SpecV: Specification Verification for Robust Unified Multimodal Evaluation

Weihao Yu[1⋆†], Rongyao Fang[2⋆], Yuxuan Cai[2⋆], Linjiang Huang[3], Yuhuan Yang[2], Xianwei Zhuang[2], Junyang Lin[2], Yixuan Yuan[1‡], and Shuai Bai[2‡]

[1]The Chinese University of Hong Kong
[2]Qwen Team, Alibaba Inc. [3]Beihang University
yuweihao@link.cuhk.edu.hk baishuai.bs@alibaba-inc.com

Abstract. Unified Multimodal Models (UMMs) consolidate visual understanding, image generation, editing, and interleaved image-text interaction, yet their evaluation increasingly depends on VLM-as-a-judge scoring that shifts as judges, prompts, or model versions evolve — a problem we term evaluation drift. We present **SpecV**, a specification-verification framework for robust unified multimodal evaluation. SpecV replaces holistic scoring with the **Specification Verification Protocol**, which decomposes each prompt into atomic, binary specifications and verifies them against model outputs, improving cross-judge agreement and stability while preserving alignment with human judgments. To produce reliable specifications, we propose **Specification Ensemble and Refinement**, a multi-model pipeline that aggregates candidate specifications, deduplicates them semantically, and filters for verifiability and relevance. We also introduce **SpecV-Bench**, a 1,200-instance benchmark covering six core UMM tasks with sub-tracks of increasing constraint complexity, enabling fine-grained analysis of capability trade-offs and failure modes. Across extensive experiments with multiple judge models, SVP consistently reduces ranking flips and improves evaluation reproducibility. Benchmark is available at `https://github.com/yuyouxixi/SpecV`.

Keywords: Unified multimodal models · Evaluation drift · Specification verification

1 Introduction

Unified multimodal models (UMMs) [6,9,10,13,16,17,33,36] are converging previously isolated capabilities into single systems, spanning visual understanding, text-to-image generation, image editing, and interleaved image-text generation. Recent models such as Nano Banana [13] and Bagel [9] can already perform many of these tasks within a single architecture. However, this convergence makes evaluation harder, not easier: the majority of core UMM tasks are inherently open-ended with no single correct answer, requiring quality judgments

⋆ Equal contribution. ‡Corresponding author. †Work done during internship at Qwen Team.

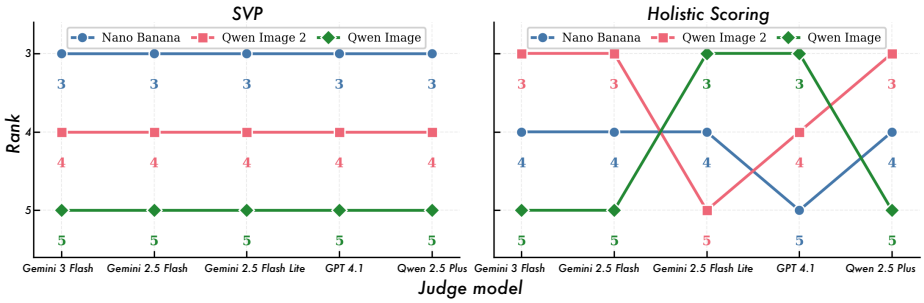

Fig. 1: Showcases of ranking changes using different evaluation VLMs under SVP and holistic scoring evaluation in the SpecV-Bench editing task.

that are far harder to stabilize than reference-based checking. Since these tasks dominate the UMM capability landscape, the fragility of open-ended evaluation becomes a central bottleneck for the credibility of unified benchmarking. This raises a natural question: *how should we evaluate these models both holistically and reliably?*

Recent unified benchmarks [11, 34, 38, 42, 43] have made substantial progress on task coverage, bringing understanding, generation, and coupled tasks into single evaluation suites. Their evaluation protocols, however, remain fragmented, each inheriting the conventions of its originating sub-community. MME-Unify [38] converts understanding tasks to multiple-choice format and uses CLIP-based [31] matching for generation; RealUnify [34] pairs direct assessment with stepwise diagnostic decomposition. These inherited protocols carry well-known limitations: discriminative formats suffer from choice-set bias, while reference-based metrics such as CLIP score [20] and FID [21] remain insensitive to fine-grained compositional semantics. These shortcomings have driven the community toward *VLM-as-a-Judge* evaluation, in which a vision-language model directly scores or ranks model outputs, valued for its semantic flexibility and ability to handle diverse output formats without hand-crafted candidate sets.

VLM-as-a-Judge addresses the flexibility problem, but it does not resolve the fragility of open-ended evaluation identified above; in fact, it introduces new instabilities. Because holistic scoring compresses a multi-dimensional output into a single scalar, the resulting score is sensitive to superficial variation: minor prompt rewording alone can shift rankings, and the judge model itself may hallucinate or misinterpret visual content. These problems compound in practice, as judge models are periodically retrained and their system prompts revised, meaning a newer version may weight evaluation criteria differently without any explicit notification. We call this phenomenon **evaluation drift**: benchmark scores shift not because the models under test have changed, but because the evaluation instrument itself has. Any robust unified evaluation framework must explicitly guard against it.

In this paper, we introduce **SpecV** (**Spec**ification **V**erification), a unified evaluation framework designed to address this challenge by reformulating open-ended assessment as a *specification verification* problem. Rather than asking

a judge model to assign a holistic score, SpecV decomposes each evaluation instance into a set of fine-grained, binary questions derived from the original prompt, each targeting a single verifiable aspect of the expected output. We refer to each such question as a *specification*, and the evaluation reduces to checking whether the model output satisfies each one. We operationalize this idea through the **Specification Verification Protocol (SVP)**, the core scoring mechanism within SpecV. Through large-scale experiments with multiple judge models, SVP yields substantially higher cross-judge agreement and stability than holistic scoring, while maintaining strong correlation with human judgments. As illustrated in Fig. 1, SVP produces identical model rankings across all five judge models, whereas holistic scoring yields highly inconsistent orderings under the same set of judges.

Constructing high-quality specifications is itself nontrivial. Specifications produced by a single model may exhibit systematic blind spots: some models emphasize color and style, others focus on spatial layout or text rendering. To mitigate such bias, we propose **Specification Ensemble and Refinement (SER)**, a scalable pipeline that leverages the complementary strengths of multiple VLMs to independently generate candidate specifications, and then consolidates them through semantic deduplication and multi-dimensional quality filtering to obtain the final specification set.

