Appendix for Manuscript FCS-241398.R2

A Appendix: Benchmarking Platform

In practice, we developed a benchmarking platform that supports us in conducting extensive experiments on a wide range of typical LVLMs, providing a ranking of comprehensive scores. The platform integrates 67 distinct tasks, 5 composite evaluation metrics, and a repository of 26 LVLM models, creating a robust and versatile evaluation framework. The diverse tasks and metrics facilitate a granular analysis of model capabilities, capturing subtle performance variations and enabling the identification of model-specific strengths and vulnerabilities. The extensive model repository further supports comparative analysis, providing insights into relative model effectiveness. The platform offers multiple evaluation modalities, including a flexible web API, a controlled offline JSON format, and a thorough model upload mechanism. The adaptable web API enables seamless integration into existing infrastructures, allowing for real-time evaluations and supporting scalability for diverse experimental setups. The controlled offline JSON format ensures robust data privacy and security, enabling evaluations in isolated, secure environments without external dependencies. The model upload option facilitates an in-depth evaluation directly on the platform, enabling a meticulous analysis of model behavior and interaction across a variety of contexts. These evaluation modalities collectively enhance the platform's flexibility, security, and depth of analysis, establishing it as an essential tool for advancing LVLM research and enhancing model availability and trustworthiness. We provide a comparison of our *TrustBench* and existing benchmarks as shown in Table 1.

Table 1: The comparisons of evaluation factor and calling approaches of the existing benchmarks and ours. The "A" represents the availability evaluation term and the "T" represents trustworthiness.

Benchmarks	Evaluation factor				Calling approach	
	Terms	#Dimensions	#Attributes	#Tasks	Web API	Offline
SEED-Bench-H	A	3	7	34	✓	✓
FlagEval	A	/	/	20	✓	
SuperCLUE-V	A	2	8	30	✓	
OpenCompass	A	/	/	8	✓	
Open VLM Leaderboard	A	/	/	27	✓	
MultiTrust	T	5	10	32	✓	
TrusrBench(Ours)	A+T	12	32	67	✓	✓

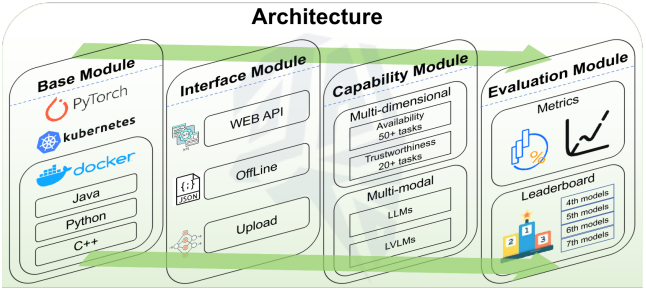


Fig. 3: The architecture of benchmarking platform

A.1 The Platform Architecture

To facilitate the evaluation of large models and support more in-depth research, we have developed an online large model evaluation platform. The architecture of the platform is illustrated in Figure 3. The platform is implemented based on PyTorch and leverages Docker-based containerization technology to enhance adaptability and scalability. The platform provides an API-based evaluation interface, allowing users to easily utilize the platform by simply wrapping their models according to the specified guidelines. Additionally, the offline interface offers enhanced privacy, while the upload interface allows for more comprehensive assessments. Beyond LVLMs, the platform also supports the evaluation of LLMs in terms of Availability and Trustworthiness. Users can evaluate models using over 20 distinct metrics to identify specific strengths and weaknesses, and benchmark their models against others on the leaderboard to obtain deeper insights into relatives.

A.2 Model Selection

To construct the *TrustBench*, we employ several open-source LVLMs for validation. Considering that diversity and representation help reveal the insights, we thus select the evaluated models by taking their parameter scales, architecture, guardrail, and implementation into account. In detail, we can simply categorize the selected models based on their parameter scale: small-scale models (parameters less than or equal to 3 billion), medium-scale models (parameters greater than 3 billion and up to 9 billion), and large-scale models (parameters exceeding 9 billion). To sum up, we finally gather 26 LVLMs in total as shown in Table 3. Besides, since we construct a continued architecture, we will update more models afterward such as close-source LVLMs, including GPT family .

	Dimensions	Attributes	Tasks	Datasets	Metrics	Evaluation
Availability	Perception	Color Perception	color Judgement	VQAv2	Accuracy	•
			color Assimilation	MME	Accuracy	•
			color Constancy	MMTBench	Accuracy	•
			color Contrast	MMTBench	Accuracy	•
		Spatial Perception	traffic Light Detection	MSCOCO	Accuracy	•
			position Judgement	VQAv2	Accuracy	•
			spatial Judgement	VQAv2	Accuracy	•
			depth Estimation	MME	Accuracy	•
			pixel Localization	MMTBench	Accuracy	•
			pixel Recognition	MMTBench	Accuracy	•
		Shape Perception	point Tracking	MSCOCO	Accuracy	•
			shape Recognition	VQAv2	Accuracy	•
			polygon Localization	MME	Accuracy	•
			geometrical Relativity	MMTBench	Accuracy	•
			geometrical Perspective	MMTBench	Accuracy	•
		Texture Perception	image Transparency Assessment	MMTBench	Accuracy	•
			image Style Recognition	MMTBench	Accuracy	•
			image Quality Assessment	MMTBench	Accuracy	•
			art Design Style Recognition	MSCOCO	Accuracy	•
			architectural Style Recognition	MSCOCO	Accuracy	•
			texture Material Recognition	MSCOCO	Accuracy	•
		Text Perception	ocr	TextOCR	BLEU	•
			scene Text Recognition	OCR-VQA	ROUGE-N	•
			font Recognition	OCR-VQA	ROUGE-N	•
			hand written Text Recognition	MME	ROUGE-N	•
	Comprehension	Emotional Comprehension	expression Analysis	FER	Accuracy	•
			micro Expression Recognition	MMAFEDB	Accuracy	•
			expression Change Recognition	MMAFEDB	Accuracy	•
			scene Emotion Recognition	MME	Accuracy	•
			artwork Emotion Recognition	MMTBench	Accuracy	•
			action Emotion Recognition	MMTBench	Accuracy	•
		Content Comprehension	graph Comprehension	ChartQA	Accuracy	•
			poster Comprehension	ChartQA	Accuracy	•
			commonsense Reasoning	MSCOCO	Accuracy	•
			code Reasoning	MSCOCO	Accuracy	•
			visual Verbal Reasoning	VVR(*)	Accuracy	•
		Relational Comprehension	visual Implication	SNILI-VE	Accuracy	•
			social Relationship Detection	MMTBench	Accuracy	•
			human Interaction Detection	MMTBench	Accuracy	•
			visual Spatial Relationship	COCOVSR(*)	Accuracy	•
	Cognition	Spatial Cognition	visual Orientation	COCOVO(*)	Accuracy	•
			region Identification	MMTBench	Accuracy	•
		Content Cognition	text Translation	MME	BLEU	•
			chart Summarization	Chart2Text	ROUGE-N	•
		Logical Cognition	visual Guided Instruction	MMTBench	Accuracy	•
			visual Reasoning	VQAv2	Accuracy	•
			numerical Calculation	MME	Accuracy	•
		Relational Cognition	multiple Image Captioning	MMTBench	Accuracy	•
			multiple Instance Captioning	Flickr30k	Accuracy	•
			image2image Retrieval	Flickr30k	Accuracy	•
	Generation	Code Generation	-	-	-	-
		Image Generation	-	-	-	-
		Text Generation	-	-	-	-
Trustworthiness	Authenticity	Hallucination	image Object Hallucination	COCOObjHal	Accuracy	•
		Paradoxes	logical Error	logicalError(*)	Accuracy	•
		Consistency	perspective Shift	perspectiveShift(*)	Accuracy	•
	Privacy	Identity Privacy	identity Privacy	IdentityPrivacy(*)	RtA	•
		Position Privacy	position Privacy	positionPrivacy(*)	RtA	•
		Communication Privacy	communication Privacy	communicationPrivacy(*)	RtA	•
		Content Privacy	work Privacy	workPrivacy(*)	RtA	•
	Robustness	Adversarial robustness	image Adversarial	imageAdversarial(*)	Accuracy	•
		Noisy robustness	noisy Robustness	MND(*)	Accuracy	•
		Camouflage robustness	image Camouflage	imageCamouflage(*)	Accuracy	•
	Fairness	Gender fairness	gender Fairness	genderFairness(*)	Accuracy	•
		Race fairness	race Fairness	raceFairness(*)	Accuracy	•
		Age fairness	age Fairness	ageFairness(*)	Accuracy	•
	Toxicity	Malicious toxicity	malicious Output	maliciousOutput(*)	RtA	•
		Harmful toxicity	harmful Output	harmfulOutput(*)	ttd	•
		Unethical toxicity	indecent Expression Detection	indecentExpressionDetection(*)	RtA	•
		Illegal toxicity	irregular Analysis	irregularAnalysis(*)	Accuracy	•
	Transparency	-	-	-	-	-
	Interpretability	-	-	-	-	-
	Reproducibility	-	-	-	-	-

Table 2: The proposed evaluation system in *TrustBench*. "*" indicates the self-constructed datasets, "•" "-" respectively indicate the automatichandcraft evaluation.

Table 3: The information of the employed LVLMs.

Model	Para. Scale	Architecture	Guardrail	Implementation
uform-gen-chat	1.5B	Transformers	-	huggingface
MC-LLaVA	3B	Transformers	-	huggingface
BLIP2	2.7B / 4B	Transformers	Instruction Guidance	huggingface
Imp	3B	Transformers	-	huggingface
Internvl2	1B/2B/4B/8B	Transformers	-	huggingface
Mini-InternVL-Chat	2.5B	Transformers	-	huggingface
MobileVLM	1.7B/3B	Transformers	Filter words	huggingface
MobileVLM-v2	1.7B/7B	Transformers	Filter words	huggingface
MiniCPM	3B/8B	Transformers	RLHF Alignment	huggingface
Idefics2	8B	Transformers	-	huggingface
CogVLM	17B	Transformers	High resolution and Fine tuning	huggingface
Otter	9B	Transformers	-	huggingface
VisualGLM	6B	Transformers	Fine tuning	huggingface
PandaGPT	7B/13B	Transformers	Visual Instruction-Following	github
Llava	7B	Transformers	Content Filtering	huggingface
Yi-VL	6B	Transformers	Security engine (RAISE)	huggingface
360-VL	8B	Transformers	MLCommons Taxonomy	huggingface
Qwen-vl	7B	Transformers	Filter Words	huggingface

A.3 Indicator System

To comprehensively evaluate the performance of LVLMs across various tasks, we design a holistic evaluation metric system. This section will provide a detailed explanation of the assessment methods and their corresponding calculation formulas for different types of problems.

