FlagEvalMM: A Flexible Framework for Comprehensive Multimodal Model Evaluation

Zheqi He, Yesheng Liu, Jing-shu Zheng, Xuejing Li, Jin-Ge Yao, Bowen Qin, Richeng Xuan, Xi Yang BAAI FlagEval Team {zqhe, yangxi}@baai.ac.cn

Abstract

We present FlagEvalMM, an open-source evaluation framework designed to comprehensively assess multimodal models across a diverse range of vision-language understanding and generation tasks, such as visual question answering, text-to-image/video generation, and image-text retrieval. We decouple model inference from evaluation through an independent evaluation service, thus enabling flexible resource allocation and seamless integration of new tasks and models. Moreover, FlagEvalMM utilizes advanced inference acceleration tools (e.g., vLLM, SGLang) and asynchronous data loading to significantly enhance evaluation efficiency. Extensive experiments show that FlagEvalMM offers accurate and efficient insights into model strengths and limitations, making it a valuable tool for advancing multimodal research. The framework is publicly accessible at https://github.com/flagevalbaai/FlagEvalMM.

1 Introduction

With the rapid advancement of large language models (LLMs) (Brown et al., 2020), multimodal models, which integrate multiple forms of input or output data such as text, images, and videos, have experienced significant development in recent years. Currently, vision-language models (VLMs) (OpenAI, 2023; Anthropic, 2024) are among the most prominent multimodal models. These models typically accept textual and visual inputs-such as images or videos-and generate textual outputs, thus primarily addressing multimodal understanding tasks. Concurrently, text-to-image (T2I) (Labs, 2024; Esser et al., 2024) and text-to-video (T2V) (Kong et al., 2024; OpenAI, 2024) generation tasks, where textual as inputs and generate visual outputs, have also garnered substantial attention, highlighting multimodal generation tasks. Recently, there has been growing interest in developing unified



Figure 1: Framework of FlagEvalMM

multimodal models capable of integrating both understanding and generation functionalities (Chen et al., 2025; Wang et al., 2024b).

These developments underscore the need for efficient and comprehensive evaluation frameworks assess multimodal models' diverse capabilities. An ideal evaluation framework should accurately, efficiently, and conveniently assess various capabilities across diverse model architectures. For evaluating VLMs, several frameworks, such as Lmms-Eval (Zhang et al., 2024c) and Vlmevalkit (Duan et al., 2024), have been proposed and widely adopted. Similarly, for evaluating T2I and T2V generation models, CompBench(Huang et al., 5555) and VBench (Huang et al., 2024) are popular choice. However, existing evaluation frameworks typically target specific multimodal tasks, lacking a comprehensive evaluation system capable of supporting a wide array of multimodal tasks uniformly.

Furthermore, current evaluation frameworks generally perform model inference and evaluation within a single runtime environment. With the increasing complexity of evaluation methods, such as use LLM as a judge (Gu et al., 2024), this architectural choice has revealed several limitations. This tight coupling may lead to conflicts between model inference and evaluation environments, and it can can also impede efficient resource usage.

In this work, we propose **FlagEvalMM**, a novel

multimodal evaluation framework that addresses existing limitations by decoupling model inference from the evaluation process. As illustrated in Figure 1, FlagEvalMM separates the inference environment (*Model Runner*) from an independent evaluation service (*Evaluation Server*). Both components communicate through a lightweight protocol, effectively resolving environment conflicts and enabling more flexible resource allocation. The modular design allows developers to easily add new tasks or models as plugins without modifying the existing framework code.

Since model inference typically dominates the evaluation time, FlagEvalMM utilizes state-of-theart inference acceleration libraries (e.g., vLLM (Kwon et al., 2023), SGLang (Contributors, 2024), LMDeploy (Contributors, 2023)) to significantly speed up computation. Additionally, it employs asynchronous data loading techniques, such as data prefetching, to further reduce waiting times.

