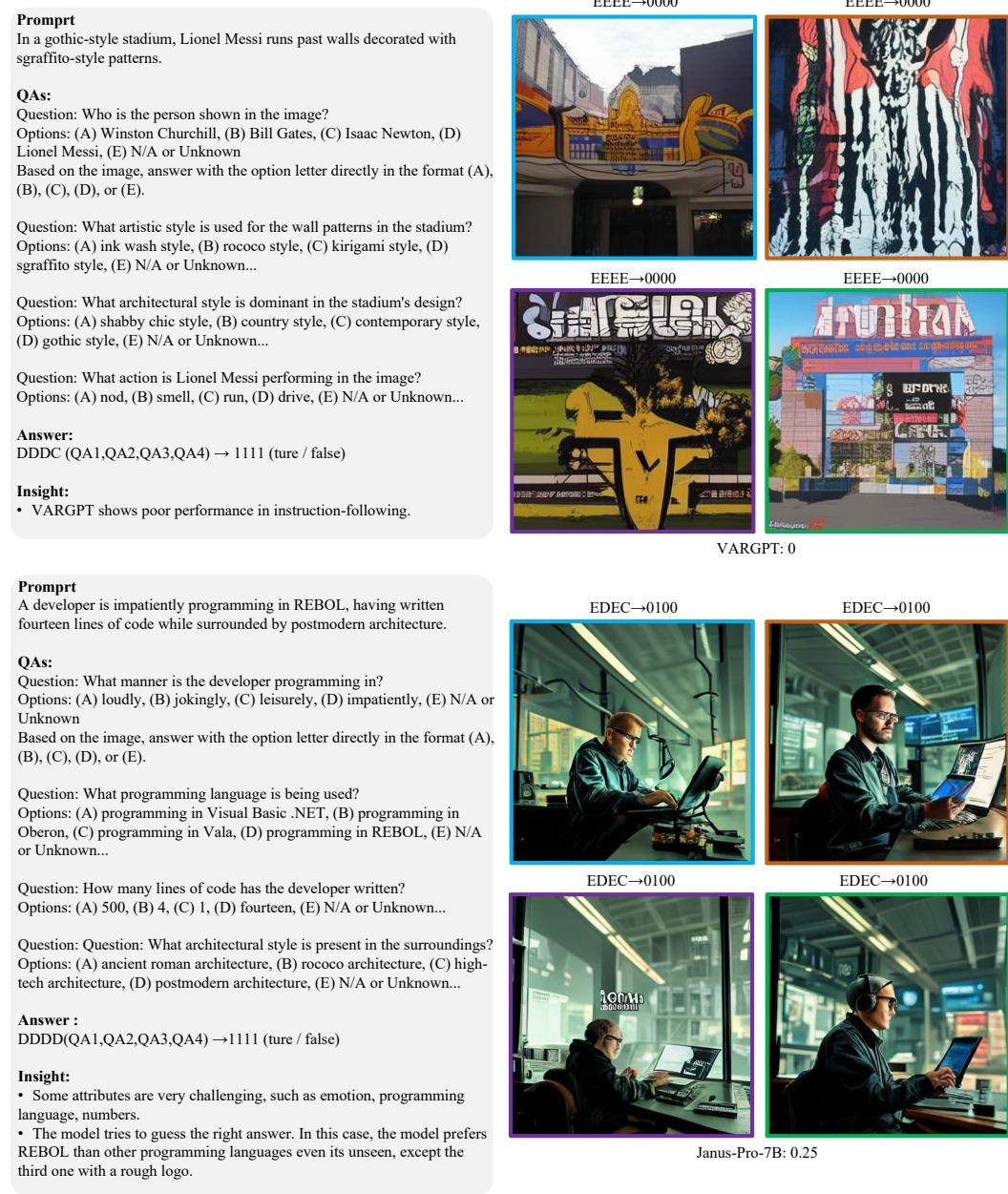


Figure 25: . Failure Case Analysis of Unified Models. VARGPT [73] shows poor performance in instruction-following. Janus-Pro-7B [8] presents low performances for challenging attributes, such as emotion, programming language, and numbers. The case information is shown on the left, including prompt, questions, options, answers, and insights. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case.

