Evaluations and Benchmarks in Context of Multimodal LLM

https://mllm2024.github.io/CVPR25/

CVPR Nashville JUNE 11-15, 2025

NUS National University of Singapore

Carnegie Mellon University

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

香港科技大學 THE HONG KONG CYBER UNIVERSITY OF SCIENCE AND TECHNOLOGY

Tencent

UTD THE UNIVERSITY OF TEXAS AT DALLAS

Hao Fei

National University of Singapore

Xiang Yue

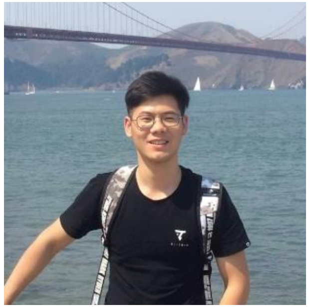
Carnegie Mellon University

Kaipeng Zhang

Shanghai AI Lab

Long Chen

HKUST



Jian Li

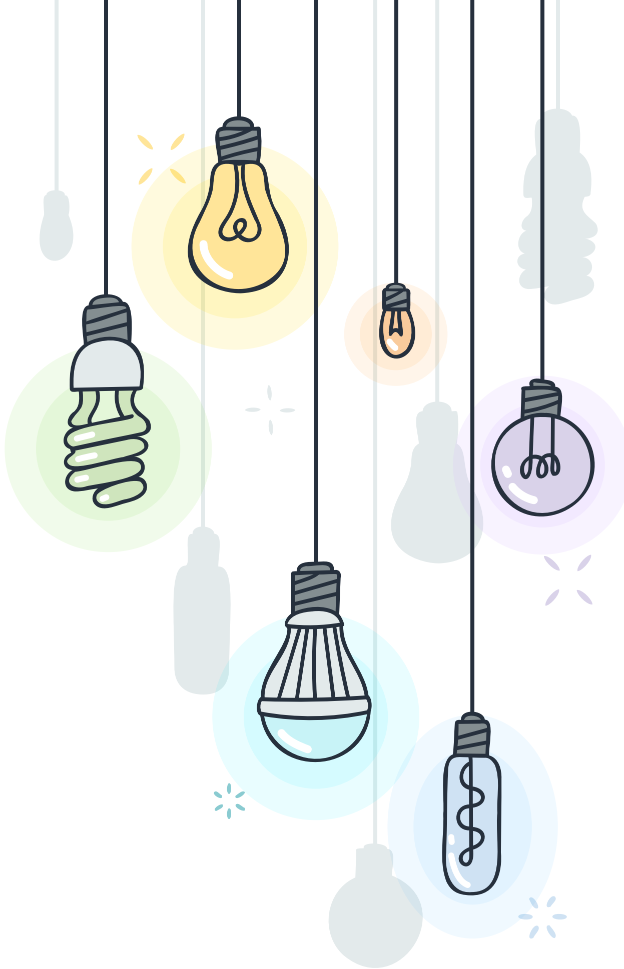
Tencent YoutuLab

Xinya Du

University of Texas at Dallas

Part-VI

Beyond Evaluation: Path to Multimodal Generalist



Hao Fei

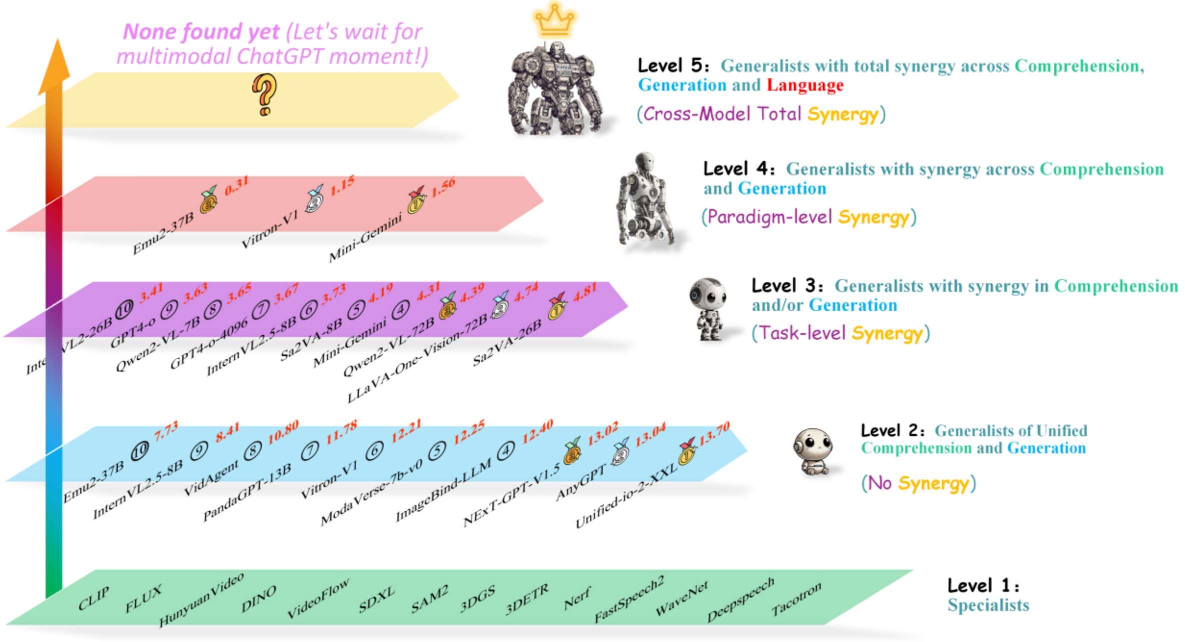
Senior Research Fellow

National University of Singapore

http://haofei.vip/

On Path To Multimodal Generalist: General-Bench & General-Level