To validate SpecV on realistic unified multimodal tasks, we build **SpecV-Bench**, a comprehensive benchmark of **1,200** test instances spanning six core tasks of UMMs: multimodal understanding, text-to-image generation, image editing, many-to-one generation, interleaved image-text generation, and thinking with images. SpecV-Bench organizes each task into fine-grained sub-tracks with increasing constraint complexity, enabling targeted diagnosis of where models succeed or fail as requirements become more compositional. Our experiments reveal clear capability trade-offs across tasks and a pronounced sensitivity to constraint complexity, underscoring that progress on any single sub-community metric does not necessarily translate to robust unified performance.

Our contributions are threefold, forming a coherent evaluation framework:

- We introduce the **Specification Verification Protocol (SVP)**, which reformulates open-ended multimodal evaluation as binary specification checking. Large-scale experiments across multiple judge models show that SVP achieves substantially higher cross-judge agreement and stability than holistic scoring, while maintaining strong correlation with human judgments.
- We propose **Specification Ensemble and Refinement (SER)**, a scalable pipeline that aggregates candidate specifications from multiple VLMs and consolidates them through semantic deduplication and multi-dimensional quality filtering, mitigating the systematic blind spots of any single model.
- We build **SpecV-Bench**, a benchmark of **1,200** instances spanning six core tasks with fine-grained sub-tracks of increasing constraint complexity. Experiments on SpecV-Bench provide actionable insights into cross-task trade-offs and failure modes under compositional constraints.

2 Related Works

2.1 Benchmarks for Unified Multimodal Models

The rapid development of UMMs has spurred dedicated evaluation benchmarks [23, 34, 38, 42, 43]. MME-Unify [38] jointly assesses comprehension and generation across 10 tasks by sampling from 12 existing datasets and standardizing scoring. UniEval [23] leverages the dual capabilities of unified models to evaluate themselves, using the understanding module to judge the model's own generations. RealUnify [34] shifts the focus to bidirectional synergy, asking whether understanding genuinely enhances generation and vice versa. Despite progress in task coverage, existing benchmarks share two limitations. First, their evaluation protocols remain fragmented — multiple-choice accuracy, CLIP-based matching, holistic VLM scoring, and trained judges coexist even within a single benchmark, making cross-benchmark comparison difficult. Second, none systematically addresses the stability of the evaluation instrument itself: whether model outputs would receive consistent ranks when the judge is updated or replaced. SpecV addresses both gaps through a unified specification verification protocol that applies consistently across all tasks and is explicitly designed to resist evaluation drift.

2.2 Evaluation Methods for Multimodal Generation

Early evaluation relied on distributional metrics such as FID [21] and IS [32], which capture aggregate statistics but are insensitive to individual compositional errors. CLIPScore [20] introduces embedding similarity as a proxy for prompt faithfulness, yet single-vector representations obscure fine-grained semantics [7]. Learned reward models such as ImageReward [39] and HPSv2 [37] improve human correlation but require task-specific training and remain opaque about which aspects succeed or fail. To achieve finer-grained evaluation, methods like TIFA [22] and DSG [7] decompose prompts into verifiable sub-questions, representing an important step toward decomposed assessment but limited to text-to-image faithfulness. The LLM-as-a-Judge paradigm [41] and its multimodal extension VLM-as-a-Judge [4] offer the semantic flexibility needed for diverse outputs and are increasingly adopted by modern benchmarks. However, they compress multi-dimensional quality into holistic scores, making them sensitive to prompt phrasing, judge version, and visual hallucinations. Our specification verification protocol retains this semantic flexibility but constrains each judgment to a single binary check, substantially reducing the degrees of freedom available for drift.

3 Specification Verification

Evaluating Unified Multimodal Models (UMMs) across diverse tasks and modalities requires an evaluation framework that retains the semantic flexibility of VLM-based judgment while providing the reproducibility that holistic scoring lacks. In this section, we present the **SpecV** evaluation framework. We address

the evaluation challenge through two complementary components: the *Specification Verification Protocol*, which decomposes each evaluation into a set of atomic binary checks (§3.2), and the *Specification Ensemble and Refinement* pipeline, which constructs the specification sets underpinning SVP through multi-model consensus (§3.3). We begin by formalizing the evaluation setup and the drift phenomenon that motivates our design (§3.1).

3.1 Holistic Scoring and Evaluation Drift

Setup. A test instance $x = (t, \mathcal{I})$ consists of a textual instruction t and a (possibly empty) set of reference images $\mathcal{I}$. An under-tested model $\mathcal{M}$ produces output $y = \mathcal{M}(x)$, which may be text, image(s), or an interleaved text-image sequence depending on the task. A judge model $\mathcal{J}$ is responsible for assessing the quality of y with respect to x.

Holistic VLM scoring. Under the prevailing VLM-as-a-Judge paradigm, the judge receives a natural-language scoring prompt p_{H} that describes the desired quality criteria and returns a scalar assessment:

$$\mathrm{Score}_{\mathrm{H}}(y;\, x,\, \mathcal{J}) \;=\; \mathcal{J}(x,\, y,\, p_{\mathrm{H}}) \;\in\; [0,\, N], \tag{1}$$

where N is the upper bound of the scale. This formulation is attractive for its generality: p_{H} can express arbitrary quality dimensions and any output format. Yet it conflates two distinct responsibilities within the judge: deciding *which* aspects of y matter and deciding *how heavily* each aspect should influence the final score. Because the relative weighting is never made explicit, it becomes an implicit function of the judge's training data, system prompt, and decoding strategy, all of which may change across model updates or changes.