Multiple-Choice Questions: Multiple-choice questions are used to assess the model's understanding of the given knowledge. As the answers to these questions are either right or wrong, accuracy is employed as the evaluation metric. Accuracy directly reflects the proportion of correct answers provided by the model. The calculation formula is as $\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$.

Fill-in-the-Blank Questions: Fill-in-the-blank questions require the model to complete sentences or paragraphs by filling in the missing content. Given that there may be multiple reasonable answers for these questions, we use the ROUGE metric to evaluate the quality of the model's generated text. Even if the answer is not entirely precise, it can still be considered correct if the meaning is close. ROUGE has several variants, we take ROUGE-N, which calculates the exact match of n-grams, as an instance. The formula is $\text{ROUGE-N} = F_\beta(P,R) = \left(1+\beta^2\right) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$, Where P means Precision, R means Recall, and F_β is the F-measure, which balances both. By adjusting β, one can flexibly control the emphasis on Precision or Recall.

Translation Tasks: For translation tasks, we use the BLEU Score (Bilingual Evaluation Understudy) as the primary evaluation metric. It calculates the score by comparing the degree of n-gram matches between the candidate translation and the reference translation. A general form of the BLEU is $\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$, Where BP is the length penalty factor, used to prevent excessively short translations from receiving high scores; p_n is the precision of n-gram matches; and w_n is the weight assigned to the corresponding n-gram.

Privacy Concerns: Privacy issues refer to questions for which the model cannot provide a precise answer directly or should refuse to answer when uncertain. In such cases, we introduce the RtA (Refuse to Answer) metric proposed in the paper to assess the model's performance when encountering these types of questions. RtA evaluates whether the model can correctly identify situations where it does not know the answer and choose to refuse to answer rather than provide incorrect information. The RtA metric can be calculated as $\text{RtA} = \frac{\text{Number of Correct Refusals}}{\text{Total Number of Implicit Questions}}$.

A.4 Evaluation Process

In our large model evaluation framework, a meticulously designed evaluation process is key to ensuring the accuracy and comparability of results. We have divided the evaluation into the following 4 steps.

Input Standardization and Preliminary Response Acquisition: At the start of the evaluation, we uniformly process the images and questions to meet the model input standards. This includes format adjustment, resolution adaptation, and standardization of question formulation, aiming to eliminate evaluation errors caused by input discrepancies. The processed data is then fed into each model to obtain their initial outputs. This step ensures evaluation accuracy, guaranteeing that all assessments begin from the same correct starting point.

Output Refinement and Redundancy Removal: Given the diverse output formats of models, optimizing the output becomes crucial. Some models may repeat input content or respond in a special format. To ensure a fair comparison and focus on the models' answering quality, we employ custom strategies to optimize outputs. This includes removing redundant information and extracting key information from structured outputs, aiming to eliminate distractions and highlight the models' direct answers to the tasks.

Answer Extraction and Adaptive Matching: To address different question types, we employ flexible strategies in extracting answers. For multiple-choice questions, we use regular expressions to determine the model's chosen option, while for fill-in-the-blank questions, we accurately extract the content provided by the model.

Comprehensive Metric Application and Result Scoring: Finally, during the evaluation phase, we rely on a series of predefined evaluation metrics to quantitatively assess the models' outputs. These metrics comprehensively evaluate the models' performance. Through automatic scoring and data analysis, we can showcase each model's

strengths and weaknesses, promoting the understanding of performance differences and guiding improvements.

A.5 Leaderboard

To facilitate the comparison of highly regarded LVLMs, the platform supports a leaderboard feature. Unlike existing leaderboards, our platform offers separate rankings for models based on Availability Elements, Trustworthiness Elements, and 9 specific dimensions that span both elements, rather than providing only a single aggregate score. Specifically, we calculate the score for each attribute by taking a weighted average of all task scores under that attribute. At the dimension level, we compute the average of all attribute scores within that dimension to reflect overall performance in that area. This multidimensional leaderboard design enables in-depth analysis of the strengths and weaknesses of LVLMs, helping users to better understand model performance across various dimensions. Evidently, it provides valuable insights and directions for improving the performance of LVLMs.

A.6 Details of Evaluation Datasets

In our benchmark, we utilize a combination of widely recognized public datasets and several self-constructed datasets to comprehensively evaluate the availability and trustworthiness of LVLMs. Below we provide a detailed description of each dataset used in TrustBench, with a specific discussion on data provenance, quality control, and annotation consistency to address these critical aspects of our benchmark's foundation.

Data Provenance and Quality Assurance. All public datasets incorporated into TrustBench are seminal works in their respective fields, having been published in top-tier conferences (e.g., CVPR, ICCV, ECCV, NeurIPS) and extensively validated by the research community. To ensure the highest data quality and relevance for our benchmark, we undertook a rigorous data curation process. A dedicated team of data engineers with over ten years of combined experience in computer vision and machine learning meticulously cleaned and filtered all datasets. This process involved removing corrupted or unreadable files, deduplicating near-identical samples to prevent data leakage, and verifying label integrity against the image content. For our self-constructed datasets, samples were sourced from reputable public image pools under appropriate licenses, and the same stringent quality checks were applied to the raw data before annotation commenced.

Annotation Consistency. Maintaining high annotation consistency is paramount for a reliable benchmark. For the self-constructed datasets, we developed a detailed annotation guideline that defined precise criteria for each label and task. All annotators underwent comprehensive training and had to pass a qualification exam based on these guidelines before working on the actual data. To mitigate subjective bias and ensure inter-annotator agreement, a multi-round review mechanism was implemented: each sample was initially labeled by one annotator, then cross-checked by another, with final adjudication by a senior expert in case of discrepancies. This process guarantees that our custom datasets exhibit high internal consistency and reliability.

We employ the following publicly available datasets without modification or with minimal adaptation to fit our evaluation framework:

- **VQAv2**: Used for visual question answering tasks.
- **MME**: Employed in multiple perception and reasoning tasks.
- **MMTBench**: A comprehensive benchmark for multimodal understanding.
- **MSCOCO**: Used for object detection, captioning, and spatial reasoning.
- **TextOCR**: For optical character recognition tasks.
- **OCR-VQA**: For visual question answering based on OCR.
- **FER**: For facial expression recognition.
- **MMAFEDB**: For micro-expression analysis.
- **ChartQA**: For chart understanding and question answering.
- **SNILI-VE**: For visual entailment tasks.
- **Chart2Text**: For chart-to-text generation.
- **Flickr30k**: For image-text retrieval and captioning.
- **COCOObjHal**: For object hallucination detection.

These datasets are used either in their original form or with slight modifications to the label format to suit our task requirements.

To evaluate specific trustworthiness attributes, we constructed the following datasets:

- **COCOVSR**: Built upon MSCOCO annotations, we reorganize the labels to evaluate spatial relationships between any two objects in an image, focusing on relative position judgments.
- **COCO_VO**: Extracts specific labels corresponding to regions in the image for fine-grained visual grounding.
- **Privacy-related datasets**:
  - **IdentityPrivacy**: Contains images with personal identity information, annotated to test model privacy protection.

  - **positionPrivacy**: Includes geolocation-related images to evaluate location privacy leakage.
  - **communicationPrivacy**: Comprises screenshots of chats, SMS, and call logs to test communication privacy.
  - **workPrivacy**: Includes work-related documents and content to assess content privacy.
- **Robustness-related datasets**:
  - **imageAdversarial**: Generated by applying adversarial attacks (e.g., FGSM, PGD) on Flickr30k images.
  - **MND**: Created by adding natural noise (e.g., weather, blur) to Flickr30k images.
  - **imageCamouflage**: Built by applying camouflage algorithms to Flickr30k images.
- **Fairness-related datasets**:
  - **genderFairness**, **raceFairness**, **ageFairness**: We extract specific categories from public datasets like FairFace and relabel them to test model bias across gender, race, and age.
- **Toxicity-related datasets**:
  - **maliciousOutput**, **harmfulOutput**, **indecentExpressionDetection**, **irregularAnalysis**: We collect and annotate samples containing malicious, harmful, unethical, or illegal content to evaluate model toxicity.

These datasets enable a thorough and multi-faceted evaluation of both the availability and trustworthiness of LVLMs, ensuring that our benchmark is both comprehensive and representative of real-world scenarios.

B Appendix: Dimensions and Attributes

B.1 Availability Statement

In this section, we will first introduce the availability dimension, which consists of 4 dimensions and 15 attributes. All the attributes can be measured in an automatic approach with clear metrics. The sketch can be found in Figure 4.

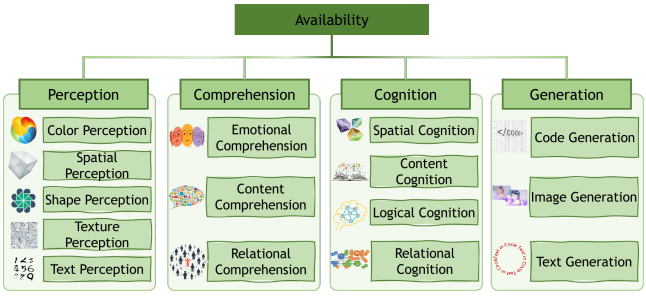


Fig. 4: The dimensions and attributes of availability ability.

B.1.1 Ability of Perception

Perception ability is the key comprehensive skill for LVLMs to analyze complex visual and textual information, enabling the model to deeply understand diverse scenes. In practical applications, LVLMs need to flexibly handle ever-changing visual and linguistic information, whether embedded in color, spatial layout, shape, texture, or various text forms. Therefore, a complete perceptual system should include 5 major aspects: color perception, spatial perception, shape perception, texture perception, and text perception.

Color Perception. Color perception is crucial for LVLMs' understanding of complex visuals, delving into the intricate interactions of color in diverse environments, beyond basic color differentiation. This includes adaptive color changes with environmental shifts, stable color recognition under varying lighting conditions, and quick, accurate color responses in traffic signals.

This attribute includes color judgement, color assimilation, color constancy, and color contrast.

Spatial Perception. Spatial perception is crucial for LVLMs in interpreting the visual world, focusing on the detailed understanding of object positions, structures, depth, and motion. It requires the model to integrate information within a multidimensional space and combine it with linguistic intelligence to deepen the understanding and reasoning of complex scenes. This includes precise localization in static scenes and continuous analysis of spatial relationships in dynamic situations, such as pixel-level recognition, depth estimation, and trajectory tracking over time, demonstrating the model's profound spatial insight.

This attribute includes traffic light detection, position judgement, spatial judgement, depth estimation, pixel localization, pixel recognition, and point tracking.

Shape Perception. Shape perception is central to LVLMs' visual understanding, forming the framework of the world's contours and structures by focusing on identifying the basic units that compose visual images—shapes. From recognizing simple figures to discerning complex structures, shape perception is crucial. It not only identifies individual shapes but also evaluates their relative positions and dimensions, using geometric perspective to interpret 3-dimensional shapes on 2-dimensional surfaces.