Furthermore, FlagEvalMM provides a comprehensive suite of evaluation paradigms for multimodal understanding and generation tasks, including but not limited to: (a) vision-language understanding (e.g., VQA), (b) text-to-image (T2I) and text-to-video (T2V) generation, and (c) imagetext retrieval. Due to its modular architecture, FlagEvalMM easily supports the addition of new task extensions and evaluation metrics, enhancing its versatility and applicability.

To demonstrate its utility, we integrate FlagEvalMM with the Flageval platform¹ and Huggingface Spaces², enabling users to efficiently deploy new models and conduct comprehensive evaluations. We maintain leaderboards categorized by various multimodal tasks, ranking models according to our meticulously designed capability frameworks. We have cumulatively evaluated hundreds of multimodal models, providing a comprehensive capability analysis of mainstream multimodal models. Our experiments on diverse tasks (visionlanguage understanding, text-to-image/video generation, and image-text retrieval) highlight the framework's flexibility and extensibility.

In summary, our main contributions are:

 We introduce FlagEvalMM, an open-source multimodal evaluation framework that handles both understanding and generation tasks

²https://huggingface.co/spaces/BAAI/open_ flageval_vlm_leaderboard under a unified platform.

- By employing a decoupled architecture with an independent evaluation service, FlagEvalMM resolves environment conflicts, enhances flexibility, and improves efficiency in the evaluation process.
- We provide extensive empirical results on various tasks, illustrating FlagEvalMM's capability to deliver detailed insights into different model strengths and limitations.

2 Related Work

With the significant progress of multimodal models, numerous evaluation frameworks have emerged to assess their capabilities. Specifically, for evaluating vision-language models (VLMs), several benchmarks focus on distinct aspects of performance. For instance, MMMU (Yue et al., 2024a) evaluates college-level subject knowledge; CMMU (He et al., 2024b) assesses Chinese K-12 educational content; Blink (Fu et al., 2024) tests visual perception abilities; MathVerse (Zhang et al., 2024d) and MathVision (Wang et al., 2024a) measure mathematical reasoning; OcrBench (Liu et al., 2024) examines text recognition accuracy; and Charxiv (Wang et al., 2024c) evaluates chart comprehension skills.

To facilitate convenient and evaluation across these benchmarks, several evaluation frameworks have been proposed. For instance, Vlmevalkit (Duan et al., 2024) is a pioneering open-source multimodal evaluation toolkit. However, its lack of flexibility requires intrusive code modifications for adding new benchmarks or models, making it unsuitable for plug-and-play integrations. VHELM (Lee et al., 2024) aggregates multiple datasets to evaluate nine aspects of model performance but suffers from several limitations: first, as an extension of HELM (Liang et al., 2022), its architecture is complex, hindering the integration of new models and the expansion of datasets; second, it primarily relies on API calls and has limited support for open-source models. Lmms-Eval (Zhang et al., 2024c), an excellent and widely-used VLM evaluation framework following the Harness (Gao et al., 2024) paradigm, only supports Transformers and vLLM as inference frameworks, thus restricting its flexibility. Furthermore, it does not accommodate evaluations of multimodal generation tasks, limiting its applicability to unified multimodal models.

¹https://flageval.baai.ac.cn/



Figure 2: Components and workflow of the evaluation server

Regarding the evaluation of multimodal generation tasks, benchmarks are fewer, and the evaluation methods, especially for image or video outputs, are inherently more complex. HEIM (Lee et al., 2023) is a comprehensive framework for evaluating text-to-image generation, but similar to VHELM, it is built upon HELM and presents usability challenges. VBench (Huang et al., 2024) systematically evaluates video generative models across 16 hierarchical and disentangled dimensions, yet it is exclusively tailored to video generation tasks. In contrast to these existing frameworks, our proposed FlagEvalMM offers enhanced flexibility and ease of use, supporting a wide range of multimodal understanding and generation tasks through a unified, user-friendly interface.

3 System Design

In this section, we present the system design of our proposed framework, FlagEvalMM. As illustrated in Figure 1, the system comprises two main components: an evaluation server and a model runner, which communicate through a carefully designed protocol via HTTP. The demonstration video of FlagEvalMM is available is available online.³ We will discuss the design of each component in detail.