Evaluation drift. This conflation gives rise to what we term *evaluation drift*. Let $\mathcal{J}_a$ and $\mathcal{J}_b$ denote two judge variants—successive releases within the same model family, or models from different providers. Given a fixed collection of outputs $\{y_1, \ldots, y_M\}$ from M models on instance x, the two judges induce respective model rankings σ_a and σ_b. Evaluation drift manifests as low rank correlation between σ_a and σ_b: the relative ordering of models shifts not because the models have changed, but because the evaluation instrument has. Crucially, this instability is not a consequence of random noise in individual judgments, it arises from a structural source. When the evaluation criteria are implicit, different judges can reasonably adopt different internal weightings. One may penalize spatial errors more heavily, another may prioritize color fidelity. Even if their perceptions largely agree, this can lead to systematically different rankings. As demonstrated in §5.3, this phenomenon is pervasive: swapping one frontier VLM for another as the holistic judge can substantially alter model rankings, calling into question the reproducibility of benchmark conclusions. Therefore, a robust evaluation framework must define explicit criteria and restrict each judgment to a narrow scope, minimizing the potential for weighting disagreements.

3.2 Specification Verification Protocol

To fulfill the above requirements, we introduce the Specification Verification Protocol (SVP). SVP addresses evaluation drift by separating the two respon-

sibilities that holistic scoring conflates. *What to assess* is fixed by an explicit specification set before any judge is invoked; *how to assess* each aspect is reduced to a binary verification that admits minimal ambiguity.

Specification sets. For each test instance x we construct a specification set $\mathcal{S}_x = \{s_1, s_2, \ldots, s_K\}$, where each specification s_k is a natural-language statement describing a single, independently verifiable property of the expected output. A well-formed specification should simultaneously possess atomicity (targeting a single aspect for binary judgment), objectivity (focusing on factual attributes rather than subjective aesthetics), and determinacy (ensuring consistent conclusions across different judges).

Verification and scoring. Given a model output y and a specification s_k, the judge performs a binary check:

$$v_k \;=\; \mathcal{J}(x,\, y,\, s_k) \;\in\; \{0,\, 1\}, \tag{2}$$

where $v_k = 1$ indicates that specification s_k is satisfied and $v_k = 0$ indicates a violation. The SVP score for instance x is the specification satisfaction rate:

$$\mathrm{Score}_{\mathrm{SVP}}(y;\, x,\, \mathcal{J}) \;=\; \frac{1}{K}\sum_{k=1}^{K} v_k\,. \tag{3}$$

This yields a value in $[0, 1]$ with a transparent interpretation: the fraction of explicitly stated requirements that the output fulfills. By decomposing the underdetermined holistic mapping into K independent binary channels, SVP limits each judge decision to a narrow, well-defined scope; any single disagreement alters the aggregate by at most $1/K$, so cross-judge discrepancies are bounded and attenuated by averaging (see §5.3). While specification content is task-dependent, the verification mechanism remains identical across all tasks, enabling meaningful cross-task comparison on a common scale and sidestepping the calibration difficulties of heterogeneous protocols. The per-specification breakdown further serves as a built-in diagnostic, allowing practitioners to pinpoint failure modes such as spatial-relation violations or text-rendering errors without additional annotation.

3.3 Specification Ensemble and Refinement

The effectiveness of SVP rests on the quality of the specification sets, they should be comprehensive enough to cover all salient aspects of the expected output, yet precise enough to support reliable binary verification. However, creating such high-quality specifications is challenging. If a single VLM generates the entire set, the criteria will inevitably reflect that model's specific biases. Therefore, relying on a single source creates systematic blind spots, which silently narrow the scope of evaluation.

To mitigate this limitation, we propose the SER pipeline, which harnesses the complementary strengths of multiple frontier VLMs through three automated stages followed by a human quality-assurance step. The underlying principle is analogous to ensemble methods in machine learning: aggregating predictions

from models with diverse inductive biases yields a result that is less biased than any individual contributor.

Stage 1: Multi-model generation. To put this ensemble principle into practice, we begin by independently prompting L frontier VLMs $\mathcal{M}_1, \ldots, \mathcal{M}_L$ to produce candidate specification sets for each test instance x:

$$\mathcal{S}^{(i)} = \text{Generate}(\mathcal{M}_i,\, x), \quad i = 1, \ldots, L. \tag{4}$$

All models receive the same structured prompt requesting atomic, binary, and objectively verifiable specifications, but their different training distributions lead to naturally complementary coverage of quality dimensions. We use $L = 3$ frontier models from distinct providers (*i.e.*, Gemini 3 Pro [15], GPT 5.2 [25], Claude Opus 4.5 [1]) to maximize distributional diversity.

Stage 2: Semantic consolidation. However, the union $\bigcup_i \mathcal{S}^{(i)}$ typically contains substantial redundancy: different models often express the same underlying requirement in different phrasings. We employ a dedicated VLM (*i.e.*, Gemini 3 Pro) to perform semantic deduplication and consolidation. Specifically, the model merges paraphrased or near-duplicate specifications into a single canonical form, while flagging any logically contradictory items for downstream review. This process yields a consolidated set, $\mathcal{S}_{\text{merged}}$, which preserves the diverse coverage of the original union without artificially inflating the specification count.

Stage 3: Cross-model quality filtering. Not all consolidated specifications are equally amenable to reliable binary verification: some may be too vague, others may be tangential to the core intent of the instance. To filter for quality, each of the L VLMs independently rates every specification in $\mathcal{S}_{\text{merged}}$ on a scale of 0 to 10, assessing relevance to the test instance, clarity of formulation, and suitability for unambiguous binary judgment. We retain only those specifications whose cross-model average meets a quality threshold τ:

$$\mathcal{S}_q = \left\{\, s \in \mathcal{S}_{\text{merged}} \;\middle|\; \tfrac{1}{L} \textstyle\sum_{i=1}^{L} \text{Score}(\mathcal{M}_i,\, s) \;\geq\; \tau \,\right\}. \tag{5}$$

The use of multiple raters mirrors the multi-model generation strategy: it prevents any single model's quality preferences from dominating the filtering step. We set $\tau = 8$ in all experiments. The importance of this step is studied in §5.4.

Human quality assurance. As a final safeguard, all specifications surviving the automated pipeline undergo review by trained annotators, who verify that each item is faithful to the intent of the test instance, unambiguously verifiable from the model output alone, and non-redundant with other retained specifications. Items that fail any criterion are revised or removed. This human-in-the-loop stage imposes a quality ceiling that purely automated methods cannot guarantee, while the preceding automated stages keep the annotation burden manageable.

Together, SVP and SER form a complete evaluation pipeline: SER constructs the specification sets offline, once per benchmark release, and SVP applies them at evaluation time for each model output and judge combination.