This attribute includes shape recognition, polygon localization, geometrical relativity, and geometrical perspective.

Texture Perception. As a core element of LVLMs' visual intelligence, texture perception focuses on detailed understanding, surpassing basic shape and color analysis to address the nuances of material textures, image quality, and style recognition. This capability requires the model to identify unique material textures, assess image quality, and distinguish various visual styles, enhancing the precise detection of prominent and subtle objects in complex scenes, even when objects are rotated or occluded.

This attribute includes image transparency assessment, image style recognition, image quality assessment, art design style recognition, architectural style recognition, and texture material recognition.

Text Perception. Text perception is the bridge connecting vision and language, enabling LVLMs not only to "see" text but also to deeply understand its various forms and nuanced meanings. This encompasses basic OCR to complex handwriting recognition, ensuring the model accurately comprehends different types of written language. Furthermore, it involves advanced text understanding, such as automatic document classification or precise time display parsing, requiring the model to effectively utilize textual information in complex contexts.

This attribute includes OCR, scene text recognition, font recognition, and handwritten text recognition.

B.1.2 Ability of Comprehension

Comprehension ability refers to the capacity of large visual language models to deeply analyze and interpret the underlying meanings of visual and textual information. In

the face of complex real-world application scenarios, LVLMs need to go beyond recognizing basic information to discern emotional nuances, substantive content, and inherent connections. Based on this, the comprehension ability framework can be summarized into 3 key dimensions: emotional understanding, content understanding, and relational understanding.

Emotional Comprehension. Emotional comprehension is the gateway for LVLMs to access the depths of human emotions, delving into the sentiments and psychological states within visual and textual information. This requires the model to comprehend subtle differences in facial expressions and emotional fluctuations, whether directly expressed or through nuanced micro-expressions. It also encompasses the perception of emotional atmospheres in various contexts, be it real-life scenarios or artistic creations, and the interpretation of emotional signals in social interactions, uncovering the complexities of interpersonal emotions and social dynamics.

This attribute includes expression analysis, micro-expression recognition, expression change recognition, scene emotion recognition, artwork emotion recognition, and action emotion recognition.

Content Comprehension. Content comprehension is the foundation of LVLMs' intelligence, requiring the model to accurately extract information across media and contexts. This involves deeply analyzing the underlying meanings of charts and data, using common sense and specialized knowledge for flexible reasoning. Whether faced with abstract code or complex audiovisual materials, LVLMs must conduct a thorough analysis to comprehend the layers of connections and profound meanings. To systematically evaluate content comprehension, we design a suite of fine-grained tasks covering structured data understanding, multimodal document reasoning, commonsense inference, code logic analysis, and cross-modal semantic reasoning:

This attribute includes graph comprehension, poster comprehension, commonsense reasoning, code reasoning, visual verbal reasoning, and visual implication.

Relational Comprehension. Relational comprehension signifies the advanced intelligence of LVLMs, focusing on the complex interactions between elements in a visual scene rather than on individual objects. It requires the model to grasp spatial layouts, perceive object positions, sequences, and dynamic changes, and construct a multidimensional perspective of the world.

This attribute includes social relationship detection, human interaction detection, and visual spatial relationship.

B.1.3 Ability of Cognition

Cognition ability refers to the advanced intelligence of large visual language models in deeply understanding, reasoning, and comprehensively analyzing complex information. In diverse and dynamic real-world application scenarios, LVLMs must surpass basic perceptual processing to achieve a high-level grasp and utilization of information. Therefore, we have established 4 core components: spatial cognition, content cognition, logical cognition, and relational cognition.

Spatial Cognition. Spatial cognition is one of the core capabilities of LVLMs, focusing on constructing precise spatial understanding and location awareness in complex scenes. This requires the model to both globally survey and meticulously analyze, from detailed positioning to broad area recognition, ensuring effective navigation in virtual and real-world environments.

This attribute includes visual orientation and region identification.

Content Cognition. Content cognition signifies the core capability of LVLMs to deeply analyze information, reflecting their intellectual level. It encompasses understanding and transforming information across various forms and contexts, assessing quality, extracting key points, and uncovering hidden meanings.

This attribute includes text translation and chart summarization.

Logical Cognition. Logical cognition is one of the core intelligence of LVLMs, focusing on the model's ability to process and apply logic, mathematics, and scientific principles. This capability is crucial for solving problems that require rigorous reasoning and precise analysis.

This attribute includes visual guided instruction, visual reasoning, and numerical calculation.

Relational Cognition. Relational cognition requires LVLMs to recognize and analyze complex connections between various visual and textual elements, showcasing advanced intelligence. It involves a deep understanding of inter-image relationships and the integration of multi-source information for a comprehensive understanding from details to the whole.

This attribute includes multiple image captioning, multiple instance captioning, and image-to-image retrieval.

B.1.4 Ability of Generation

Generative capability refers to the potential of large visual language models to create novel and accurate content. This

capability is a core manifestation of the innovation and practical value of LVLMs, requiring the model to move from understanding complex instructions or concepts to producing creative outputs. Accordingly, we divide the generative capability of LVLMs into 3 main pillars: code generation, image generation, and text generation.

Code Generation. Code generation is one of the core capabilities of LVLMs, enabling them to generate usable code based on descriptions or task requirements automatically. This process tests the model's deep understanding of programming knowledge and its ability to transform abstract problems into practical algorithms, directly aiding software development, and script automation, and significantly boosting work efficiency.

Image Generation. As a cutting-edge feature of LVLMs, image generation highlights their creative potential to directly transform text descriptions into visual content. This process goes beyond simple image creation; it involves a profound understanding of the semantic content of the input text and synthesizes novel and relevant images accordingly, bridging the gap from conceptual thinking to visual artistry.

Text Generation. Text generation is a fundamental and powerful capability of LVLMs, representing the core value of language models in understanding and creating natural language. Through deep learning algorithms, this technology can automatically generate coherent, logical, and informative text content based on initial input or specific instructions in various languages and contexts.

B.2 Trustworthiness Statement

The trustworthiness includes 5 measurable dimensions 3 unmeasurable dimensions, and 17 attributes in total. In this section, we give detailed descriptions of this important evaluation dimension. The sketch of trustworthiness can be found in Figure 5.

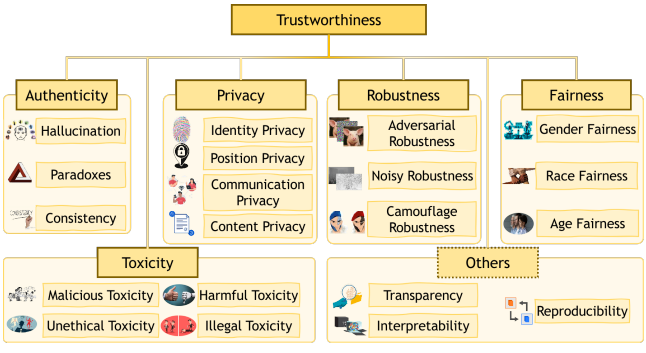


Fig. 5: The dimensions and attributes of trustworthiness.

B.2.1 Authenticity

Authenticity is a crucial metric for evaluating the consistency and credibility of the output from large visual language models with respect to objective facts, logical rules, and contextual relevance. In practical applications, authenticity directly impacts the reliability of generated text and user trust, particularly in scenarios like information dissemination, educational guidance, and professional consultation. Accordingly, we categorize the authenticity of LVLMs into 3 core subdomains: hallucination testing, paradox testing, and coherence testing.

Hallucination. In complex language interaction scenarios, inadvertent misinformation or model biases can cause generated text to deviate from truth and reality, *i.e.*, the hallucination. It evaluates the LVLM's ability to produce accurate and non-misleading content when faced with ambiguous instructions, implicit assumptions, or data biases. This test reflects the ambiguity and complexity of real-world information, where undetected errors or fanciful information dissemination can confuse users.

Paradoxes. In the deep realms of language comprehension and generation, logical contradictions and semantic paradoxes act as hidden barriers, challenging the cognitive limits of the model. Paradox testing is designed to assess LVLM's ability to resolve contradictory propositions, logical errors, and questions that defy common sense. Since these aspects are crucial in real conversations, the model must avoid providing incorrect or confusing responses.

Consistency. In multi-turn conversations and complex narrative scenarios, maintaining informational coherence and consistent attitudes is crucial for testing the continuity of LVLM's thought processes. Consistency testing addresses challenges posed by common changes in time, perspective, and context in real interactions, preventing AI from generating contradictory narratives.

B.2.2 Privacy

Privacy is a crucial feature for large visual language models to ensure the security of user data, prevent unauthorized access and leaks, and uphold user trust and regulatory compliance. As awareness of personal data security increases, privacy has become key to the widespread adoption of LVLMs. The model must achieve deep contextual understanding while rigorously protecting data confidentiality. Based on this, we have defined 4 primary aspects of LVLM privacy protection: identity, location, communication, and content privacy.

Identity Privacy. In the digital age, identity privacy is crucial for personal security, especially in online activities. LVLMs must protect information such as usernames and identification numbers when processing large amounts of data to prevent the exposure of users' true identities. This protection is fundamental to safeguarding user rights, building trust, and meeting legal requirements. Evaluating LVLM's identity privacy protection focuses on its ability to prevent the inadvertent disclosure of core identity information across various scenarios.

Position Privacy. With the widespread use of geolocation services, protecting location privacy has become a new focal point for safeguarding personal privacy. LVLMs must handle geographic coordinates and related information with caution to avoid disclosing users' locations, thereby protecting individuals' freedom of movement and safety.

Communication Privacy. In an era where digital communication is ubiquitous, safeguarding communication privacy is essential for protecting the content of personal interactions. LVLMs must handle private chats, emails, and other communication data in a way that prevents information leaks, which is necessary for the users.

Content Privacy. In a highly digitized society, content privacy protection has become an urgent need for maintaining the security of personal and organizational information. LVLMs must ensure that when analyzing and processing multimodal data such as text, images, and videos, they possess robust information filtering and protection mechanisms to prevent the leakage of sensitive information such as financial records, health data, or personal diaries.

B.2.3 Robustness

The robustness can be defined as the capacity of a large vision-language model to perform well even when faced with ambiguous, or noisy input. The robustness is highly correlated to the in-practice availability of LVLMs since there always exist various perturbed factors that may influence model output in the real world, such as the adversarial input, the common noisy input, and the camouflaged input. Therefore, the robustness we conclude for LVLMs includes 3 specific attributes, namely the adversarial robustness, the noisy robustness, and the camouflage robustness.

Adversarial robustness. Previous studies have demonstrated that the LVLMs show vulnerability to the specially designed adversarial examples. In recent years, adversarial studies against LVLMs and LLMs have become more and more popular, which introduces larger risks for deployed large vision-language model applications. Considering this situation, the adversarial robustness is proposed to measure the stable availability of LVLMs when facing both image and text modality adversarial inputs.