3.1 Evaluation Server

As illustrated in Figure 2, the evaluation server provides data to the model runner and evaluates model performance. A **Task** serves as the smallest executable unit within the evaluation server, consisting of three core components:

- **Processor**: Performs data preprocessing, converting datasets from various sources into a standardized format, stored persistently.
- ³Video available at: https://youtu.be/L7EtacjoM0k

- **Config**: Provides configuration parameters such as evaluation metrics and prompt template.
- Evaluator: Evaluates model outputs and generates performance metrics.

The workflow for each task is as follows: read configurations to acquire metadata, distribute data to models, await model outputs, and finally evaluate the generated results. The evaluation server is designed with scalability in mind and can be deployed on cloud platforms to decouple evaluation and inference. While predefined Dataset types and Evaluators are provided, users can also define and register customized Datasets and Evaluators for specific tasks.

3.2 Model Runner

The Model Runner is responsible for executing model inference, offering significant flexibility while following the defined Communication Protocol with the evaluation server (see Section 3). As illustrated in the right part of Figure 1, the Model Runner consists of two primary components: the **Model Adapter** and the **Backend**.

Model adapter plays as the bridge between the evaluation server and the model inference backend, It fetches data from the evaluation server, schedules tasks, and invokes backend processes for model inference. For convenience, commonly used model adapters are provided within our model zoo, including support for OpenAI-style REST API, and popular services such as Gemini and Anthropic (further details are provided in Appendix §A). Users may directly utilize these predefined adapters or develop custom adapters tailored to their specific requirements.



Figure 3: Communication protocol between evaluation server and model runner

The Backend is the inference engine responsible for executing the model computation, user can choose the backend according to their own needs. To optimize inference efficiency, FlagEvalMM officially supports high-performance backends like vLLM, SGLang and MLDeploy. Alternatively, users can directly leverage popular libraries, such as Transformers, Diffusers, PyTorch, or other APIs for inference. To enhance efficiency and reduce redundant computations, we implement a caching mechanism based on SQLite (Gaffney et al., 2022), a lightweight database system. When caching is enabled, the system computes a hash value for input data (including text, images, and parameters) and uses this hash as a unique key to store inference results. Subsequent identical requests retrieve the stored results directly from the cache, significantly reducing processing overhead.

3.3 Communication Protocol

The communication protocol between the evaluation server and the model runner is designed to be simple, modular, and extensible. As illustrated in Figure 3, the protocol supports the complete evaluation lifecycle, including task retrieval, metadata provisioning, data access, and result submission. All interactions between the evaluation server and model runner adhere to a RESTful HTTP pattern (Fielding, 2000), with each evaluation step corresponding to a dedicated API.

The protocol starts with the model runner requesting the available tasks via get_tasks, and then querying detailed task information with task_info. After selecting a task, the runner retrieves task-level metadata meta_info using get_meta. These metadata include the number of samples, task type (e.g., VQA, T2I), output directory, and other necessary settings.

Once the task setup is complete, the model runner requests specific evaluation items using the get_data(i).The returned data_info includes necessary details like image paths, textual prompts, and unique question identifiers. After inference, the runner submits its predictions back to the evaluation server via the submit(result).

Each step in the communication protocol supports distributed and parallelized model evaluation. The protocol's modular design also enables easy integration of new task types or data formats without requiring modifications to the core communication logic. As a result, FlagEvalMM remains flexible and easily adaptable to various multimodal evaluation scenarios.

4 Evaluation Results and Analysis

We have evaluate more than 50 multimodal understanding models and 30 multimodal generation models on the FlagevalMM leaderboard. In this paper, we focus on the performance of VLMs and textto-image models. we select some frontier models from various companies and research institutions for detailed analysis.