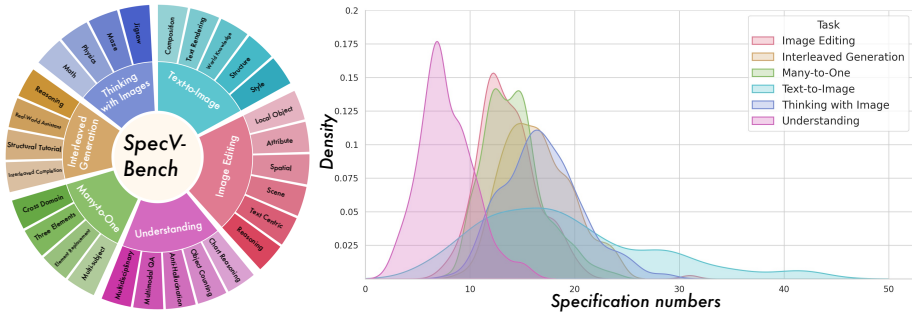

Fig. 2: Statistical analysis of SpecV-Bench. Left: 6 tasks and 28 sub-tracks of SpecV-Bench. Right: Specification numbers distribution of different tasks.

4 SpecV-Bench

Built upon the SpecV evaluation framework introduced in §3, we present SpecV-Bench, a comprehensive benchmark comprising 1,200 test instances across six core capability axes of UMMs. We first articulate the design principles that guide benchmark construction (§4.1), then provide a detailed account of each task category and its sub-tracks (§4.2), and conclude with the data construction pipeline and quality assurance measures (§4.3). An overview of the benchmark is provided in Fig. 2. After the full SER pipeline, SpecV-Bench contains **17,604** specifications across its 1,200 test instances, yielding an average of **14.67** specifications per instance. Fig. 3 displays data examples in SpecV-Bench.

4.1 Design Principles

The design of SpecV-Bench is guided by three principles that collectively shape a benchmark suited to the unique demands of UMMs.

Full-spectrum task coverage. A unified multimodal model is expected to understand, generate, and transform visual content within a single architecture. Its evaluation should therefore be equally unified. SpecV-Bench encompasses six capability axes, spanning understanding, text-to-image generation, image editing, many-to-one generation, interleaved generation, and thinking with images, within a single coherent framework. This design enables direct comparison of model performance across the full range of multimodal abilities and, more importantly, reveals cross-task trade-offs that would otherwise remain hidden when capabilities are assessed independently.

Task-specific sub-track taxonomy. Within each of the six capability axes, SpecV-Bench further introduces fine-grained sub-tracks tailored to the distinct characteristics of the task at hand. For example, text-to-image generation is decomposed along dimensions such as object counting, spatial layout, attribute binding, and stylistic control, whereas image editing is organized by edit granularity ranging from local object manipulation to global style transfer. This task-aware decomposition serves two diagnostic purposes: it pinpoints the specific scenarios in which a model excels or struggles, and it enables meaningful

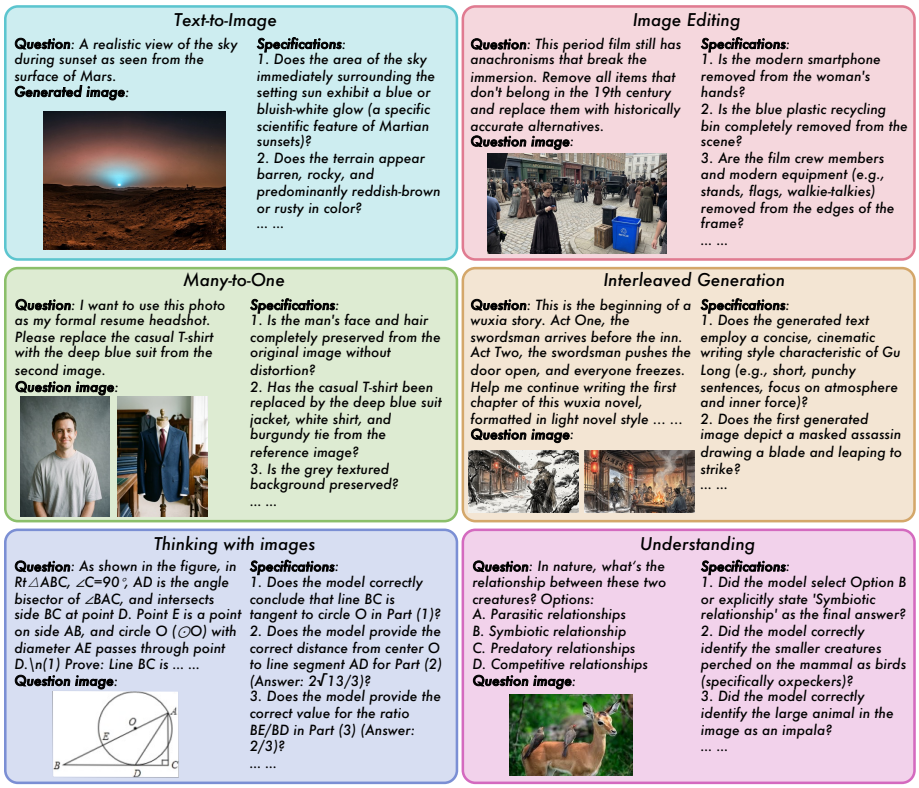

Fig. 3: Examples of the 6 tasks in SpecV-Bench. SpecV-Bench contains 1,200 test instances and 17,604 specifications.

cross-task correlation analysis, for instance, revealing whether a model's strength in spatial reasoning during understanding transfers to accurate spatial layout in generation.

Specification-driven construction. SpecV-Bench incorporates the evaluation framework from the very beginning of the design process. Each test question is crafted so that its requirements can be naturally decomposed into a set of atomic, independently verifiable specifications, deliberately avoiding items that rely on subjective aesthetic judgment or are otherwise difficult to assess in a reproducible manner. Rather than pursuing sheer volume, we invest in annotation depth, with each instance associated with an average of 14.67 fine-grained specifications, yielding 17,604 individual binary verification checks across the entire benchmark. This specification-driven philosophy ensures tight alignment between what a question demands and how the output is assessed, and delivers diagnostic granularity that exceeds what the instance count alone would suggest.