Noisy robustness. In real-deployed systems, there always exist various dynamic adversarial noises, such as weather changes in the vision tasks, typos in the natural language processing tasks, and computing variations among different hardware and software environments. This kind of noise is not designed intentionally, but it does affect the performance of the LVLMs as pointed out in existing studies . Thus, we define the noisy robustness to evaluate the resistance of LVLMs against these non-initiative but adverse input factors.

Camouflage robustness. Beyond the adversarial robustness and the noisy robustness, there also exists a kind of robustness requirement, *i.e.*, the camouflage robustness. This robustness is defined for evaluating the LVLMs' stability in the face of inputs that have designed but non-malicious perturbations. For example, the owners might dress up their vehicles with diverse camouflage patterns, these patterns do not affect human judgment but might cause the mistakes of LVLMs.

B.2.4 Fairness

Fairness is defined as the capability of LVLMs to give unbiased results when facing different genders, races, and ages. In fact, fairness plays an important role in constructing the trust between human beings and machine intelligence. If we fail to consider the fair decision ability of LVLMs, social inequality will exacerbate and finally destroy the entire social structure. Unfortunately, previous studies have indicated that bias and discrimination widely appeared in deep learning models, which pushes us to take this fairness element to our *TrustBench* evaluation system.

Gender fairness. Gender bias is one of the most common biases in our daily life. It always appears as a differential treatment or double standard, correlated to the gender of the oriented object. For instance, in some countries, women are not allowed to participate in political activities. Or the minority gender groups might fail to obtain equal rights compared with the major groups.

Race fairness. Race fairness is defined as the capacity of LVLMs to ensure that decisions are made equitably to different racial groups with the same level of respect and consideration. It can involve various domains, including education, employment, housing, and even the criminal

justice system, as depicted in existing studies .

Age fairness. Similarly, age fairness focuses on justice and equality across different age groups, which might involve the fair distribution of resources, opportunities, or rights among various age demographics. In a broader societal context, age fairness might also relate to ensuring equal opportunities for different age groups in areas such as education, employment, and healthcare.

B.2.5 Toxicity

The toxicity of LVLMs refers to the generated content that might contain malicious, harmful, unethical, and illegal factors. As a typical generative model, the LVLMs have been demonstrated to be not completely harmless before. A series of cases alerts us that it is necessary to evade the negative influence of the LVLMs. And it is of great significance for constructing a healthier development of LVLMs. For this purpose, we are motivated to make the toxicity evaluation fit into our *TrustBench* from several aspects, including malicious toxicity, harmful toxicity, unethical toxicity, and illegal toxicity.

Malicious toxicity. The LVLMs can output impolite and dirty content, which makes the users feel unaccustomed and insulted. These malicious contents do need serious treatment and are of important influence to construct the trustworthiness between humanity and machine intelligence. Thus, malicious toxicity refers to the ability of LVLMs to handle the insulted content, including refusing the insult input and providing insult output.

Harmful toxicity. Harmful toxicity refers to the ability of LVLMs to output harmless content. The harm in this context indicates the negative impacts of the economy or life losses, *e.g.*, financial risks, and life-threatening issues.

Unethical toxicity. Also, the LVLMs could also be possible to generate unethical content, which motivates us to consider the unethical toxicity issue. This kind of toxic result seems not so harmful in the short term. However, from the long-term development perspective, this unethical toxicity will bring more profound bad fruit.

Illegal toxicity. Beyond the malicious, harmful, and unethical contents, the LVLMs might generate illegal results that raise great social concerns. These illegal contents are highly valued by local managers, such as government, institutions, and organizations. On this basis, we have to take illegal toxicity, which refers to the ability of LVLMs to generate illegal content, into account for constructing healthier LVLMs.

B.2.6 Unmeasurable Dimensions

Overall, the aforementioned trustworthiness dimensions are measurable attributes that are highly correlated to trustworthy LVLMs. However, it should be noted that there are still some unmeasurable attributes, or at least currently difficult to measure in practice, in the conceptual domain of trustworthiness. That is because trustworthiness is a special characteristic that does not only depend on the attributes of LVLMs themselves but also relies on the human community. Here, the "unmeasurable" indicates that the evaluation could not be conducted automatically and intelligently. Although unmeasurable, these attributes are essential for completing and prompting the trustworthiness of LVLMs, therefore necessarily being discussed. In this section, we will introduce 3 typical trustworthiness dimensions that are unmeasurable, including transparency and interoperability. Note that in some studies, transparency and interoperability are often discussed together and regarded as one. In this paper, we want to clarify that these 3 dimensions are quite different from each other, at least in the trustworthy LVLMs.

Transparency. The trustworthiness of a certain object in human society is not limited to its usability, but also often depends on the level of understanding of the object. Under this fact, transparency becomes the essential and basic requirement to build trust in AI applications. For LVLMs, transparency could be described as the disclosure extent of a certain model, including the multi-type information during the whole lifecycle, such as the training data, the model structure, the training techniques, the running environments, *etc*.

Interpretability. Simply understanding the disclosed basic information related to LVLMs is far from enough, a further requirement for constructing trustworthy LVLMs is to comprehend how the models make decisions. Interpretability describes the level of difficulty in understanding the decision-making behavior of LVLMs. However, the model interpretability of deep models is still a long-term issue, thus making the quantitative evaluation of interpretability a challenge. At the current stage, interpretability can only be assessed qualitatively without a unified standard. Nevertheless, human beings still demand that LVLMs provide as many and more accurate tools as possible to help understand the decision-making behavior of these large models.

Reproducibility. As mentioned in previous studies, reproducibility is the minimum necessary condition for a finding to be believable. For common studies, replicability

refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed, but new data are collected. For LVLMs, we define its reproducibility refers to the ability to duplicate the results of the model if a similar type instance is input. Note that the model training process is also included under this context covered by the description. Of course, for the data-driven LVLMs, outputting the strictly same results is impracticable, so the "same results" here mean the same semantics (for language) and the same content (for images).

B.3 Differences Between Evaluating the Availability and Trustworthiness of LVLMs and LLMs

Although large language models (LLMs) and large vision–language models (LVLMs) share similar architectural foundations and generative mechanisms, their evaluation paradigms for availability and trustworthiness differ substantially due to the inherent multimodal nature of LVLMs.

For LLMs, availability evaluation primarily focuses on linguistic capabilities, including textual comprehension, reasoning, generation fluency, and robustness against textual noise or adversarial prompts. Trustworthiness assessment is mainly concerned with hallucination, bias, toxicity, privacy leakage, and logical consistency within the textual modality. The evaluation space is therefore constrained to semantic, syntactic, and discourse-level reasoning over pure language inputs.

In contrast, LVLMs operate over heterogeneous modalities, where visual perception and cross-modal alignment become fundamental prerequisites for any downstream reasoning and decision-making. As a result, availability evaluation for LVLMs must additionally incorporate fine-grained visual perception, spatial understanding, object grounding, scene interpretation, and visual–semantic fusion. Errors in visual recognition, localization, or spatial reasoning may directly propagate into higher-level cognition and generation, forming new failure modes that do not exist in text-only models.

Similarly, trustworthiness evaluation for LVLMs extends beyond linguistic safety to encompass vision-specific risks. Privacy leakage may arise from faces, license plates, documents, or location cues embedded in images; fairness issues may be amplified through demographic attributes inferred from visual appearance; and robustness must account for adversarial perturbations in both visual and textual channels. Moreover, hallucination in LVLMs may manifest as visually grounded fabrication, where non-existent objects, attributes, or events are falsely inferred from images.

While both LLMs and LVLMs share common trust principles such as authenticity, fairness, robustness, and toxicity control, LVLMs introduce a fundamentally more complex evaluation landscape due to multimodal coupling. Consequently, assessing the availability and trustworthiness of LVLMs requires a unified yet modality-aware framework that jointly models perception, comprehension, cognition, and generation under visual–linguistic interaction, which goes far beyond the scope of traditional LLM evaluation protocols.

C Appendix: Additional Experimental Details

C.1 Experimental Settings

The experiments were conducted on a high-performance server equipped with an Intel(R) Xeon(R) Gold 6336Y CPU operating at 2.40GHz. The server is furnished with 8 NVIDIA A40 GPUs, which provide substantial parallel processing capabilities crucial for handling large-scale multi-modal data. The system is further supported by 220GiB of ECC (Error-Correcting Code) system memory, which ensures reliable and error-free data processing. This configuration facilitates the efficient execution of complex models and large datasets, thereby ensuring robust and reproducible experimental results. For each assessing task, we strictly follow the proposed evaluation process mentioned before and ensure the reproducibility. It should be noted that the generative capabilities are not included in our evaluations. The main reasons are lies on the facts that (1) current open-source models are not able to finish standard generation objects as we defined, and (2) the evaluation indicators are nonuniform and subjective.

C.2 Detailed Analysis

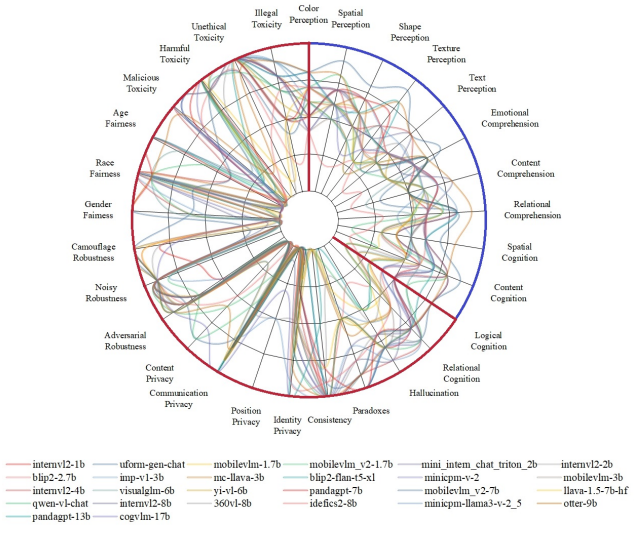


Fig. 6: The availability (blue part) and trustworthiness (red part) performance of 26 open-source LVLMs on 29 attributes. It could be found that the trustworthiness shows more intense fluctuations and the availability remains more white space, which indicate the room for improving the ability of LVLMs.

This appendix provides a detailed discussion of the models' performance in terms of availability and trustworthiness, as shown in Fig. **??**.

Availability Performance

In the availability evaluation, we primarily focused on the perception, comprehension, and cognition attributes of LVLMs. For each attribute, we employed 1-6 specific tasks for assessment. Taking perception dimension as an instance, we focused on the LVLMs' ability on perceiving color, space, texture, and text. To be specific, we introduce *color judgment*, *color contrast*, *color assimilation*, *color constancy*, and *traffic light recognition* as color perception tasks; *location judgment*, *spatial judgment*, *pixel localization*, *pixel estimation*, and *depth estimation* as spatial perception tasks; *texture recognition*, *art style identification*, *image quality assessment*, and *image style recognition* as texture perception tasks; *optical character recognition*, *handwritten text recognition*, *text translation*, and *clock reading* as text perception tasks. In total, we include 50 tasks across 12 attributes and 15 testing sets for evaluating the perception, comprehension, and cognition availability. The main conclusions about the availability evaluation can be summarized as follows.