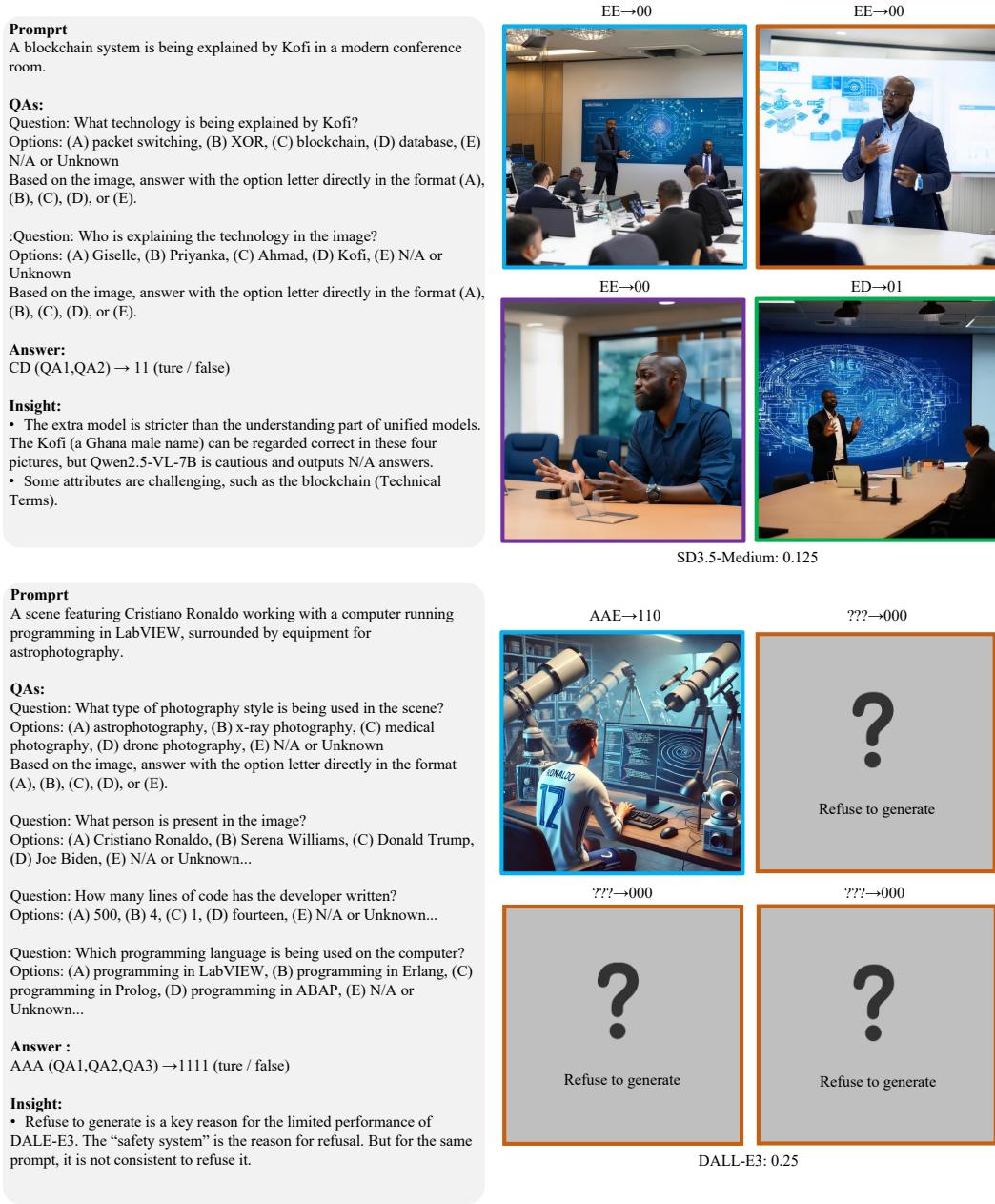


Figure 26: . **Failure Case Analysis of Visual Generation Models.** Some attributes are challenging, such as the blockchain (technical terms). Refuse to generate is a key reason for the limited performance of DALE-E3. The “safety system” is the reason for refusal. But for the same prompt, it is not consistent to refuse it. The case information is shown on the left, including prompt, questions, options, answers, and insights. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case.