4.2 Task and Sub-track Overview

Following the principles outlined above, SpecV-Bench organizes its 1,200 instances into six task categories that together cover the full capability landscape of UMMs. We summarize the task taxonomy in Fig. 2 and briefly describe each category below. Detailed sub-track definitions and example instances are provided in supplementary materials.

Text-to-Image Generation spans five sub-tracks—compositional generation, text rendering, world knowledge, structural fidelity, and style adherence—that collectively probe the breadth of visual synthesis capabilities.

Image Editing covers six sub-tracks ranging from atomic operations (object addition, removal, attribute modification) to holistic changes (background manipulation, style transfer) and reasoning-driven editing where the intent is expressed at a high semantic level.

Many-to-One Generation requires synthesizing a single coherent image from multiple heterogeneous reference images and a text prompt, with sub-tracks varying the degree of cross-reference visual and semantic consistency.

Interleaved Generation asks the model to produce outputs that alternate coherently between text and images, with sub-tracks progressing from constrained gap-filling to open-ended, knowledge-intensive sequence generation.

Thinking with Images targets an emerging capability in which image generation serves as a cognitive tool within multi-step reasoning, spanning geometric, scientific, and open-ended visual thinking problems.

Multimodal Understanding draws instances from five established benchmarks [19, 24, 28, 35, 40] to directly quantify whether a model's generative capability comes at the expense of comprehension, or whether the two reinforce each other.

4.3 Data Construction and Quality Control

All test instances in SpecV-Bench are manually designed by five trained annotators who are permitted to use AI assistive tools for inspiration and drafting but retain full responsibility for the final content. Throughout the authoring process, each problem, in addition to conforming to the characteristics of the track itself, must also take into account its downstream specification decomposition, so as to ensure a smooth alignment with the evaluation framework described in §3

For tasks which need for reference or context images (*e.g.*, image editing, many-to-one or interleaved generation), we obtain such images from two complementary sources: (i) synthetic images produced by state-of-the-art generation and editing models, which offer fine-grained control over desired visual properties, and (ii) royalty-free images collected from the web, which provide naturalistic diversity. For thinking with images task, data for the *math* and *physics* sub-tracks are commissioned from domain experts to ensure the correctness of the reasoning problems and the necessity of image-based reasoning. The *maze* and *jigsaw* sub-tracks are generated programmatically, enabling precise control over problem parameters and guaranteeing the existence of unique solutions. As noted in §4.2, understanding instances are sampled from five established benchmarks, *e.g.*, Charxiv-RQ [35] and MMMU [40]. We apply stratified sampling to

Table 1: Comparison of different models on SpecV-Bench.

Category	Model	Understanding	T2I	Editing	Many to One	Interleave Gen.	Think w/ Images	Average
Unified Multimodal Models	Nano Banana	78.21	68.03	74.95	68.74	82.04	51.00	70.50
	Nano Banana Pro	83.25	90.23	88.42	74.31	74.63	61.81	78.78
	Nano Banana 2	83.25	90.45	90.34	76.33	80.27	69.97	81.77
	Qwen-Image-2	71.00	75.28	73.30	67.73	74.56	32.47	65.72
	Qwen-Image	57.00	64.60	69.16	60.82	34.03	19.84	50.91
	Bagel w/o CoT	63.75	28.09	54.45	47.35	35.90	16.81	41.06
	Bagel w. CoT	69.00	26.39	56.88	45.96	45.06	18.34	43.61
	Emu 3.5	45.99	67.14	64.53	66.53	62.23	14.18	53.43
Vision Language Models	Qwen 3.5 Plus	85.14	-	-	-	-	-	-
	Claude Opus 4.5	78.25	-	-	-	-	-	-
	GPT 5.2	87.29	-	-	-	-	-	-
	Gemini 3 Pro	88.25	-	-	-	-	-	-
	Gemini 3.1 Pro	89.00	-	-	-	-	-	-
Image Generation / Editing Models	GPT Image 1	-	66.79	68.64	73.23	-	-	-
	GPT Image 1.5	-	80.24	81.14	77.64	-	-	-
	Seedream 5.0	-	81.30	85.14	73.10	-	-	-
	GLM-Image	-	45.49	-	-	-	-	-
	Z-Image-turbo	-	54.14	-	-	-	-	-
	BLIP3o-NeXT	-	25.60	-	-	-	-	-

preserve the topical and difficulty distribution of the original benchmarks. Every test instance undergoes cross-checking by at least one annotator other than its author, covering question clarity, image adequacy, and specification decomposability. Instances that fail any of these criteria are revised or discarded, ensuring that the released benchmark maintains a consistently high standard of quality. The detailed per-task construction procedure is provided in the supplementary material.

5 Experiments

We organize our experiments around four objectives. First, we establish the experimental setup and models under evaluation (§5.1). Second, we present benchmark results on SpecV-Bench to illustrate the diagnostic value of our task taxonomy (§5.2). Third, we validate the SVP by comparing its stability and human correlation against holistic scoring (§5.3). Finally, we ablate the SER pipeline to quantify the contribution of each stage (§5.4).

5.1 Experimental Setup

We evaluate a total of 19 models organized into three categories. The first category consists of eight UMMs that support both visual understanding and generation within a single architecture, including Nano Banana [13], Nano Banana Pro [16], Nano Banana 2 [17], Qwen-Image [36], Qwen-Image-2 [30], Bagel [9] (w. and w/o CoT), and Emu 3.5 [8]. The second category comprises five VLMs, namely Gemini 3 Pro [15], Gemini 3.1 Pro [18], GPT 5.2 [25], Claude Opus 4.5 [1], and Qwen 3.5 Plus [29]. Since these models do not natively generate images, they are evaluated only on the understanding task and serve as an upper-bound reference for comprehension capability. The third category includes six dedicated image generation or editing models, namely GPT Image 1 [26], GPT Image 1.5 [27], Seedream 5.0 [2], GLM-Image [12], Z-Image-turbo [3], and BLIP3o-NeXT [5], each tested on the generation and editing tasks its architecture sup-

ports. Unless otherwise noted, all evaluations use Gemini 3 Flash [14] as the default judge model.

5.2 Main Results on SpecV-Bench

Tab. 1 summarizes overall performance across the six task axes, and Tab. 2 provides fine-grained sub-track breakdowns.