Is your MLLM a well-rounded generalist?



Project

Leaderboard

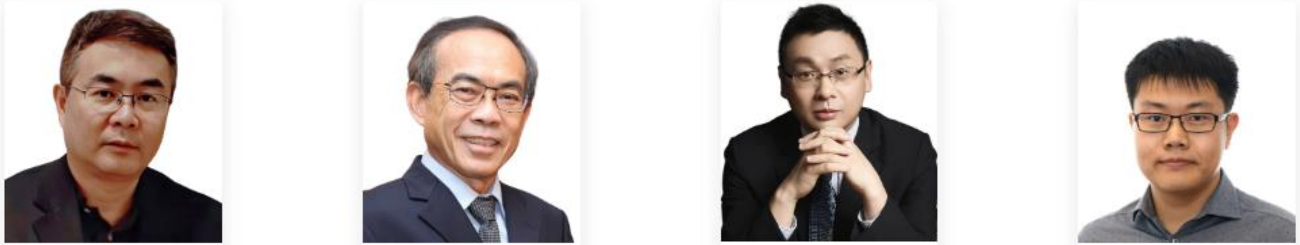
Benchmark

Project: https://generalist.top/

Paper: https://arxiv.org/abs/2505.04620

Benchmark: https://generalist.top/leaderboard

- **Hao Fei**, Yuan Zhou, ···, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, Hanwang Zhang. "On Path to Multimodal Generalist: General-Level and General-Bench". **ICML (Spotlight)**. 2025



Jiebo Luo	Tat-Seng Chua	Shuicheng Yan	Hanwang Zhang
University of Rochester	*NUS*	*NUS*	*NTU*
Advisory	Advisory	Project Supervision	Project Supervision

3D Group	Image Group	Image Group	Image Group	Image Group	Image Group	Video Group

Tao Zhang	Tianjie Ju	Zixiang Meng	Shilin Xu	Liyu Jia	Wentao Hu	Meng Luo
WHU	*SJTU*	*WHU*	*PKU*	*NTU*	*NTU*	*NUS*
Video Group	NLP Group	Video&Image Group	Video Group	Image Group	Image Group	Video Group