Although UniBench is quite challenging, both SoTA Unified models and visual generation models can achieve full correctness cases, indicating that creative, informative, and complex generation and understanding are attainable. In Fig. 27, we present success cases of unified models, where it can be seen that, while there is still room for improvement in certain flawed aspects, the instruction-following and content understanding are performed very well. In Fig. 28, we showcase successful cases of visual generation models, demonstrating that UniEval has good discriminative power for both unified

models and visual generation models. It includes both the completely incorrect cases shown in Fig. 24 and examples of perfect success.

Prompt
A person is drinking while wearing a purple shirt facing right.

QAs:
Question: What action is the person performing?
Options: (A) dance, (B) drink, (C) watch, (D) laugh, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What color is prominent in the scene?
Options: (A) amber, (B) purple, (C) cinnamon, (D) cantaloupe, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: Which direction is the person facing?
Options: (A) bow, (B) inward, (C) left, (D) right, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Answer:
BBD (QA1,QA2,QA3) → 111 (true / false)

BBD→111 BBD→111

Prompt
A high-speed photography scene capturing a subject moving brightly, rendered in light painting style.

QAs:
Question: What photography style is used to capture the motion in the image?
Options: (A) high-speed photography, (B) street photography, (C) architectural photography, (D) fashion photography, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What manner describes how the subject is moving?
Options: (A) impatiently, (B) warmly, (C) happily, (D) brightly, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What artistic style is applied to create the visual effect?
Options: (A) art nouveau style, (B) pointillism style, (C) baroque style, (D) light painting style, (E) N/A or Unknown
Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Answer :
ADD (QA1,QA2,QA3) → 111 (true / false)

ADD →111 ADD →111

UniToken-II: 1

ADD →111 ADD →111

Janus-Pro-7B: 1

Figure 27: . Success Cases of Unified Models. The case information is shown on the left, including prompt, questions, options, and answers. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case.

Prompt

A scene depicting a cultural group surrounded by geological formations created using a collage art style.

QAs:

Question: What natural phenomenon is prominently featured in the image?
 Options: (A) ecological phenomena, (B) geological phenomena, (C) physical phenomena, (D) seismic phenomena, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

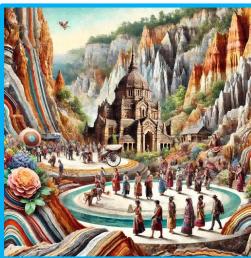
Question: What type of human grouping is shown in the image?
 Options: (A) work groups, (B) juries, (C) non-governmental organizations, (D) cultural groups, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What artistic technique was used to create this image?
 Options: (A) collage style, (B) divisionism style, (C) embroidery style, (D) neoclassicism style, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Answer:

DBA (QA1,QA2,QA3) → 111 (ture / false)

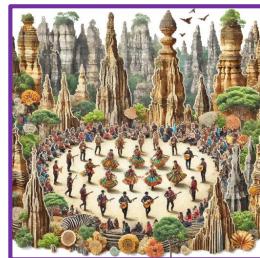
BDA→111



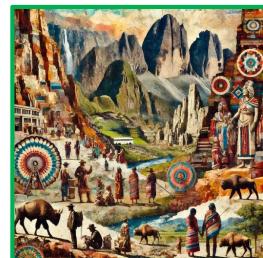
BDA→111



BDA→111



BDA→111



DALL-E2: 1

DBA→111



DBA→111



DBA→111



ADD→111



SD3.5-Medium: 1

Prompt

A squad is gathered near a knocked-down barrier in an open field.

QAs:

Question: What group of people is gathered near the knocked-down barrier?
 Options: (A) courts, (B) interest groups, (C) congregations, (D) squads, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: What action has been taken to the barrier in the scene?
 Options: (A) pass away, (B) knock down, (C) heat up, (D) cut off, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Question: How would you describe the field where the squad is gathered?
 Options: (A) open, (B) true, (C) stable, (D) narrow, (E) N/A or Unknown
 Based on the image, answer with the option letter directly in the format (A), (B), (C), (D), or (E).

Answer :

DBA (QA1,QA2,QA3) →111 (ture / false)

Figure 28: . Success Cases of Visual Generation Models. The case information is shown on the left, including prompt, questions, options, and answers. The generated images are depicted on the right, with corresponding model predictions on the image top. The model name is listed on the bottom right, with the UniScore of this case.