Table 2: Detailed comparison of different models on six tasks of SpecV-Bench.

(a) Text-to-Image

Model	Comp.	Text	Konw.	Struct.	Style	Aver.
Nano Banana	78.04	36.58	78.59	61.00	85.96	68.03
Nano Banana Pro	89.83	81.30	90.73	93.84	95.44	90.23
Nano Banana 2	89.34	83.19	91.84	91.31	96.58	90.45
GPT Image 1	70.84	43.75	79.47	59.03	80.85	66.79
GPT Image 1.5	83.61	68.75	84.99	75.33	88.52	80.24
Seedream 5.0	86.53	58.03	87.95	83.35	90.64	81.30
GLM-Image	46.33	26.39	61.05	46.13	47.03	45.49
Z-Image-turbo	70.47	43.10	52.23	39.37	65.55	54.14
Qwen-Image-2	83.99	67.07	68.98	68.39	87.97	75.28
Qwen-Image	79.04	55.18	60.74	47.74	80.29	64.60
Bagel w/o CoT	40.45	15.44	43.15	13.91	27.48	28.09
Bagel w. CoT	29.66	12.29	46.98	16.37	26.63	26.39
Emu 3.5	80.25	51.19	63.20	55.16	85.92	67.14
BLIP3o-NeXT	33.34	8.86	40.27	8.71	36.80	25.60

(b) Understanding

Model	ChartR.	Count	Anti-H.	MMQA	MDK	Aver.
Nano Banana	60.00	91.03	75.00	95.00	70.00	78.21
Nano Banana Pro	73.75	95.00	80.00	92.50	75.00	83.25
Nano Banana 2	71.25	92.50	82.50	97.50	72.50	83.25
Qwen-Image-2	58.75	86.25	72.50	90.00	47.50	71.00
Qwen-Image	31.25	72.50	65.00	87.50	28.75	57.00
Bagel w/o CoT	46.25	82.50	67.50	81.25	41.25	63.75
Bagel w. CoT	55.00	92.50	75.00	87.50	35.00	69.00
Emu 3.5	16.25	73.75	55.00	47.44	37.50	45.99
Gemini 3 Pro	78.75	95.00	85.00	97.50	85.00	88.25
Gemini 3.1 Pro	87.50	90.00	85.00	97.50	85.00	89.00
GPT 5.2	94.87	85.00	80.00	95.00	81.58	87.29
Opus 4.5	70.00	86.25	75.00	95.00	65.00	78.25
Qwen 3.5 Plus	78.21	90.00	87.50	90.00	80.00	85.14

(c) Many-to-One

Model	Cross	Repl.	3-Integ.	Multi	Aver.
Nano Banana	67.35	67.06	71.66	68.90	68.74
Nano Banana Pro	72.38	66.80	82.21	75.85	74.31
Nano Banana 2	76.67	66.42	86.49	75.73	76.33
GPT Image 1	76.60	63.34	78.28	74.71	73.23
GPT Image 1.5	81.16	68.79	81.45	79.14	77.64
Seedream 5.0	74.94	64.28	77.12	76.07	73.10
Qwen-Image-2	69.18	65.68	62.38	73.67	67.73
Qwen-Image	69.87	56.91	63.80	52.69	60.82
Bagel w/o CoT	50.64	33.15	49.95	55.67	47.35
Bagel w. CoT	50.65	35.10	49.76	48.34	45.96
Emu 3.5	71.03	56.74	71.17	67.19	66.53

(d) Image Editing

Model	Attr.	Scene	Local Reas.	Spatial	Text	Aver.
Nano Banana	93.28	83.28	85.53	70.70	89.71	70.69	74.95
Nano Banana Pro	93.62	95.19	89.86	89.60	82.10	80.12	88.42
Nano Banana 2	96.04	96.88	90.90	86.96	89.31	81.97	90.34
GPT Image 1	78.59	82.96	74.16	54.45	82.37	39.34	68.64
GPT Image 1.5	83.39	87.19	85.13	70.70	89.71	70.69	81.14
Seedream 5.0	96.92	92.37	94.33	75.05	90.54	61.65	85.14
Qwen-Image-2	94.43	93.08	90.70	43.67	73.74	44.19	73.30
Qwen-Image	88.72	84.08	91.07	38.53	76.03	36.53	69.16
Bagel w/o CoT	75.95	71.89	65.10	41.27	48.89	23.61	54.45
Bagel w. CoT	77.92	70.99	53.29	42.28	56.27	40.50	56.88
Emu 3.5	79.38	86.07	77.95	47.42	69.92	26.43	64.53

(e) Interleaved Generation

Model	Compl.	Tutor.	Real.	Reas.	Aver.
Nano Banana	87.65	77.47	87.99	75.05	82.04
Nano Banana Pro	84.73	67.39	77.98	68.42	74.63
Nano Banana 2	76.04	78.18	85.33	81.52	80.27
Qwen-Image-2	81.11	67.40	80.80	68.92	74.56
Qwen-Image	41.41	25.85	42.16	26.70	34.03
Bagel w/o CoT	33.82	27.69	51.75	30.35	35.90
Bagel w. CoT	52.84	45.90	41.34	40.16	45.06
Emu 3.5	67.22	64.15	69.45	48.11	62.23

(f) Thinking with Images

Model	Math	Phys.	Maze	Jigsaw	Aver.
Nano Banana	67.11	77.09	25.31	34.16	51.00
Nano Banana Pro	78.66	78.96	49.09	40.55	61.81
Nano Banana 2	87.32	93.28	46.74	52.53	69.97
Qwen-Image-2	42.90	46.50	13.84	26.65	32.47
Qwen-Image	27.14	20.34	11.77	20.09	19.84
Bagel w/o CoT	23.42	26.42	13.55	3.85	16.81
Bagel w. CoT	25.40	26.31	8.34	13.32	18.34
Emu 3.5	15.28	20.25	15.61	5.59	14.18

Overall performance. Among UMMs, Nano Banana 2 leads with an average score of 81.77, followed by Nano Banana Pro at 78.78, while Bagel and Emu 3.5 trail below 55 — indicating that the "unification tax" varies dramatically across architectures. Dedicated models remain competitive on individual axes: GPT Image 1.5 edges out all UMMs on many-to-one generation (77.64), Seedream 5.0 leads on editing (85.14), and Gemini 3.1 Pro tops understanding at 89.00, surpassing all UMMs. Visual results can be found in supplementary materials.