- Hao Fei, … ral-Bench” .
  ICML（

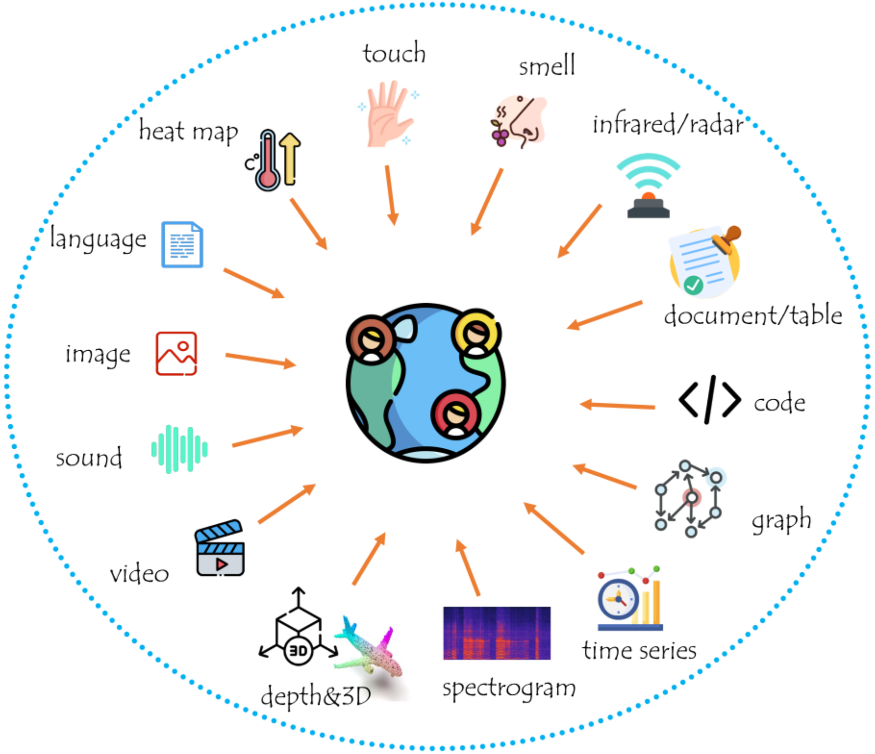
Table of Content

Table of Content

Path to Multimodal Generalist: General-Level

Multimodal AI

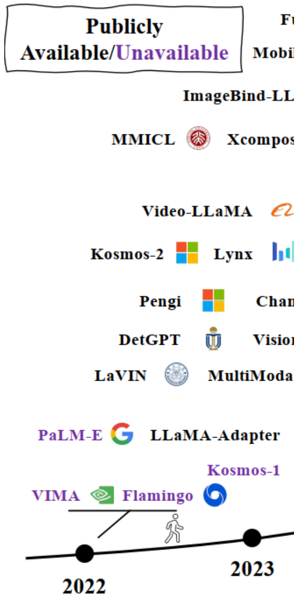
☞ The world functions with varied multimodal information and signals