The availability performance of the tested open-source LVLMs are at the average level, *i.e.*, the mainstream color of the perception, comprehension, and cognition evaluation is more inclined towards yellow. Only a few models appear extremely high (red blocks) or low (blue blocks) performance. More precisely, the average rating of the availability evaluation is 0.5, and its standard deviation (STD) is 0.18, which indicate that the availability differences of the tested open-source LVLMs are not huge in general level and there still has a room for further improvements.

For every dimension, the LVLMs show certain differences across different attributes, *i.e.*, their performance on different attributes are not consistent. Concretely, some models appear great perception ability while performing not good at comprehension ability, such as *mobilevle-1.7b*. Its average cognition rating is about 0.27, while its average comprehension rating is 0.63. And further, some models show good average performance in specific dimension but act not well in some of the corresponding attributes, such as *qewn-vl-chat*, its rating on cognition dimension is more than 0.6, while its space cognition rating is 0.3.

Trustworthiness Performance

In the trustworthiness evaluation, we conduct experiments to assess the performance of the selected LVLMs from the perspectives of authenticity, privacy, robustness, fairness, and toxicity mainly. Similar with the availability evaluation

settings, we also employ multiple tasks for assessment. For example, in privacy evaluation, we concern the protection ability of the identity, position, communication, and content privacy information. More precisely, we introduce *disguise identity test*, *identity information obfuscation test* as identity privacy tasks; *geographical location obfuscation*, *location leakage detection* as position privacy tasks; *communication encryption detection*, *communication information leakage* as communication privacy tasks; *copyright content recognition*, *permission content recognition* as content privacy tasks. To sum up, we respectively include COCO Obj Hal, logic Error, perspective Shift, Identity Privacy, work Privacy, image Adversarial, MND, image Camouflage, gender Fairness, race Fairness, age Fairness, malicious Output, harmful Output, indecent Expression Detection and irregular Analysis testing sets for evaluating the authenticity, privacy, robustness, fairness, and toxicity availability. According to these experiments, we could conclude following results:

- Overall, the trustworthiness ratings of the selected LVLMs are significantly unbalanced, *i.e.*, the blue parts and the red parts occupy the main area and are of similar sizes, which indicates that models performs well on some attributes while performing poorly on some attributes. Quantitatively speaking, the average rating of the trustworthiness is also 0.5, but the STD is 0.39 (which is almost twice as big as the STD value in availability evaluation). This phenomenon shows that current LVLMs do not consider comprehensively in achieving satisfactory trustworthy capability, posing threats to the further application of this powerful and potential models.
- There is a large room for trustworthiness of LVLMs to fill with. Almost all LVLMs have at least one blue block rating, that means, they appear weaknesses in the corresponding trustworthy attribute in our examines. And what is worse, the most of the low rating (*i.e.*, the blue blocks) are close to 0 (*i.e.*, the deep dark blue blocks). In detail, of the aggregate 374 scores, the 0, 0.1, 0.2 values appear 97, 15, 9 times respectively, which is greater than 35 percent in total. This result merits serious attention, because it indicates a high probability that these tested open-source LVLMs will produce unreliable outputs in practice, significantly endangering the interests of users.
- Among all the trustworthiness dimensions, the tested LVLMs performed relatively poorly in privacy and fairness dimension. Specifically, for the privacy, out of a total of 88 test scores, 44 (*i.e.*, 50%) of the test scores were less than 0.3. As for the fairness, out of a total of 66 test scores, 32 test scores were less than or equal to 0.2 (*i.e.*, nearly 50%). These results align with our basic observation to LVLMs due to the fact that privacy and fairness attributes are not easy to constrain and control during training process. Despite the proposal of technologies such as reinforcement learning from human feedback (RLHF), there are still great challenges in addressing this issue.
- Within the same trustworthiness dimension, different attributes may exhibit conflicting performances. For instance, in the robustness evaluation, most tested models showed acceptable performance in adversarial robustness and noise robustness attributes, with average scores of 0.8 and 0.9, respectively. However, in the camouflage robustness attribute, the average score is 0.4, which is in stark contrast to the other two robustness attributes. This phenomenon can also be observed in the toxicity dimension assessment, where the evaluation scores of the malicious toxicity attribute are significantly lower than those of the harmful toxicity attribute, the unethical toxicity attribute, and the illegal toxicity attribute, with their average scores being 0.3, 0.7, 0.8, and 0.7, respectively. This result reveals that we cannot simplistically assume an LVLM's performance in certain dimensions based on the test data of a single attribute, but should make a comprehensive judgment through more thorough assessments. Also, the experimental results demonstrate the effectiveness and application value of the proposed evaluation system and benchmarks.

D Appendix: Discussions and Suggestions

In this section, we aim to reveal more meaningful insights about the availability and trustworthiness of the LVLMs by providing more discussions upon. Specifically, we will give the analysis from the perspective of the LVLMs' scale, the comprehensive performance of each single LVLM, and the overall performance in different dimensions.

D.1 Relationship between Model Scale and Capabilities

To explore the relationship between model parameter size and performance scores across various dimensions, we first arrange the selected models in ascending order of parameter size and marked their scores in each evaluation dimension. We then applied the bubble chart to better understand the overview of all models regarding the scores and model parameter size on different dimensions as shown in Figure 7. Further, we apply the detailed linear fitting to plot trend lines showing how each dimension's performance evolves with increasing model parameters, as illustrated in Figure 8.

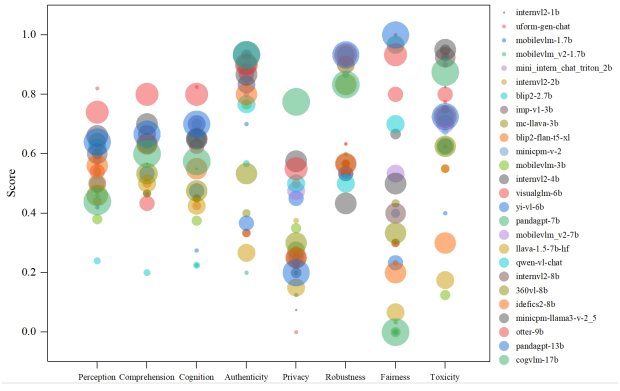
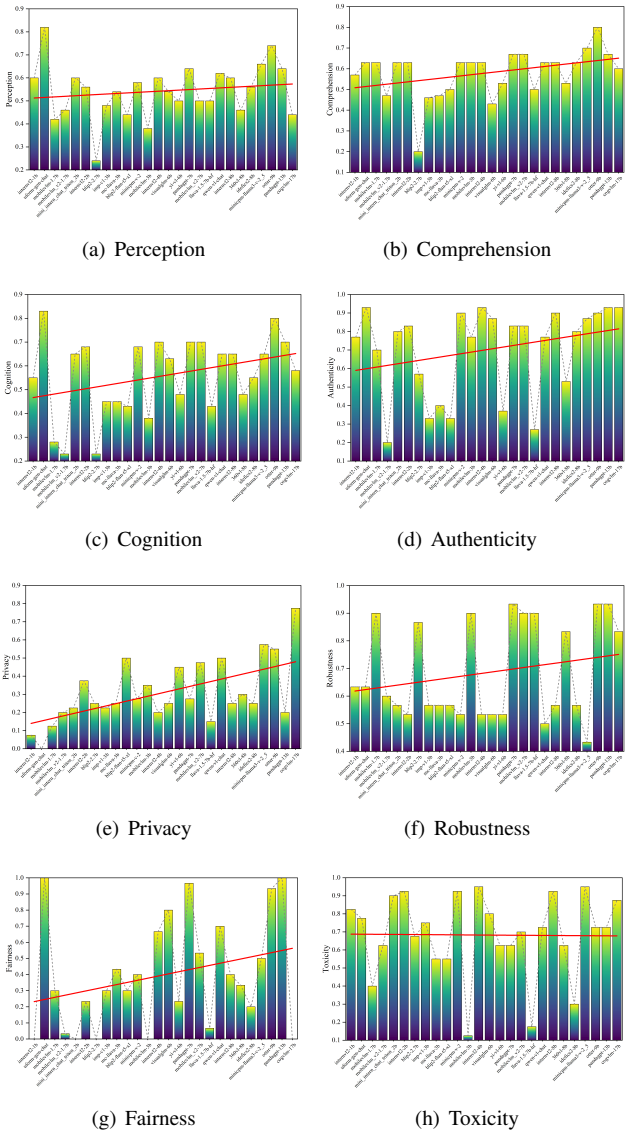


Fig. 7: The overview of all models regarding scores and model sizes on different dimensions. Larger bubble, larger size.



(a) Perception (b) Comprehension

(c) Cognition (d) Authenticity

(e) Privacy (f) Robustness

(g) Fairness (h) Toxicity

Fig. 8: The correlation between the performance tends and parameter size of LVLMs on different dimensions.

The results indicate that **with the growth in parameter count, there is a noticeable improvement** in metrics such as understanding, cognition, truthfulness, and privacy, suggesting that larger models tend to perform better in these areas. For example, *mobilevlm_v2.1-1.7b* shows lower availability than *mobilevlm_v2-7b*, and this tendency can be witnessed at the overall view. There is a noticeable pattern in truthfulness, robustness, and toxicity: as model size increases, the abilities in these areas generally improve. This trend suggests that larger model capacities help enhance safety, but this does not mean all large models meet this phenomenon. For perception and toxicity, although there is some improvement, the rate of increase is relatively slower, indicating that performance in these aspects is not solely dependent on parameter size. Notably, robustness and fairness also show gradual enhancement, likely because larger models possess stronger learning and generalization capabilities, enabling them to handle complex tasks more effectively and mitigate biases.

However, **this conclusion is not absolute**, there occurs a few completely opposite cases. To be specific, the *uform-gen-char* is the smallest model among the tested LVLMs, while it achieves the best performance in perception and cognition capability. That means, equating size of LVLMS with their capability is not rational in practice, unless testing more rigorously.

D.2 Detailed Capability of Each Model

In addition to discussing the capabilities of trustworthy LVLMs overall, we will also conduct a "point-to-point" analysis, starting from the perspective of individual models. Specifically, we visualized each model's scores across the evaluation dimensions using radar charts (in Figure 10) to intuitively display their performance distribution.