Sub-track analysis and cross-task trade-offs. Sub-track breakdowns expose persistent weaknesses that isolated evaluations would miss. Text rendering remains a universal bottleneck in T2I (Tab. 2a), with even the best model at 83.19 while weaker ones fall below 30; similarly, reasoning-driven edits in the editing category (Tab. 2d) prove far harder than explicit attribute modifications, and element replacement in many-to-one generation (Tab. 2c) consistently lags behind multi-subject composition. The unified evaluation surface also reveals instructive cross-task trade-offs: Bagel reaches 69.00 on understanding yet only 26.39 on T2I, while Emu 3.5 shows the opposite profile (67.14 generation vs. 45.99 understanding). Nano Banana 2 substantially closes the generation gap relative to the base model (90.45 vs. 68.03 on T2I) while maintaining strong interleaved performance (80.27), illustrating that architectural improvements can alleviate, though not eliminate, these tensions.

Table 3: The ranking stability of SVP and holistic scoring under intra-family and cross-family VLM evaluators.

Method		Understanding		T2I		Editing		Many to One		Interleave Gen.		Think w. Images		Average	
		τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
Intra-family	Holistic	0.839	0.931	0.906	0.978	0.685	0.849	0.749	0.898	0.842	0.936	0.810	0.929	0.805	0.920
	SVP	0.971	0.994	0.927	0.985	0.806	0.927	0.758	0.891	0.873	0.952	0.869	0.941	0.867	0.948
Cross-family	Holistic	0.941	0.988	0.902	0.969	0.758	0.888	0.807	0.914	0.847	0.940	0.880	0.952	0.856	0.942
	SVP	0.961	0.992	0.912	0.978	0.903	0.967	0.830	0.927	0.870	0.952	0.937	0.976	0.902	0.965

Table 4: The correlation between automatic evaluation metrics and human evaluation.

Method	Understanding		T2I		Editing		Many to One		Interleave Gen.		Think w. Images		Average	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
Holistic	0.358	0.388	0.432	0.501	0.419	0.453	0.261	0.305	0.583	0.656	0.699	0.788	0.459	0.515
SVP	0.391	0.421	0.557	0.628	0.471	0.519	0.496	0.548	0.601	0.692	0.743	0.824	0.543	0.605

5.3 Stability and Reliability of SVP

We now turn to validating the stability and reliability of SVP by comparing it against holistic scoring under systematic judge variation.

Ranking stability across judge models. We employ five judge models from three provider families, including GPT 4.1 from OpenAI, Gemini 2.5 Flash Lite, Gemini 2.5 Flash and Gemini 3 Flash from Google, and Qwen 3.5 Plus from Alibaba. Each judge independently scores all model outputs under both SVP and holistic scoring. For holistic scoring, we follow common practice by prompting the same judge to produce a scalar quality rating on a 0 to 10 scale. Implement details of holistic scoring can be found in the supplementary materials. Intra-family stability is computed as the average pairwise rank correlation among judges within the same provider family (*i.e.*, Gemini 2.5 Flash Lite, Gemini 2.5 Flash and Gemini 3 Flash), while cross-family stability averages all pairwise correlations across different families (*i.e.*, GPT 4.1, Qwen 3.5 Plus, and Gemini 3 Flash). We report Kendall's τ and Spearman's ρ. Tab. 3 shows the results.

SVP achieves consistently higher agreement than holistic scoring across both settings and nearly all tasks. On the cross-family setting, SVP raises the average Kendall's τ from 0.856 to 0.902 and Spearman's ρ from 0.942 to 0.965, with the largest gains on high-diversity tasks such as editing (τ: 0.758 $\rightarrow$ 0.903) where ambiguous instructions amplify divergent weighting preferences among judges. Notably, intra-family agreement under holistic scoring is actually lower than cross-family agreement, suggesting that version updates within a provider family can introduce more evaluation drift than baseline differences across providers; SVP narrows this gap by raising intra-family τ to 0.867. These results confirm that decomposing evaluation into independent, single-aspect binary checks is an effective countermeasure against **evaluation drift**.

Correlation with human judgments. To assess how well automatic metrics align with human quality perception, three trained annotators independently evaluate a randomly sampled subset of 80 instances per task, covering 4 models (*i.e.*, Nano Banana, Nano Banana 2, Qwen-Image, Qwen-Image-2). More details about human evaluation are in the supplementary materials. Tab. 4 compares

Table 5: Ablation studies on specification ensemble and refinement (SER).

Method	Understanding		T2I		Editing		Many to One		Interleave Gen.		Think w. Images		Average	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
Only Claude	0.347	0.385	0.561	0.634	0.379	0.435	0.312	0.380	0.547	0.661	0.684	0.766	0.471	0.543
Only GPT	0.368	0.402	0.480	0.560	0.386	0.441	0.290	0.344	0.516	0.620	0.637	0.733	0.446	0.517
Only Gemini	0.319	0.345	0.556	0.620	0.453	0.492	0.228	0.260	0.468	0.564	0.627	0.704	0.442	0.497
w/o filtering	0.318	0.352	0.541	0.599	0.375	0.435	0.308	0.374	0.470	0.574	0.655	0.750	0.444	0.514
SER	0.329	0.360	0.516	0.585	0.408	0.455	0.442	0.494	0.548	0.640	0.716	0.794	0.493	0.555

SVP and holistic scoring on their correlation with human ratings. SVP outperforms holistic scoring on every task, raising the average Kendall's τ from 0.422 to 0.493 and Spearman's ρ from 0.477 to 0.555. The gain is largest on many-to-one generation (τ: 0.214 $\rightarrow$ 0.442), where reconciling multiple reference images produces outputs with many independently assessable aspects that holistic scoring collapses into a single scalar, and smallest on understanding (τ: 0.319 $\rightarrow$ 0.329), where the evaluation space is inherently narrower. Together with the stability analysis above, these results confirm that SVP not only yields more consistent scores across judges but also produces ratings that are consistently *closer to human judgment*.