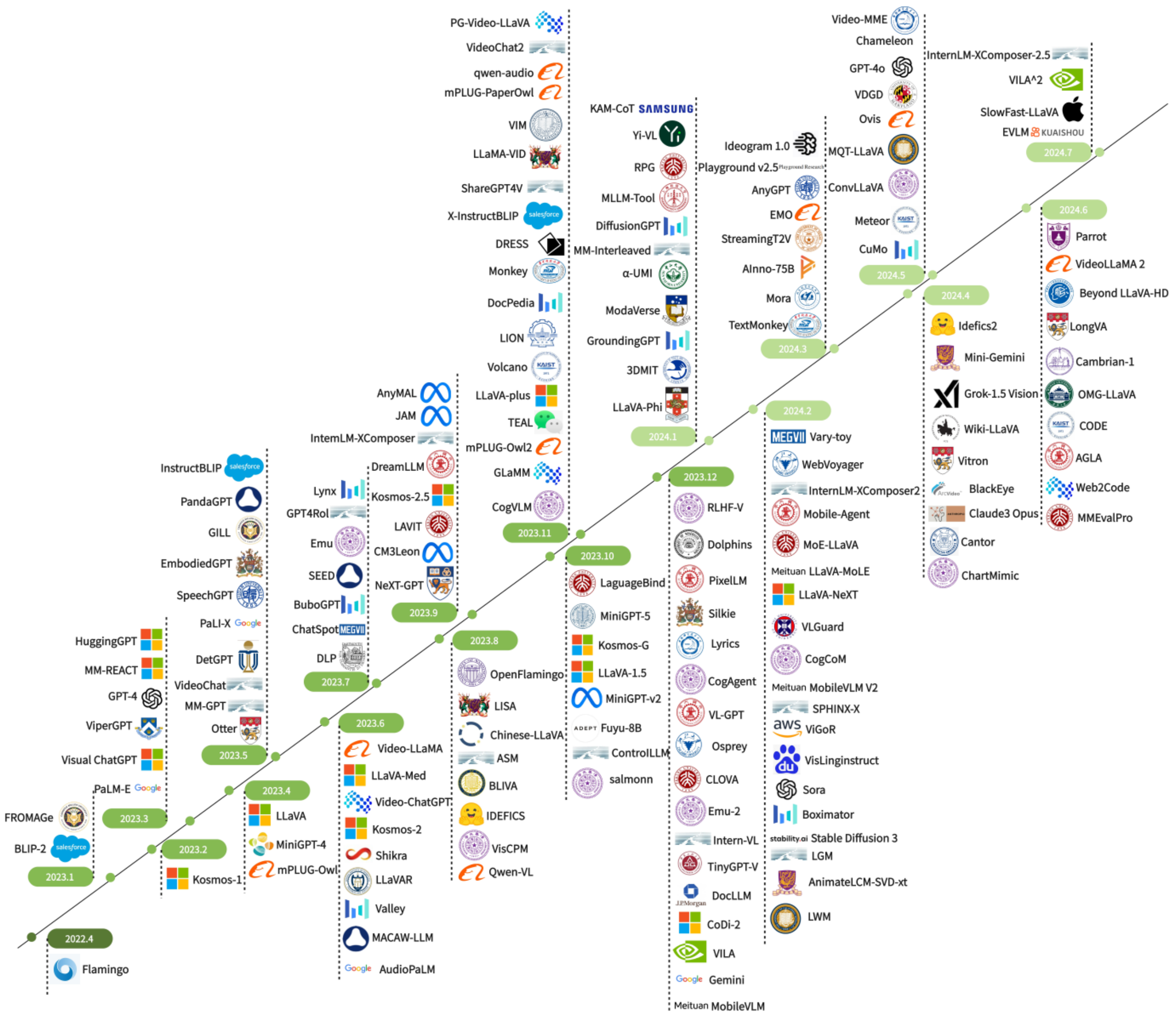


Path to Multimoc

Development of Mul

☞ Hot Trends

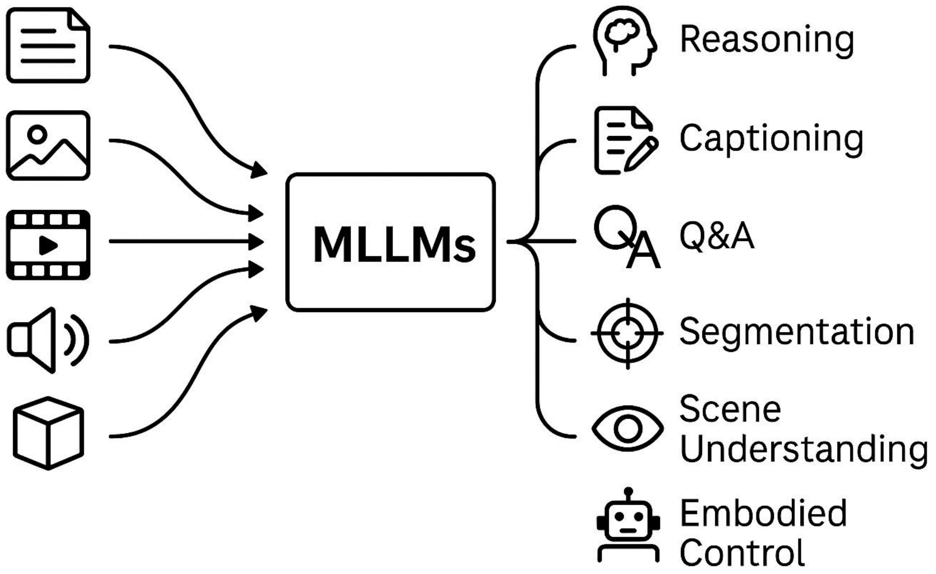




Path to Multimodal Generalist: General-Level

Background

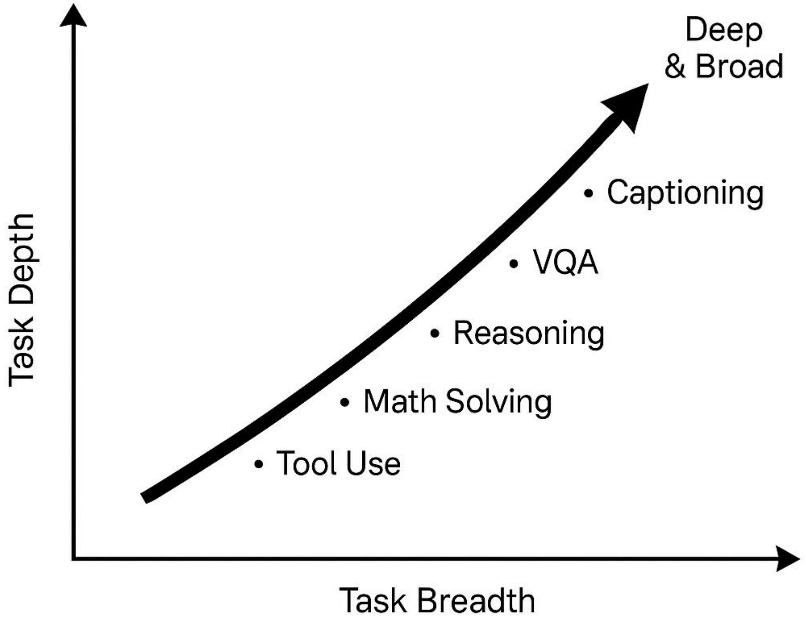
- Expansion of MLLMs: More modalities, More Tasks



