Muse Spark 1.2 Multimodal

Evaluation Methodology

We evaluate Muse Spark 1.2 on multimodal benchmarks spanning visual perception and reasoning, visual knowledge and intelligence, spatial reasoning and user-interface grounding, and chart understanding.

Overall Methodology

- **Model configurations.** Muse Spark 1.2 results are generated through the Meta Model API. We include direct-response and tool-enabled configurations and compare the tool-enabled configuration with Muse Spark 1.1, Claude Opus 5, GPT-5.6 Sol, and Gemini 3.7 Flash when comparable results are available. We use the maximum available reasoning effort for each model: xhigh for Muse Spark 1.2 and Muse Spark 1.1, max for Claude Opus 5 and GPT-5.6 Sol, and high for Gemini 3.7 Flash.
- **Tooling and preprocessing.** Tooling includes containers that support headless coding execution and GUI interaction. Within these containers, internet access is blocked to prevent potential cheating behaviors. Images are downscaled so their longest side is at most 2,000 px and re-encoded at Pillow quality 90.
- **Evaluation and grading.** We adapt each benchmark's official setup as closely as possible to a common evaluation framework, preserving the benchmark's tasks, media inputs, grading logic, and primary metric. Within each benchmark, compared models use the same evaluation data and scoring definition. A model response that does not produce a gradable final answer within the evaluation limits receives no credit. Departures from the official setup are described in the corresponding benchmark section below.
- **Third-party models.** Third-party results represent our best-effort evaluations, and runtime details may vary. Where applicable, we use a common evaluation framework and align benchmark settings as closely as practical. Our prompts and tool environments may not be specifically tuned for proprietary third-party models, so the results may not reflect their best performance in environments tailored to their strengths.
- **Multimodal intelligence.** The multimodal intelligence score is the average of the scores for BabyVision, PerceptionBench, ZeroBench, WorldVQA, SimpleVQA, ERQA, OmniSpatial, CharXiv Reasoning, ChartMuseum, and ChartQAPro. Each benchmark contributes equally to the aggregate.

Visual perception and reasoning

BabyVision: BabyVision contains 388 visual-reasoning questions spanning fine-grained discrimination, spatial perception, visual tracking, and visual pattern recognition, all of which are included in the reported score. The set comprises 135 multiple-choice and 253 free-form questions. GPT-OSS-120B with high reasoning assigns a binary grade based on semantic equivalence to the reference answer. The primary metric is mean accuracy across valid graded responses.

PerceptionBench: PerceptionBench contains 3,000 verified open-ended visual-perception questions, all of which are included in the reported score. GPT-OSS-120B with high reasoning assigns a binary correct-or-incorrect grade against the reference answer. The primary metric is mean accuracy across valid graded responses.

Visual knowledge and intelligence

ZeroBench: ZeroBench contains 100 deliberately difficult multimodal-reasoning questions. Responses are generated using stochastic sampling. GPT-OSS-120B with high reasoning and container access compares the final response with the reference answer and assigns a binary semantic-correctness grade. The primary metric is the percentage of questions for which at least one evaluated response is judged correct over five samples, consistent with the leaderboard.

WorldVQA: WorldVQA contains 3,500 image-question pairs across nine categories. We use the benchmark's official headline subset: 3,000 questions across eight categories, excluding all 500 questions in the Notable People category. GPT-OSS-120B with high reasoning grades correctness against the reference answer. The primary metric is mean question-level accuracy across the 3,000-question subset.

SimpleVQA: SimpleVQA contains bilingual visual-factuality questions in English and Chinese. We use a 2,024-question evaluation set: 1,013 English questions and 1,011 Chinese questions. GPT-OSS-120B with high reasoning grades correctness against the reference answer. The primary metric is raw accuracy across the 2,024-question evaluation set.

Spatial reasoning and UI grounding

ERQA: ERQA contains 400 interleaved image-and-text, four-choice embodied-reasoning questions. We extract the final A–D option programmatically. No model-based judge is used. The primary metric is exact-choice accuracy across the 400 questions.

OmniSpatial: OmniSpatial contains 1,533 question-answer pairs over 1,387 images and clips, spanning 50 spatial tasks. A deterministic programmatic grader extracts the last recognized answer option from the final response and compares it case-insensitively with the reference option. A response with no recognized option receives no credit. No

model-based judge is used. The primary metric is mean accuracy, computed by averaging correctness within each question and weighting all 1,533 questions equally.

Chart understanding

CharXiv Reasoning: CharXiv Reasoning contains 1,000 chart-reasoning questions from the CharXiv validation set, all of which are included in the reported score. GPT-OSS-120B with high reasoning assigns a binary grade based on semantic and mathematical equivalence between the model's response and the ground-truth answer. The primary metric is mean binary accuracy across valid graded responses. The Gemini 3.7 Flash score was obtained directly from the model card.

ChartMuseum: ChartMuseum contains 1,000 questions in its test split, all of which are included in the reported score. GPT-OSS-120B with high reasoning assigns a binary grade based on answer equivalence to the reference answer. The primary metric is mean binary accuracy across valid graded responses.

ChartQAPro: ChartQAPro contains 1,948 test prompts, all of which are included in the reported score. Grading follows the benchmark's programmatic rules. Choice, fact-checking, and year answers use exact matching. Other numerical answers receive credit within 5% relative error. Text answers use thresholded normalized similarity, and list answers may receive fractional credit. The primary metric is the benchmark's Overall mean score across the 1,948 prompts. No model-based judge is used.

Multimodal agents

Wild Artifact Bench: For each task in Wild Artifact Bench, we first run each model to generate its output artifacts, then sample pairs of artifacts and ask a judge to award a binary score for which artifact is better. The pairwise scores are then used to fit an Elo rating for each model. We use both an agentic auto-judge and a human judge, and report their results in two separate figures. Both the auto-judge results and the human results are based on approximately 2,000 pairwise comparisons. All models compared are run at maximum available reasoning effort with a 200-turn cap on every trajectory. The models run on the same type of Linux sandbox with a desktop GUI and have access to tools for bash operations, multimedia viewing, and desktop computer use.

The average cost of running one task on Wild Artifact Bench is computed as input tokens per task × the model's uncached input rate, plus output tokens per task × its output rate. Token counts are per-task averages measured from the evaluation trajectories. We treat every input token as uncached for a fair cold-start comparison. We use each provider's base on-demand rates and exclude cache discounts, batch discounts, contributor pricing, regional uplifts, and long-context uplifts.

Design Arena

Image to Frontend, **Image to Website**, and **Video to Website** are reported directly from the public Design Arena leaderboard, read on August 18, 2026. Reasoning effort is each model's Design Arena default unless otherwise noted. This slightly departs from the reasoning effort reflected elsewhere in this methodology, since we use the maximum available reasoning effort for each model. In addition, the leaderboard is live and continues to accumulate votes. The reported values are fixed to the August 18, 2026 read and will drift from what the site shows over time.