4.1 Datasets and Evaluation Metrics

4.1.1 Multimodal Understanding

To comprehensively evaluate the multimodal understanding capabilities of models and address the dataset contamination and metric saturation issues (Chen et al., 2024), we selected multiple recent public and self-constructed evaluation datasets for this

	Average Rank			Capability Score				
Model	Overall	EN	ZH	Gen	Math	Chart	Vis	Text
gemini-2.0-pro	2.1	2.4	1.5	64.00	52.18	67.06	62.73	78.22
Qwen2.5-VL-72B	4.6	5.4	2.5	61.30	35.45	67.00	60.90	77.63
Qwen2.5-VL-32B	6.7	7.8	4.0	60.17	42.57	62.15	59.22	74.68
claude-3-7-sonnet-20250219	6.9	4.2	13.5	58.98	49.31	71.19	66.55	67.69
InternVL2_5-78B	6.9	7.2	6.0	61.31	37.80	60.14	62.97	70.87
gpt-4o-2024-11-20	8.1	7.2	10.5	58.39	30.82	65.50	62.02	70.31
claude-3-5-sonnet-20241022	8.1	6.2	13.0	59.14	45.24	71.89	62.66	67.00
Qwen2-VL-72B	10.4	12.2	6.0	57.30	32.53	60.06	54.48	71.75
gemini-1.5-pro	11.0	8.0	18.5	53.29	50.80	62.41	56.74	63.62
Mistral-Small-3.1-24B	12.6	9.6	20.0	53.36	32.40	64.94	60.49	62.25
llava-onevision-qwen2-72b	20.3	18.0	26.0	45.84	32.90	52.09	48.55	49.48
Molmo-72B-0924	22.0	18.8	30.0	43.27	26.31	54.27	50.12	44.98

Table 1: Ability evaluation of some frontier VLM models. For Gen (General Knowledge), Math, Chart, Vis (Visual Perception), Text (Text Recognition and Understanding), scores are averaged across English and Chinese evaluations.

VLM assessment: Charxiv (Wang et al., 2024c), CII-Bench (Zhang et al., 2024a), CMMMU (Zhang et al., 2024b), MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVision (Wang et al., 2024a), MathVerse (Zhang et al., 2024d), MMVETv2 (Yu et al., 2024), Blink (Fu et al., 2024), and selfconstructed subjective image-text QA dataset and text recognition and understanding dataset. These datasets cover five capabilities: general knowledge, mathematical, chart comprehension, visual perception, and text recognition and understanding, dach dataset can be mapped to one or more capabilities Additionally, we distinguished between Chinese and English capabilities based on question language and cultural type.

Except for the two self-constructed benchmarks, all datasets are publicly available academic datasets. Public datasets utilized the default prompts and accuracy calculation methods provided by their original sources. The self-constructed subjective evaluation dataset employs binary manual scoring to judge correctness. The self-constructed text recognition and understanding evaluation adopts the automatic accuracy evaluation method from OCR-Bench (Liu et al., 2024), determining correctness based on whether the manually annotated standard answer string is a subsequence of the model's response.

4.1.2 Multimodal Generation

For multimodal generation tasks, we evaluate the result for 4 aspects: consistency with the prompt, realism, aesthetic quality, and safety. In FlagevalMM, we currently support several metrics for automatic evaluation of multimodal generation models. In our leaderboard, we combined some automatic evaluation metrics with human evaluation to provide a more comprehensive evaluation, we choose VQAScore (Lin et al., 2024), Q-Align (Wu et al., 2024), VideoScore (He et al., 2024a) as the automatic evaluation metrics. In human evaluation, we employs 3 human evaluators to score the image in 4 aspects above, and the final score is the average of the 3 human scores. The detailed annotation guideline can be found in Appendix §D.

Beyond standard datasets like COCO (Lin et al., 2014) and GenAI Bench (Li et al., 2024) available in FlagEvalMM, our leaderboard uses self-constructed datasets for text-to-image and text-to-video tasks. The text-to-image dataset contains 414 self-designed high-quality prompts, while the text-to-video dataset includes 148 prompts (100 self-designed, 48 public). The self-constructed datasets are evaluated using the same automatic evaluation metrics as the public datasets.

4.2 Leaderboard

In this section, we present evaluation results for representative state-of-the-art multimodal models.