Path to Multimodal Generalist: General-Level

Background

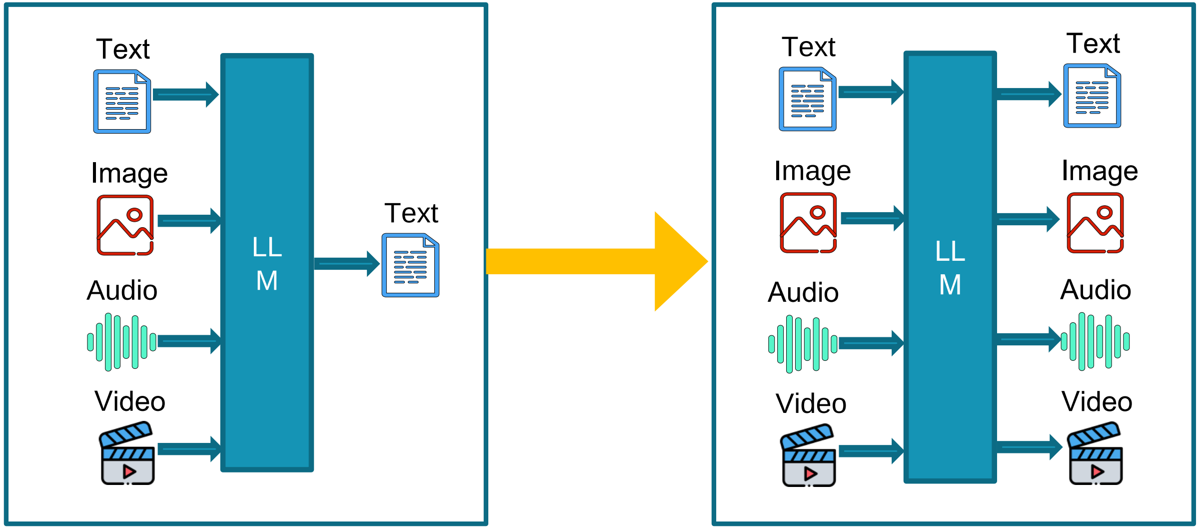
- Evolving with deeper capability



Path to Multimodal Generalist: General-Level

Background

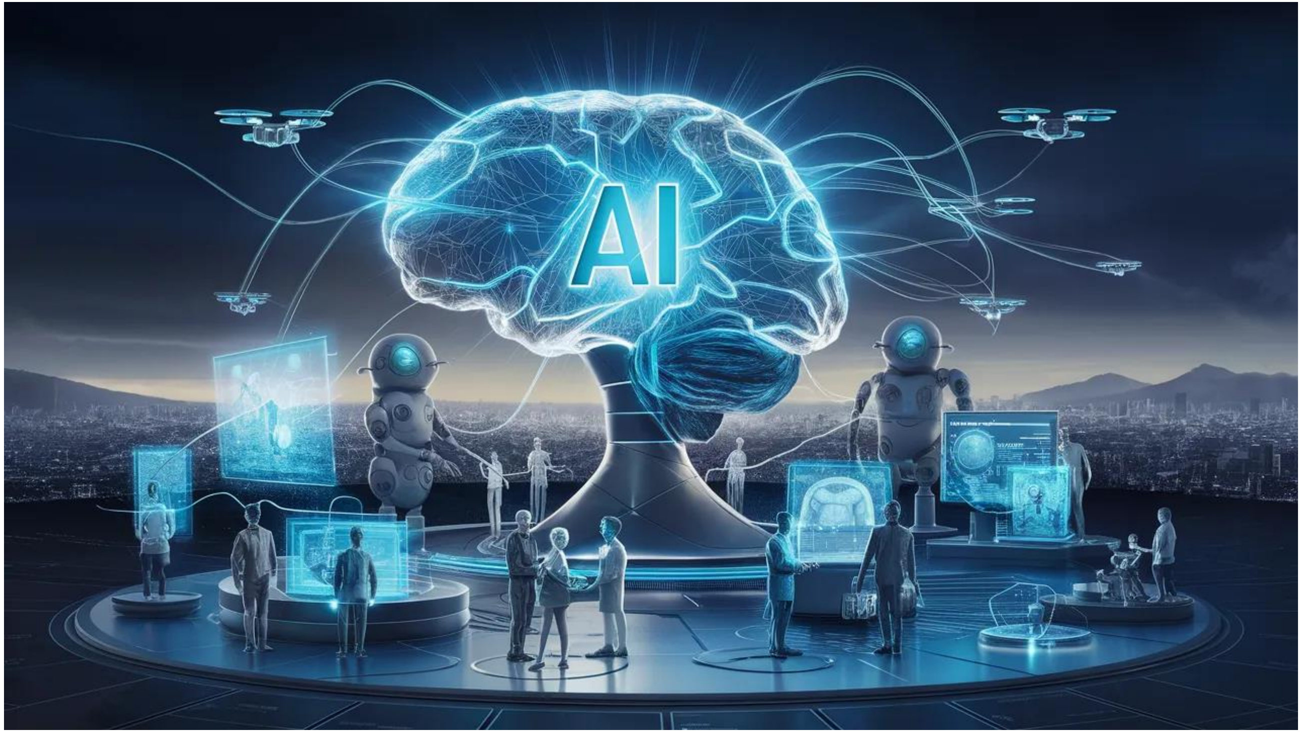
- Multimodal Comprehension vs. Unified Multimodal Comprehension & Generation