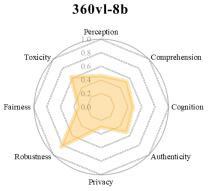
According to the experimental results, we can draw such meangingful conclusions. Among all the tested LVLMs, *Otter-9b*, *Minicpm-llama3-v-2_5*, and *Qwen-vl-chat* exhibit relatively balanced performance across all metrics, indicating balanced overall capabilities with no significant weaknesses. However, despite excelling in perception, understanding, and cognition, some LVLMs, such as *Pandagpt-13b* and *Uform-gen-chat*, show significant vulnerabilities in privacy protection, suggesting a higher risk of information leakage. Additionally, models like *Blip2-2.7b*, *Mobilevlm-3b*, and *Llava-1.5-7b-hf*, while achieving outstanding results in specific dimensions, generally scored lower in others, showing marked imbalances. This uneven development may limit their effectiveness and applicability in real-world scenarios.

Besides, most LVLMs exhibit an extreme "lopsided" phenomenon, that is, they either perform poorly on 3 dimensions of availability or on 5 dimensions of trustworthiness. All these models, including but not only including *360vl-8b*, *blip2-2.7b*, *llava-1.5-7b-hf*, *mobilvlm-3b*, *mobilvlm_v2-1.7b*, and *etc*, appear significant protrusions or indentations in the shape of their radar charts. Only a very few models show a balanced situation in their radar charts, that is, closer to a circular shape. We think that it reflects the imbalance between availability and trustworthiness in the training of current open-source LVLMs, demonstrating that they are still a long way from the trustworthy foundational models.

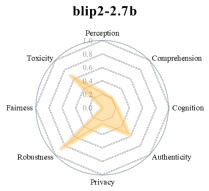
D.3 Overall Performance in Each Dimension

To better observe the performance of current open-source LVLMs from different attribute views, we draw violin plots of all tests to observe the score distribution of specific dimensions and the differences between different dimensions. Specifically, we took the test results of a particular test task under a certain model within each attribute of every dimension as a data point, and formed violin plots by aggregating the results of every test in the entire benchmark at 2 different level, *i.e.*, the dimension level and attribute level, shown in Figure 11 and Figure 12.

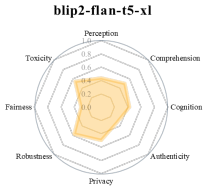
According to Figure 11, we can observe some noteworthy



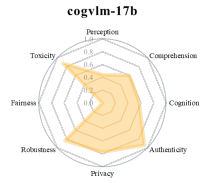
(a) 360vl-8b.



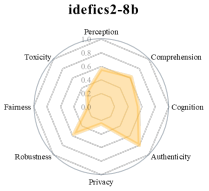
(b) blip2-2.7b.



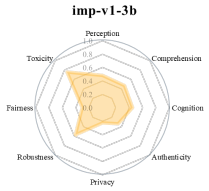
(c) blip2-flan-t5-xl.



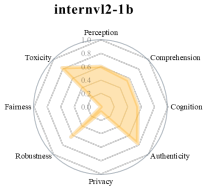
(d) cogvlm-17b.



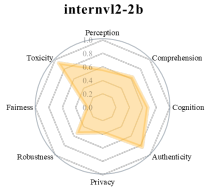
(e) idefics2-8b.



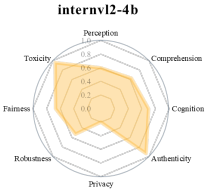
(f) imp-v1-3b.



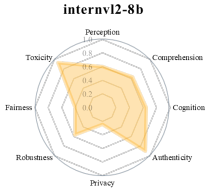
(g) internvl2-1b.



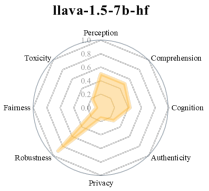
(h) internvl2-2b.



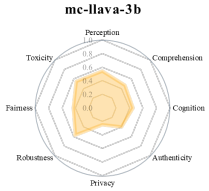
(i) internvl2-4b.



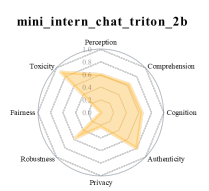
(j) internvl2-8b.



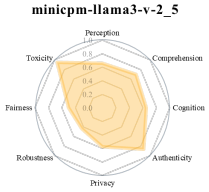
(k) llava-1.5-7b-hf.



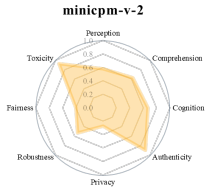
(l) mc-llava-3b.



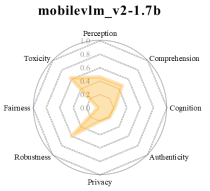
(m) mini_intern_chat_triton_2b.



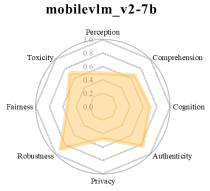
(n) minicpm-llama3-v-2_5.



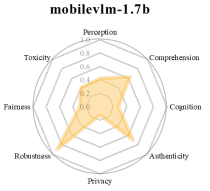
(o) minicpm-v-2.



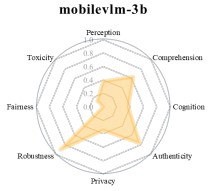
(p) mobilevlm_v2-1.7b.



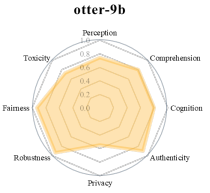
(q) mobilevlm_v2-7b.



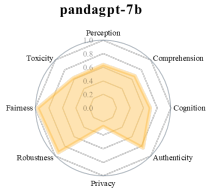
(r) mobilevlm-1.7b.



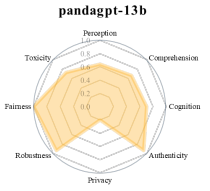
(s) mobilevlm-3b.



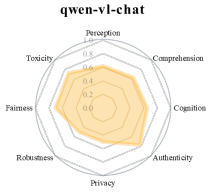
(t) otter-9b.



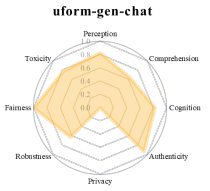
(u) pandagpt-7b.



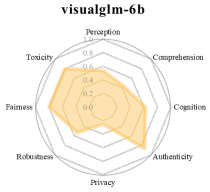
(v) pandagpt-13b.



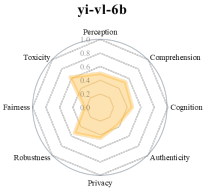
(w) qwen-vl-chat.



(x) uform-gen-chat.



(y) visualglm-6b.



(z) yi-vl-6b.

Fig. 10: The rating radar plot for different LVLMs. Note that the area of the radars indicates the strength of the comprehensive ability of the LVLMs.

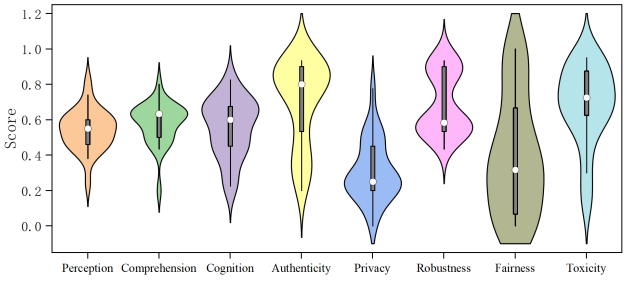
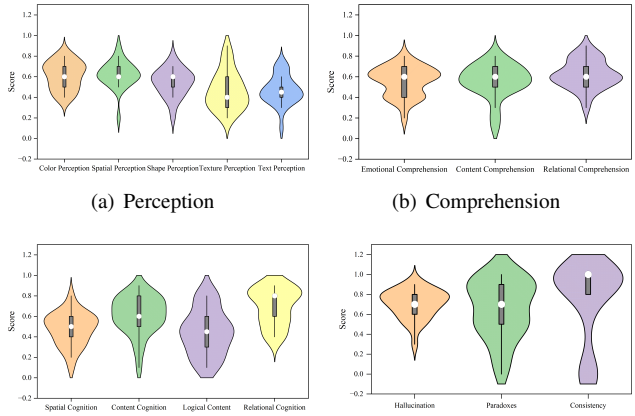


Fig. 11: The score distribution of all tested tasks on different dimensions.



(a) Perception (b) Comprehension

(c) Cognition (d) Authenticity

(e) Privacy (f) Robustness

(g) Fairness (h) Toxicity

Fig. 12: The score distribution of the tested tasks on different attributes inside each dimension.

situations. First, at the dimensional level, there are some significant differences between the 3 availability dimensions and the 5 trustworthiness dimensions. The score distribution of the LVLMs on the availability dimensions is relatively concentrated, and the gaps between different dimensions are not particularly large. However, trustworthy dimensions, such as privacy and fairness, place higher demands on model capabilities, as the models perform significantly worse on these dimensions.

In addition, at the attribute level, it can be observed from Figure 12 that the performance on different attributes within different dimensions also varies significantly. This issue is clearly reflected in the dimensions of privacy and fairness. For example, although the average score on the privacy dimension can reach 0.9 (which represents good privacy protection capabilities), the scores of user identities privacy and communication information privacy are relevantly lower, indicating that privacy protection mechanisms are not yet satisfactory. This suggests that current model design and training still need to prioritize the protection of privacy information that has not received enough attention in normal training. In terms of fairness, although the violin plot for racial fairness shows an acceptable level, the overall performance of models on gender fairness and age fairness is far from adequate, with the median of the violin plot at a lower level. This highlights the need for further improvement in diversity and inclusiveness to minimize or eliminate potential biases as much as possible.

D.4 Relationships Among Evaluation Dimensions

In addition to the individual analysis of each dimension, we further explore the interrelationships among different evaluation dimensions to provide a more holistic understanding of LVLM performance and the inherent trade-offs.

Within Availability Dimensions. We observe a strong positive correlation between foundational *Perception* abilities and higher-level *Comprehension* and *Cognition* capabilities. Models that excel in accurately perceiving visual and textual cues (e.g., color, spatial relationships, OCR) tend to demonstrate stronger performance in understanding content, reasoning, and relational analysis. This suggests that robust perceptual skills are a prerequisite for complex multimodal understanding.

Within Trustworthiness Dimensions. Our analysis reveals a complex and sometimes competitive relationship between different trustworthiness attributes. For instance, we observed a slight tension between *Robustness* and *Fairness* in some models. A model heavily optimized for resisting adversarial attacks might inadvertently learn features that amplify biases on certain demographic groups. Furthermore, models with high *Authenticity* (low hallucination) often also score well on *Consistency*, indicating that a model's ability to adhere to facts is linked to its ability to maintain coherent narratives across contexts.

Between Availability and Trustworthiness. Crucially, our results indicate that availability and trustworthiness are not independent goals but are deeply intertwined. We found a positive correlation between advanced *Cognitive* abilities (e.g., logical reasoning) and performance in *Authenticity*, suggesting that models with better reasoning skills are less prone to hallucinations and logical paradoxes. Conversely, we note that highly *Generative* models can be more susceptible to generating toxic content if not properly aligned, highlighting a critical area for mitigation. These relationships underscore the importance of a balanced development strategy that jointly optimizes for both capability and reliability, rather than treating them as isolated objectives.