4.2.1 Results of VLMs

Table 1 summarizes the performance of representative VLMs across five key multimodal capabilities. The left side of the table shows the overall average

Model	Weighted		Human I	Evaluatio	Automated Evaluation			
WIOUCI		Cons	Real	Aes	Safety	VQAS	OA-Qua	OA-Aes
Hunyuan-Image	73.00	67.93	66.67	78.50	100.00	73.76	95.36	81.00
Doubao-Image v2.1	71.74	69.79	61.90	75.00	94.64	76.69	89.96	73.24
DALL-E 3	70.12	70.24	57.51	68.38	98.21	81.82	94.42	89.92
Kolors	68.80	68.53	62.43	63.84	92.86	75.60	88.60	80.77
FLUX.1 schnell	68.39	61.95	64.34	73.18	99.11	77.95	93.53	74.60
Firefly Image 3	66.15	62.80	57.07	68.90	95.54	74.39	88.92	76.91
Midjourney v6.1	65.91	67.56	46.95	64.58	98.21	77.63	86.82	77.60
Stable Diffusion 3.5 Large	65.22	67.86	45.61	60.86	100.00	78.28	89.47	73.47
CogView-3 Plus	64.34	67.63	45.68	57.37	99.11	80.16	90.72	80.15

Table 2: Performance comparison of text-to-image models across human and automated evaluation metrics.

rank along with language-specific average ranks, while the right side details capability scores, each representing averages from evaluations conducted in both English and Chinese. Models are ranked based on their overall average rank.

Our analysis reveals substantial progress among recent open-source VLMs. Notably, the Qwen2.5 series (Team., 2025) surpasses several earlier commercial models, highlighting significant advancements within the open-source community. This improvement indicates a narrowing performance gap between open-source and proprietary solutions in multimodal understanding tasks. However, some models, such as Mistral-3.1(AI, 2025) and Claude 3.7 (Anthropic, 2025), exhibit pronounced performance discrepancies across different languages and cultural contexts, performing notably better in English than in Chinese. These results underscore persistent challenges regarding cross-lingual and cross-cultural generalization in current VLM architectures. According to some case study, we found VLMs currently exhibit instability and inaccuracies in tasks involving spatial reasoning, position estimation, and counting. Additionally, they struggle with classic computer vision challenges such as occlusion, varying illumination, deformation, and perspective changes.

4.2.2 Results of text-to-image models

Table 2 compares the performance of selected textto-image models using both human and automated evaluation metrics. Since some T2I models only support English prompts, the results presented in the table are based on a subset of English prompts. Models are ranked according to the weighted average of human evaluation scores.

The results demonstrate that commercial mod-

els, such as Hunyuan-Image (Tencent, 2024) and Doubao-Image (ByteDance, 2024), generally achieve higher performance in human evaluation compared to open-source counterparts like FLUX (Labs, 2024) and CogView-3 (Zheng et al., 2024). Notably, while automated metrics offer useful insights, they do not always align closely with human judgments. For instance, in the consistency dimension, the VQAScore exhibits a Pearson correlation coefficient (Cohen et al., 2009) of only 0.76 with human evaluation scores. Similarly, for aesthetic quality, the OneAlign-Aesthetic metric yields a moderate correlation of 0.59. These observations highlight the limitations of current automated evaluation methods and suggest the necessity for further refinement to better reflect human perception. According to some case study, we found that T2I models often struggle with generating high-quality images for human motion scenarios and accurately depicting specified object.

5 Conclusion and Future Work

In this work, we introduce **FlagEvalMM**, an opensource integrates both multimodal understanding and generation tasks within a unified platform. By decoupling model inference from the evaluation process, FlagEvalMM effectively mitigates environmental conflicts and significantly enhances flexibility. Moreover, integration with public platforms such as FlagEval and Huggingface Spaces further enhances ease of use and accessibility. In the future, we plan to incorporate additional evaluation methodologies, such as multi-round conversational tasks, interactive gameplay with vision-language models, and advanced reasoning capability assessments. These extensions aim to broaden the scope and depth of FlagEvalMM.

Limitations

Due to the rapidly evolution of evaluation methods and models, our work integrates only a selected subset of existing evaluation approaches and benchmarks. Additionally, a significant gap remains between automated evaluation and human assessment in generation tasks, necessitating continued reliance on manual evaluation.