Path to Multimodal Generalist: General-Level

Ultimate Goal

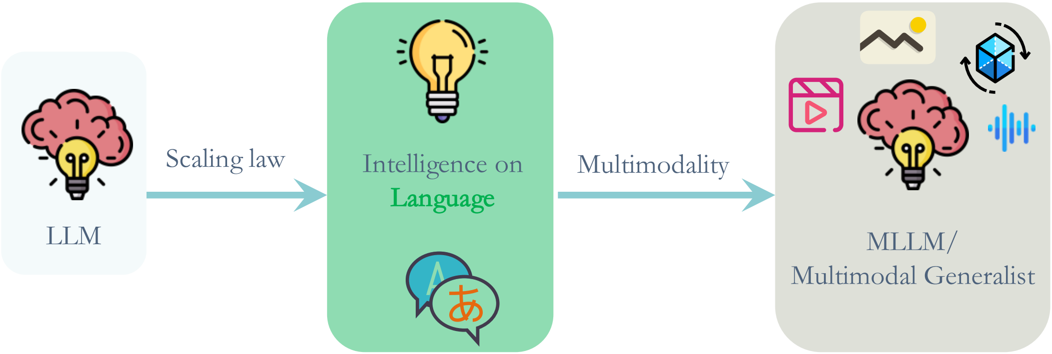
☞ What will the next-generation of **multimodal foundation models/agents** look like?



Path to Multimodal Generalist: General-Level

Motivation

- Existing issue-I: *The language intelligence of LLMs empowers multimodal intelligence.*