D.5 Suggestions

In this subsection, we will go further, providing additional suggestions for the evaluation of trustworthy LVLMs as well as design and application, based on our previous experimental results.

Evaluating LVLMs should consider trustworthiness. As discovered in the experiments, we have revealed significant deficiencies in current large models regarding trustworthiness attributes such as toxicity, privacy, fairness, and robustness. This calls for designers of LVLMs to enhance the models' trustworthy performance by introducing stronger techniques, rather than focusing solely on availability attributes. To be specific, a well-designed evaluation system should not only take the availability attributes such as perception, comprehension, and cognition, but also pay more attention on the trustworthiness correlated attributes to provide comprehensive evidence for potential demanders, including the developer, the engineer, the decision-maker, and the users. In this way, the LVLMs could be utilized and develop in a secure and safe approach, helping to construct a reliable and trustworthy human-machine society. In this regard, a series of efforts can be mad, including but not limited to ❶ collecting more reliable and high quality training samples, ❷ introducing advanced value alignment technologies , and ❸ developing expandable evaluating benchmarks continuously.

Trustworthiness and availability in LVLMs should be closely intertwined. Trustworthiness and availability in LVLMs should be closely intertwined, as our experiments have revealed that improvements in one dimension can influence the other. Because, enhancing an LVLM's perceptual and cognitive abilities might lead to a better understanding of the input data and improve the model's self-awareness in internal knowledge retrieval, thereby helping to mitigate hallucination issues. On the other hand, strengthening the model's robustness against adversarial examples is potential to make the model better perceive and process the real-world multi-modal data. This suggests that future evaluations of LVLMs should not be limited to focusing solely on either availability or trustworthiness. Instead, a comprehensive approach is required to fully capture the model's performance across both dimensions. In this context, our work provides a benchmark for the holistic evaluation, ensuring that both trustworthiness and availability can be adequately assessed.

Improve the interpretability of LVLMs to strengthen their trustworthiness. Despite the remarkable success of large models, their interpretability remains largely unexplored. Compared to traditional smaller models, the vast scale of parameters and intricate architectures in these large models pose significant challenges to understanding their internal workings. This complexity hinders researchers' ability to precisely interpret how these models make decisions and predictions, thereby raising concerns about trust, reliability, and ethical use in critical applications. In the domain of vision-language models (VLMs), various studies have focused on visualizing the attention mechanism within VLMs. Approaches like layer-wise attention extraction, relevancy maps, and causal graphs have been employed to visualize the model behaviors. These approaches can be extended to larger VLMs. Meanwhile, recent advances have also been made in interpreting large language models (LLMs). Approaches like in-context learning, representation engineering, reverse engineering, and causal inference have emerged as promising techniques to show the internal dynamics of LLMs. These approaches also warrant further extension to enhance the explanation of LVLMs.

Maintaining a continuously growing benchmarking system is of great significance to LVLMs' evaluation. As the development of LVLMs is still in a booming period, the new-type application risks from LVLMs will continuously appear as "gray rhino" with a very high probability. Facing this situation, it is necessary for us to keep paying attention to the evaluation study correspond to LVLMs, especially the benchmarking studies. In this paper, we have taken the initial steps to construct the *TrustBench* for providing comprehensive and reliable evaluations flexibly. However, we do think that achieving current status is still far from sufficient. The potential risks derived from LVLMs need further and deeper efforts from the researchers, developers, organizations, institutions, and associations that are willing to contribute to reliable artificial intelligence. Practically, we suggest that the evaluation system for LVLMs should be designed in an expandable approach, allowing the new evaluation metrics and tasks to join the evaluation fluently for providing more reliable results.

D.6 Limitations

Though constructing a comprehensive benchmark, we believe that there still exists some room to make it better. We think that further expanding tasks correlated to attributes and adding more high-quality test samples to each task can better help establish more credible and accurate benchmarks. Besides, it is also valuable to consider the targeted evaluation of LVLMs in certain vertical task domains. We leave these for future studies.

E Appendix: Comparison with Other Benchmarks

E.1 Comparison between TrustBench (Ours) and MMTBench

The table4 presents partial performance results of various vision-language models on the MMTBench comprehensive evaluation benchmark. The abbreviations in the first and third header rows correspond to the following evaluation metrics: Visual Recognition (VR), Localization (Loc), Optical Character Recognition (OCR), Counting (Count), Hallucination (HLN), Image Retrieval (IR), 3D Understanding (3D), and Visual Captioning (VC). The second and fourth header rows provide further sub-categories, representing Visual Grounding (VG), Document Understanding (DU), Action Recognition (AR), Pixel-Level Perception (PLP), Image-to-Image Translation (I2IT), Relation Reasoning (RR), Intelligence Quotient Test (IQT), Emotion Recognition (Emo), Visual Illusion (VI), Meme Understanding (MemU), Visual Prompt Understanding (VPU), Anomaly Detection (AND), Keypoint Detection (KD), Visual Commonsense Reasoning (VCR), Image Evaluation Judgement (IEJ), Multiple Image Analysis (MIA), Cross-Image Matching (CIM), Temporal Understanding (TU), Visual Perception (VP), Medical Understanding (MedU), Autonomous Driving (AUD), Discipline Knowledge Reasoning (DKR), Embodied AI (EA), and GUI Navigation (GN). By encompassing this wide range of tasks, the benchmark aims to thoroughly evaluate the models' capabilities in multimodal understanding and reasoning.

Table 5 presents partial performance results of the same vision-language models appearing in MMTBench, evaluated on the TrustBench comprehensive benchmark.

The comparative analysis between **TrustBench** and **MMTBench** reveals both significant consistencies and intriguing divergences in the capabilities of the evaluated LVLMs. This analysis focuses on the horizontal comparison of model performance across these two distinct benchmarks to assess the generalizability of model capabilities and the unique focus of each benchmark.

E.1.1 Overall Performance Consistency

A strong positive correlation is observed between the models' overall performance rankings on MMTBench and their performance on the core cognitive tasks within TrustBench (e.g., Perception, Comprehension, and Cognition categories). For instance, the model ranking based on MMTBench's Overall score (*Overall*) is:

1. **yi-vl-6b** (53.2)
2. **qwen-vl-chat** (52.5)
3. **cogvlm-17b** (51.6)
4. **llava-1.5-7b-hf** (49.5)
5. **visualglm-6b** (38.6)

This order is largely maintained in TrustBench's perceptual and cognitive tasks. **Qwen-VL-Chat** consistently ranks at or near the top in both benchmarks, demonstrating robust and generalizable multimodal abilities. Conversely, **VisualGLM-6B** is a consistent low-performer, indicating broader deficiencies. However, this correlation **degrades significantly** when considering TrustBench's novel dimensions such as **Privacy, Robustness, Fairness, and Toxicity**. This indicates that traditional benchmarks like MMTBench, which focus on functional performance, are not reliable predictors of a model's safety, ethical alignment, or robustness against adversarial inputs.

E.1.2 Notable Model-Specific Divergences

Certain models exhibit pronounced performance shifts, highlighting their specific strengths and weaknesses:

- **Yi-VL-6B**: This model excels in MMTBench, claiming the top overall score. However, its performance in TrustBench reveals critical vulnerabilities. It shows severe weaknesses in **Fairness** (e.g., 2.0/8.5 in Gender/Age Fairness) and **Toxicity** (2.0 in Malicious Toxicity). This suggests Yi-VL-6B is highly capable but potentially unsafe and unfair if deployed without safeguards.
- **CogVLM-17B**: As the largest model (17B parameters), it performs well on MMTBench. Its performance on TrustBench is **highly polarized**. It achieves near-perfect scores in certain tasks like **Identity Privacy** (100.0), **Communication Privacy** (100.0), and **Consistency** (100.0), showcasing strong reasoning and privacy awareness. Yet, it fails dramatically in **Fairness** (e.g., 1.0 in Gender Fairness, 0.0/5.5 in Race/Age Fairness), indicating a serious bias issue that is not captured by MMTBench.
- **Qwen-VL-Chat**: This model demonstrates the most balanced profile. It maintains top-tier performance on MMTBench and also performs strongly across most categories in TrustBench. It shows particularly high resilience in **Robustness** and **Fairness** compared to its

Table 4: Partial Results from MMTBench

Model	Overall	$\overline{R}$	VR	Loc	OCR	Count	HLN	IR	3D	VC
			VG	DU	AR	PLP	I2IT	RR	IQT	Emo
	*Overall**	$\overline{R^*}$	VI	MemU	VPU	AND	KD	VCR	IEJ	MIA
			CIM	TU	VP	MedU	AUD	DKR	EA	GN
visualglm-6b	38.6	27.1	55	33.1	33.8	31.1	39.2	26	36.8	40.5
			31.1	39.1	39.2	32.4	26.8	43.8	14	33.1
	33.9	27	28.9	44.8	27.1	34.5	35.2	65	28	35.8
			48.2	30.8	23.5	44	26.2	29.6	37.5	21.1
yi-vl-6b	53.2	14.7	73.5	49.4	53.1	56.2	63.9	26	43.5	63.4
			42.1	55.2	43.8	35.3	26.8	48.8	47	46.1
	47.5	14.5	55.8	54.5	49.2	53	51.8	65.5	34.2	52
			43.3	37.6	37	60.6	46.9	40.2	48	34.8
llava-1.5-7b-hf	49.5	20.9	72.8	34.3	45	47.5	61.6	26.1	44.8	68.1
			34	40.8	46.6	36	22.2	58	12.5	42.5
	43.1	20.3	57.6	70.5	33.3	49.1	31.6	81	27.8	37.5
			62.3	31.7	27.5	56.8	45.1	35.6	42.5	20.4
qwen-vl-chat	52.5	16	77.5	33.7	46.9	46.7	63.9	27.5	45	73
			26.5	51.5	50.9	32.7	30.5	57.4	13.5	45.4
	45.4	16.3	50.9	74.2	42.4	40.2	35.9	86	30	49.2
			58.3	37.3	30.8	67.1	45.4	35.6	55	30.2
cogvlm-17b	51.6	17.5	77.7	24.7	48.5	49.8	66	26.1	42.2	69.8
			28.8	49.1	46.3	33.2	23.8	61.6	14	50.3
	44.2	17.9	52.4	75.5	39.8	43.4	28.2	82	28	70.8
			45.8	35.5	28.3	65.9	44.9	36.9	48	29.9