Acknowledgments

This work was supported by the National Science and Technology Major Project of China (Grant No. 2022ZD0116306). We thank Xue Sun for his valuable contribution in adding numerous new datasets to our open-source project.

References

Mistral AI. 2025. Mistral small 3.1.

- Anthropic. 2024. Claude 3.5 sonnet. https://www. anthropic.com/news/claude-3-5-sonnet. Accessed: 2025-01-18.
- Anthropic. 2025. Claude 3.7 sonnet. https://www. anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-03-08.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.
- ByteDance. 2024. Doubao image. https://www.volcengine.com/docs/6791/1366783.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.
- Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.
- LMDeploy Contributors. 2023. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https: //github.com/InternLM/lmdeploy.

- SGLang Contributors. 2024. Sglang: A fast serving framework for large language models and vision language models. Accessed: 2025-03-23.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the* 32nd ACM international conference on multimedia, pages 11198–11201.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Roy Thomas Fielding. 2000. Architectural styles and the design of network-based software architectures. University of California, Irvine.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.
- Kevin P Gaffney, Martin Prammer, Larry Brasfield, D Richard Hipp, Dan Kennedy, and Jignesh M Patel. 2022. Sqlite: past, present, and future. *Proceedings* of the VLDB Endowment, 15(12).
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, and 1 others. 2024a. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123.
- Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024b. Cmmu: a benchmark for chinese multi-modal multi-type question understanding and reasoning. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, pages 830–838.

- Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 5555. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, pages 1–17.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.*
- Black Forest Labs. 2024. Flux. https://github.com/ black-forest-labs/flux.
- Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and 1 others. 2024. Vhelm: A holistic evaluation of vision language models. *Advances in Neural Information Processing Systems*, 37:140632–140666.
- Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, and 1 others. 2023. Holistic evaluation of text-to-image models. Advances in Neural Information Processing Systems, 36:69981–70011.
- Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and 1 others. 2024. Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

- Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.
- Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.
- OpenAI. 2023. Gpt-4v(ision) system card. OpenAI Research.

OpenAI. 2024. Sora.

- Qwen Team. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.
- Tencent. 2024. Hunyuan image. https://cloud. tencent.com/document/product/1729/105969.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*
- Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, and 1 others. 2024b. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.
- Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems, 37:113569–113697.
- Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, and 1 others. 2024.
 Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54015–54029.
- Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. 2024. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. arXiv preprint arXiv:2408.00765.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2409.02813.
- Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2024a. Can mllms understand the deep implication behind chinese images? *Preprint*, arXiv:2410.13854.
- Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, and 1 others. 2024b. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2024c. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024d. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.
- Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, pages 1–22. Springer.

A Commercial API Support

We support mainstream APIs for multimodal tasks. For Vision-Language Models (VLMs), we provide integration with OpenAI, Gemini, Claude, Hunyan, and Qwen. For Text-to-Image (T2I) models, we support DALL-E, Flux, and Kolors. Additionally, we offer OpenAI-style REST API compatibility for both types of tasks, which we highly recommend using for seamless integration and ease of deployment.

B Add A New Evaluation Task

This section describes the procedure for adding new evaluation tasks to the benchmark system. The process consists of three main steps:

B.1 Create Task Configuration

New evaluation tasks require creating appropriate configuration files in the tasks directory. For simple tasks (e.g., Visual Question Answering), developers can utilize the existing VqaBaseDataset class.

The basic configuration template includes:

- dataset_path: Path to the original dataset
- split: Dataset partition (e.g., "image")
- processed_dataset_path: Storage path for processed datasets (e.g., "CustomBench")
- processor: Data processing script (e.g., "process.py")

Developers can configure tasks in two ways:

- 1. **Default Prompt Configuration**: Uses the system's default prompt template ("Answer the question using a single word or phrase.")
- 2. Custom Prompt Configuration: Allows customization of the prompt template for specific task requirements

B.2 Implement Data Processing

Each new task requires a dedicated processing script (specified in the processor field) to transform raw data into the system's standardized format. The script should handle:

- Data loading from source files
- Format conversion
- · Quality control checks
- · Output generation in the expected structure

B.3 Register the Task

After configuration and processing implementation, the task must be registered in the system's task registry. This involves:

- Adding the task to the appropriate configuration files
- Updating any necessary dependencies
- Verifying integration through test cases

The modular design allows for seamless addition of new evaluation tasks while maintaining consistency across the benchmark system.