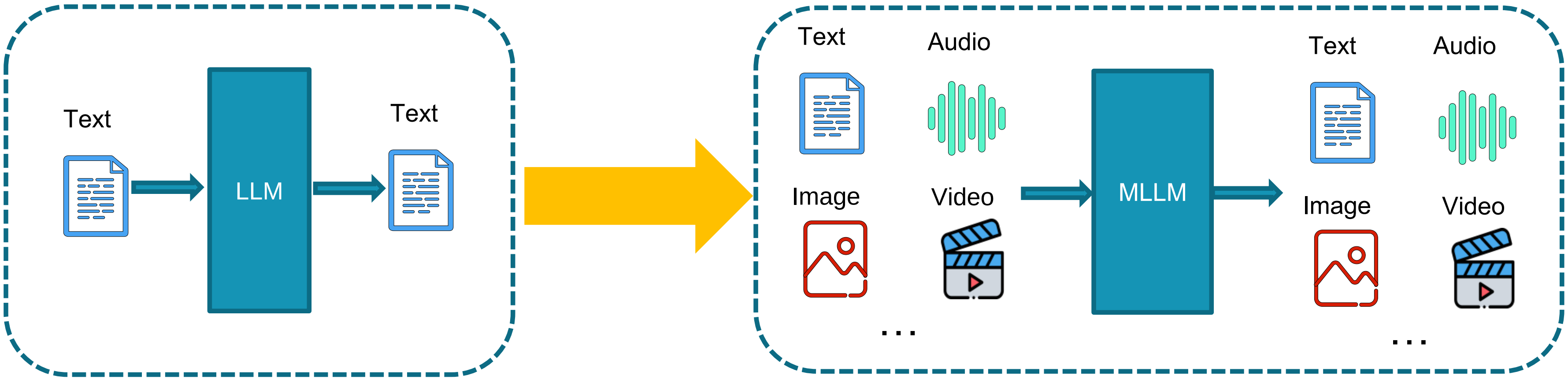


Path to Multimodal Generalist: General-Level

On Path to Multimodal Generalist: General-Level and General-Bench

- Existing intelligent pattern in multimodal generalist

*Extending **Language** LLM to **Multimodal** LLM (MLLM)*

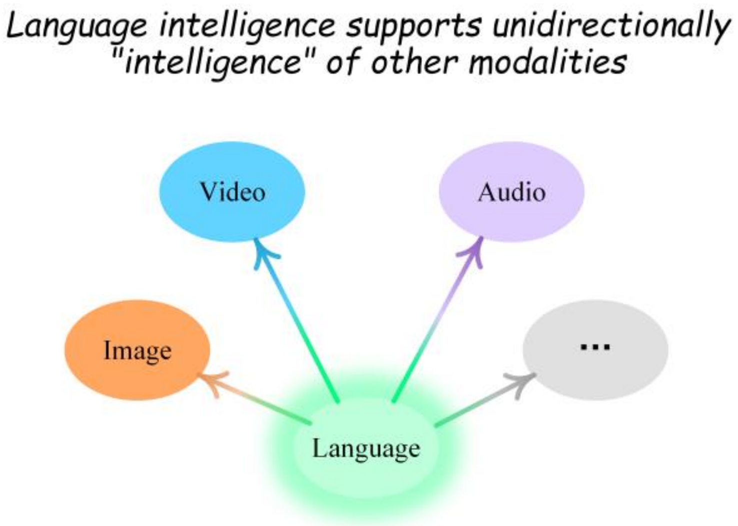


Path to Multimodal Generalist: General-Level

Motivation

- Existing issue-I: *The language intelligence of LLMs empowers multimodal intelligence.*

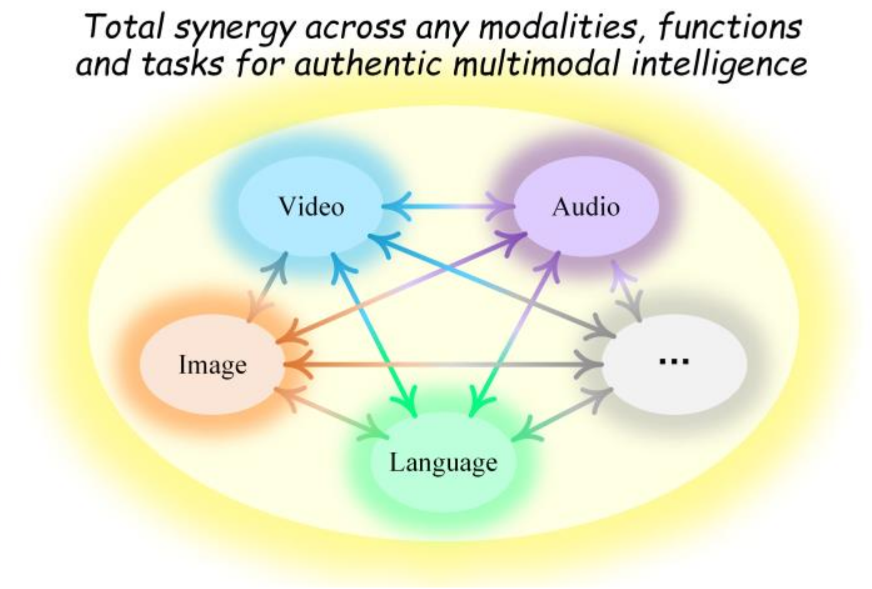
__Existing__ intelligent pattern in multimodal generalist



Motivation

- Existing issue-I: *The language intelligence of LLMs empowers multimodal intelligence.*