5.4 Ablation Studies on Specification Ensemble and Refinement

We also isolate the contribution of each SER component by comparing four alternatives against the complete pipeline in Tab. 5. The first three variants replace SER with specifications generated by a single model (Claude Opus 4.5, GPT 5.2, or Gemini 3 Pro, respectively), and the fourth variant applies multi-model generation and semantic consolidation but omits the cross-model quality filtering step. All variants are compared in terms of human correlation, using the same protocol described above.

SER: ensemble with filtering yields the best and most balanced scores. No single specification-generation model dominates all tasks: Claude leads on text-to-image ($\tau = 0.561$), Gemini on editing ($\tau = 0.453$), and GPT on understanding ($\tau = 0.368$), yet all three score below $\tau = 0.32$ on many-to-one generation. Naively merging their specifications without quality filtering actually degrades correlation compared to the best single model (average τ: 0.444 versus 0.471 for Claude alone), because the unfiltered union inevitably introduces vague or tangential items. Cross-model consensus filtering resolves this: the full SER pipeline achieves the highest average correlation ($\tau = 0.493$, $\rho = 0.555$) while also exhibiting the most uniform per-task profile, narrowing the peak-to-trough τ spread from over 0.33 for individual models to a range of 0.329–0.716. This balanced and reliable coverage across fundamentally different task types is essential for a unified benchmark.

6 Conclusion

Unified multimodal models integrate understanding, generation, and editing, yet their evaluation remains vulnerable to evaluation drift when holistic VLM judges

are updated or replaced. We introduced SpecV, which reformulates open-ended evaluation as specification verification. The Specification Verification Protocol decomposes each instance into atomic, binary requirements, improving cross-judge stability while maintaining human alignment. Specification Ensemble and Refinement further ensures specification quality through multi-VLM aggregation, deduplication, and consensus filtering. Built on this framework, SpecV-Bench provides 1,200 instances across six UMM task families with 17,604 specifications. Experiments confirm that SVP yields more stable rankings than holistic scoring under systematic judge variation, while exposing fine-grained failure modes across tasks. We hope SpecV provides a robust foundation for unified multimodal evaluation as models and judges continue to evolve.

References

1. Anthropic: Opus-4.5 (2025), https://www.anthropic.com/news/claude-opus-4-5
2. ByteDance: Seedream5.0 (2026), https://seed.bytedance.com/en/seedream5_0_lite
3. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699 (2025)
4. Chen, D., Chen, R., Zhang, S., Wang, Y., Liu, Y., Zhou, H., Zhang, Q., Wan, Y., Zhou, P., Sun, L.: Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In: Forty-first International Conference on Machine Learning (2024)
5. Chen, J., Xue, L., Xu, Z., Pan, X., Yang, S., Qin, C., Yan, A., Zhou, H., Chen, Z., Huang, L., et al.: Blip3o-next: Next frontier of native image generation. arXiv preprint arXiv:2510.15857 (2025)
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
7. Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. arXiv preprint arXiv:2310.18235 (2023)
8. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3. 5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
10. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025)
11. Fang, R., Yu, A., Duan, C., Huang, L., Bai, S., Cai, Y., Wang, K., Liu, S., Liu, X., Li, H.: Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. arXiv preprint arXiv:2509.09680 (2025)
12. GLM: Glm-image (2026), https://z.ai/blog/glm-image
13. Google: Gemini2.5-flash-image (2025), https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-image

14. Google: Gemini3-flash (2025), https://deepmind.google/models/gemini/flash/
15. Google: Gemini3-pro (2025), https://ai.google.dev/gemini-api/docs/gemini-3
16. Google: Gemini3-pro-image (2025), https://deepmind.google/models/gemini-image/pro/
17. Google: Gemini3.1-flash-image (2026), https://gemini.google/overview/image-generation/
18. Google: Gemini3.1-pro (2026), https://deepmind.google/models/gemini/pro/
19. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14375–14385 (2024)
20. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
22. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20406–20417 (2023)
23. Li, Y., Wang, H., Zhang, Q., Xiao, B., Hu, C., Wang, H., Li, X.: Unieval: Unified holistic evaluation for unified multimodal understanding and generation. arXiv preprint arXiv:2505.10483 (2025)
24. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
25. OpenAI: Gpt-5.2 (2025), https://openai.com/zh-Hans-CN/index/introducing-gpt-5-2/
26. OpenAI: Gpt-image-1 (2025), https://developers.openai.com/api/docs/models/gpt-image-1
27. OpenAI: Gpt-image-1.5 (2025), https://developers.openai.com/api/docs/models/gpt-image-1.5
28. Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3170–3180 (2023)
29. Qwen: Qwen-3.5 (2026), https://qwen.ai/blog?id=qwen3.5
30. Qwen: Qwen-image-2.0 (2026), https://qwen.ai/blog?id=qwen-image-2.0
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
33. Shi, W., Yu, A., Fang, R., Ren, H., Wang, K., Zhou, A., Tian, C., Fu, X., Hu, Y., Lu, Z., et al.: Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. arXiv preprint arXiv:2510.14958 (2025)
34. Shi, Y., Dong, Y., Ding, Y., Wang, Y., Zhu, X., Zhou, S., Liu, W., Tian, H., Wang, R., Wang, H., et al.: Realunify: Do unified models truly benefit from unification? a comprehensive benchmark. arXiv preprint arXiv:2509.24897 (2025)

35. Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., Zhu, R., Liang, K., Wu, X., Liu, H., Malladi, S., et al.: Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems **37**, 113569–113697 (2024)
36. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
37. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
38. Xie, W., Zhang, Y.F., Fu, C., Shi, Y., Nie, B., Chen, H., Zhang, Z., Wang, L., Tan, T.: Mme-unify: A comprehensive benchmark for unified multimodal understanding and generation models. arXiv preprint arXiv:2504.03641 (2025)
39. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
40. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)
41. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
42. Zhou, P., Peng, X., Song, J., Li, C., Xu, Z., Yang, Y., Guo, Z., Zhang, H., Lin, Y., He, Y., et al.: Opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 56–66 (2025)
43. Zou, K., Huang, Z., Dong, Y., Tian, S., Zheng, D., Liu, H., He, J., Liu, B., Qiao, Y., Liu, Z.: Uni-mmmu: A massive multi-discipline multimodal unified benchmark. arXiv preprint arXiv:2510.13759 (2025)