C Benchmarks for VLM evaluation

Table 3 summarizes the benchmarks utilized by the FlagEval leaderboard for evaluating visionlanguage models (VLMs). Each benchmark assesses one or more specific model capabilities, such as visual perception, general knowledge, or mathematical reasoning.

D Human Evaluation Process and Scoring Guidelines

In this evaluation, images generated by different models from the same textual prompt were simultaneously displayed to annotators in random order and position. Three trained annotators independently rated each image according to specific evaluation dimensions. Annotators sequentially completed scoring for each evaluation dimension before proceeding to the next. After completing scoring for all three dimensions, annotators repeated this process for two additional rounds. The repeated evaluation rounds were designed to measure and ensure the stability and consistency of annotator scoring criteria.

The evaluation dimensions included Text-Image Consistency, image realism, aesthetic quality, and image safety. Text-Image Consistency, realism, and aesthetic quality were scored on a 5-point scale, whereas safety was scored as a binary (0 or 1). Definitions for each evaluation dimension are provided below:

- **Text-Image Consistency**: Assesses the extent to which the generated image accurately reflects the content described by the text.
- **Realism**: Higher realism scores indicate that the image faithfully represents objects' expected shapes and characteristics rather than

Benchmark	Language	Capability
Charxiv(Val)(Wang et al., 2024c)	English	Chart Comprehension
CII-Bench(Test)(Zhang et al., 2024a)	Chinese	General Knowledge
CMMMU(Val)(Zhang et al., 2024b)	Chinese	General Knowledge
MMMU(Val)(Yue et al., 2024a)	English	General Knowledge
MMMU-Pro(Standard, Vision)(Yue et al., 2024b)	English	General Knowledge, Visual Perception
MathVision(Test)(Wang et al., 2024a)	English	Mathematical Ability
MathVerse(testmini)(Zhang et al., 2024d)	English	Mathematical Ability
MMVET-v2(Yu et al., 2024)	English	General Knowledge, Visual Perception
Blink(Val)(Fu et al., 2024)	English	Visual Perception
Self-constructed VQA Dataset	English, Chinese	General Knowledge, Visual Perception
Self-constructed Text Dataset	English,Chinese	Text Recognition and Understanding

Table 4: Scoring Guidelines for Human Evaluation of Text-to-Image Models

Score	Text-Image Consistency	Realism	Aesthetic Quality	Safety
0				Image contains unsafe content
1	Image does not reflect tex- tual description	Significant errors in shapes and characteristics	Subjects difficult to iden- tify; chaotic composition	Image is safe
2	Image includes a few ele- ments from the textual de- scription	Some noticeable errors in shapes and characteristics	Poor aesthetic quality	_
3	Image contains most ele- ments of the textual de- scription	Some obvious errors but overall coherent	Average aesthetic quality; ordinary composition and color matching	_
4	Image reflects textual de- scription well	Minor, less obvious errors in shapes and characteris- tics	Good aesthetics with slight shortcomings in composition or color matching	_
5	Image perfectly aligns with textual description	No errors; image is coher- ent and realistic	Excellent aesthetic quality with outstanding composi- tion and color matching	_

arbitrary generation. For realistic scenes, the evaluation considers conformity with the real world; for surreal scenes, it considers consistency with common knowledge from animations, sci-fi films, or other surreal contexts.

- Aesthetic Quality: Evaluates the aesthetic appeal of the image, including clarity, composition, and alignment with aesthetic standards.
- **Safety**: Ensures the generated images are free from violence, pornography, drug-related content, and anti-social themes.

Detailed scoring criteria for each evaluation dimension are summarized in Table 4.