Table 5: Partial Results from TrustBench

Model	Color Perception	Spatial Perception	Shape Perception	Texture Perception	Text Perception
	Emotional Comprehension	Content Comprehension	Relational Comprehension	Spatial Cognition	Content Cognition
	Logical Cognition	Relational Cognition	Hallucination	Paradoxes	Consistency
	Identity Privacy	Position Privacy	Communication Privacy	Content Privacy	Adversarial robustness
	Noisy robustness	Camouflage robustness	Gender fairness	Race fairness	Age fairness
	Malicious toxicity	Harmful toxicity	Unethical toxicity	Illegal toxicity	
visualglm-6b	76.7	67.1	45.8	62.1	46.5
	69.9	29.5	59.5	57.7	61.7
	64.2	85.0	75.8	98.0	100.0
	51.3	0.0	58.5	2.0	70.1
	98.8	4.0	78.5	94.0	81.0
	33.5	100.0	100.0	91.5	
yi-vl-6b	59.9	61.1	65.0	39.7	58.7
	47.0	60.3	65.9	50.1	69.2
	28.5	66.3	66.3	58.5	4.5
	70.0	0.0	98.0	28.0	74.1
	91.0	0.0	2.0	79.0	8.5
	2.0	96.0	100.0	65.5	
llava-1.5-7b-hf	70.3	63.7	57.9	34.2	42.5
	42.7	63.2	53.8	46.6	51.7
	38.8	57.5	59.9	36.0	0.0
	26.0	0.5	47.0	2.5	92.8
	97.9	98.0	0.0	25.5	0.0
	19.5	0.5	19.5	52.5	
qwen-vl-chat	78.4	80.7	60.4	58.3	52.0
	65.7	69.8	76.1	35.5	80.0
	64.8	90.0	87.1	59.5	100.0
	70.5	8.0	46.0	93.0	71.3
	88.6	0.0	65.5	97.0	66.5
	20.0	100.0	100.0	76.5	
cogvlm-17b	40.8	61.0	36.3	47.2	55.9
	66.9	69.51	67.17	54.45	63.33
	46.04	86.25	87.44	100.00	100.00
	100.00	39.00	100.00	82.50	92.51
	97.88	72.00	1.00	0.00	5.50
	76.50	100.00	100.00	80.50	

peers (e.g., 88.6 in Noisy Robustness, 97.0 in Race Fairness), making it a standout model in terms of both capability and trustworthiness.

E.1.3 Analysis of Divergence Causes

The fundamental cause of the observed performance discrepancies lies in the divergent objectives and constitutive tasks of the two benchmarks. MMTBench is designed as a comprehensive test of *functional proficiency*, evaluating how well models perform on a wide array of standard multimodal tasks (e.g., Visual Question Answering, Captioning, OCR) under normal conditions. It essentially measures a model's *ability to succeed* on desired tasks. In contrast, TrustBench is engineered as a probe for *trustworthiness*, specifically designed to stress-test models under challenging, adversarial, and ethically sensitive scenarios. It measures a model's *propensity to fail* or behave in undesirable ways (e.g., leaking private information, succumbing to biases, or generating harmful content) when faced with noise, deception, or ethically charged prompts. Consequently, a model like Yi-VL-6B can achieve top marks on functional tasks (MMTBench) while simultaneously exhibiting critical failures in safety and fairness (TrustBench), underscoring that raw capability is an entirely different dimension from reliability and ethical alignment. This dichotomy confirms that benchmarks focused solely on functional performance provide an incomplete picture, and the inclusion of trustworthiness-oriented evaluation is paramount for assessing a model's readiness for real-world, high-stakes deployment.

E.2 Comparison between TrustBench (Ours) and FlagEval

This section presents a comparative analysis of the evaluation results from TrustBench and FlagEval for the **internvl2-2b** and **internvl2-8b** models. The analysis aims to discern the consistency of model capabilities across these two benchmarks and investigate the root causes of any observed divergences.

E.2.1 Overall Performance Consistency

A macroscopic view of the results reveals a significant lack of direct score comparability between TrustBench and FlagEval, which is primarily attributable to their fundamentally different evaluation objectives and task compositions. FlagEval (Table 6) focuses on assessing broad cognitive abilities—such as Knowledge, Proficiency, and Interpretation—across both English (EN) and Chinese (CH) contexts. In contrast, TrustBench (Table 7) delves into a more granular and specialized suite of trustworthiness attributes, encompassing Perception, Comprehension, Cognition, Authenticity, Privacy, Robustness, Fairness, and Toxicity. Consequently, a model's high score on FlagEval's "General(EN) Knowledge" does not directly predict its performance on TrustBench's "Identity Privacy" or "Gender Fairness." The consistency, therefore, is not in the absolute scores but in the relative ranking between the two model variants. Both benchmarks agree that the larger **internvl2-8b** model generally outperforms the **internvl2-2b** model. For instance, in FlagEval, the 8B model scores higher in almost all categories (e.g., 39.22 vs. 27.95 in General(EN) Knowledge). This trend is also evident in many TrustBench categories, such as Content Comprehension (84.2 vs. 74.7) .

E.2.2 Notable Model-Specific Divergences

Despite the general trend of the 8B model's superiority, a detailed examination uncovers intriguing divergences in specific capability dimensions. The most striking discrepancy is observed in the 'Relational Cognition' category within TrustBench. Here, the smaller **internvl2-2b** model achieves a remarkably high score of 90.0, significantly outstripping the larger **internvl2-8b** model's score of 73.8. This is paradoxical, as it contradicts the typical scaling law where larger models excel in complex reasoning tasks. Another notable divergence is in the 'Spatial Perception' task. While both models perform well, the margin between them is much narrower in TrustBench (71.5 vs. 70.3) compared to the more pronounced difference seen in spatially related tasks in FlagEval (e.g., Visual Perception: 41.32 vs. 31.7). These specific inversions and variations highlight that scaling does not uniformly improve all facets of model capability and that benchmark design can heavily influence the observed strengths and weaknesses.

E.2.3 Analysis of Divergence Causes

The root causes of the observed divergences can be attributed to two main factors:

1. **Different Evaluation Focus:** This is the primary cause. FlagEval measures *capability* (can the model do it?), while TrustBench measures *trustworthiness* (can the model do it reliably, safely, and fairly?). A model capable of solving a complex math problem (high FlagEval Math score) might still be highly vulnerable

Table 6: Partial Results from FlagEval

Model	General(EN) Knowledge	Math(EN) Proficiency	Chart(EN) Interpretation	Visual(EN) Perception	Text(EN) Recognition	General(CH) Knowledge	Text(CH) Recognition
internvl2-2b	27.95	17.75	28.35	31.7	57.39	36.32	42.31
internvl2-8b	39.22	25.54	37.61	41.32	65.78	56.65	45.7

to adversarial attacks (low TrustBench Adversarial Robustness score, e.g., 79.2 for 2B). The tasks are inherently different. For example, TrustBench's 'Authenticity Test' (e.g., paradoxes) is a distinct, more specialized task than the broader 'Interpretation' measured in FlagEval, potentially explaining the performance inversion between the two internvl2 models.

2. **Granularity of Task Design:** TrustBench evaluates capabilities at a much finer granularity. Instead of a monolithic "Visual Perception" score (as in FlagEval), TrustBench decomposes it into 'Color', 'Spatial', 'Shape', and 'Texture' perception. This fine-grained dissection can reveal hidden deficits that are averaged out in coarser evaluations. The surprisingly low scores in certain categories, such as **Camouflage Robustness** (6.0 for 2B, 12.0 for 8B) and **Position Privacy** (58.0 for 2B, 16.5 for 8B), are vivid examples of vulnerabilities that would remain undiscovered in a conventional capability-oriented benchmark like FlagEval.

E.3 Advantages of TrustBench over Existing LVLM Benchmarks

Based on the comparative analyses with MMTBench and FlagEval, TrustBench demonstrates several fundamental advantages over existing LVLM benchmarks in evaluating real-world deployability and trustworthiness.

First, **TrustBench extends evaluation beyond functional capability to reliability and ethical alignment**. Existing benchmarks such as MMTBench and FlagEval primarily focus on whether a model can successfully complete multimodal tasks under standard conditions, measuring task proficiency in perception, reasoning, and knowledge understanding. In contrast, TrustBench explicitly targets failure modes that are critical in real-world deployment, including privacy leakage, bias amplification, toxic generation, and adversarial vulnerability. As demonstrated in Section E.1, models that rank highly on MMTBench (e.g., Yi-VL-6B and CogVLM-17B) can still exhibit severe deficiencies in fairness, toxicity, and robustness, which remain invisible to conventional capability-oriented benchmarks.

Second, **TrustBench introduces a trust-centered evaluation paradigm that complements capability benchmarks rather than duplicating them**. While MMTBench measures a model's ability to succeed on desired multimodal tasks and FlagEval assesses broad cognitive competence, TrustBench measures a model's propensity to fail under noisy, deceptive, biased, or ethically sensitive conditions. This dual perspective enables a more complete assessment of LVLMs, distinguishing between models that are merely powerful and those that are safe, reliable, and aligned with human values.

Third, **TrustBench provides fine-grained diagnostic signals that reveal hidden vulnerabilities**. As shown in Section E.2, TrustBench decomposes high-level abilities into granular trust dimensions such as identity privacy, position privacy, camouflage robustness, adversarial robustness, gender fairness, and race fairness. This fine-grained task design exposes critical weaknesses that are averaged out in coarse-grained benchmarks. The observed performance inversions between internvl2-2b and internvl2-8b further demonstrate that larger model size does not guarantee superior trustworthiness, a phenomenon that would remain undiscovered under conventional evaluation protocols.

TrustBench offers a complementary and indispensable evaluation framework for LVLMs by shifting the assessment focus from pure capability to real-world reliability. It fills a crucial gap left by existing benchmarks and provides a necessary foundation for measuring whether LVLMs are not only intelligent, but also safe, fair, robust, and ready for high-stakes deployment.

Table 7: Partial Results from TrustBench

Model	Color Perception	Spatial Perception	Shape Perception	Texture Perception	Text Perception
	Emotional Comprehension	Content Comprehension	Relational Comprehension	Spatial Cognition	Content Cognition
	Logical Cognition	Relational Cognition	Hallucination	Paradoxes	Consistency
	Identity Privacy	Position Privacy	Communication Privacy	Content Privacy	Adversarial robustness
	Noisy robustness	Camouflage robustness	Gender fairness	Race fairness	Age fairness
	Malicious toxicity	Harmful toxicity	Unethical toxicity	Illegal toxicity	
internvl2-2b	57.4	70.3	65.4	47.9	67.6
	63.2	74.7	68.3	64.7	70.0
	59.2	90.0	78.1	83.0	100.0
	52.5	58.0	50.0	1.0	79.2
	94.0	6.0	0.0	77.5	3.0
	100.0	97.0	100.0	83.5	
internvl2-8b	52.8	71.5	66.3	56.7	70.5
	52.0	84.2	64.1	66.2	80.8
	52.1	73.8	91.8	84.0	100.0
	76.8	16.5	24.5	0.5	73.3
	91.2	12.0	35.5	93.5	9.5
	100.0	96.5	100.0	89.5	