__Ideal__ intelligent pattern in multimodal generalist

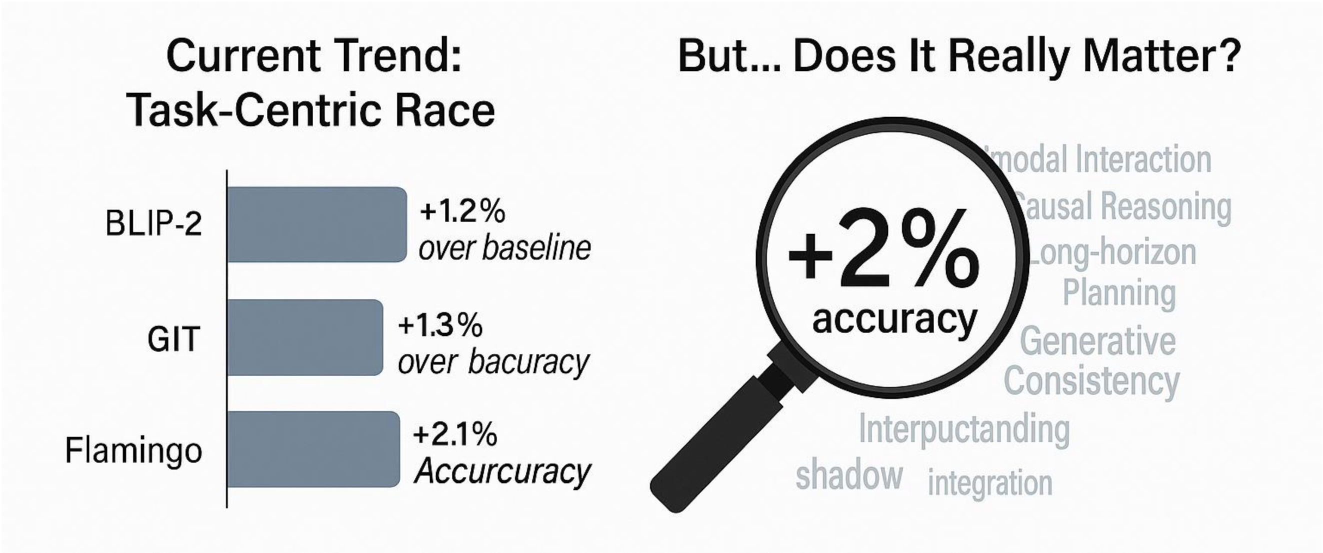


Path to Multimodal Generalist: General-Level

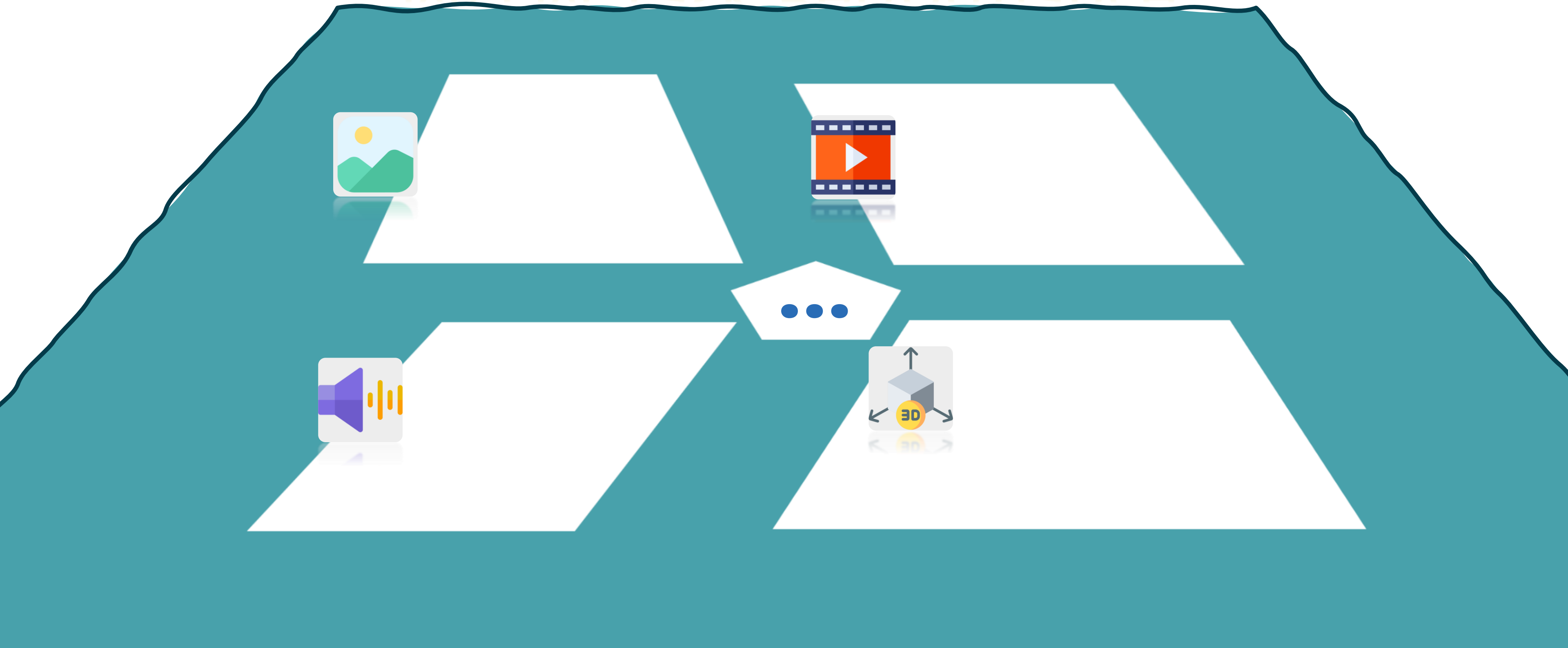
Motivation

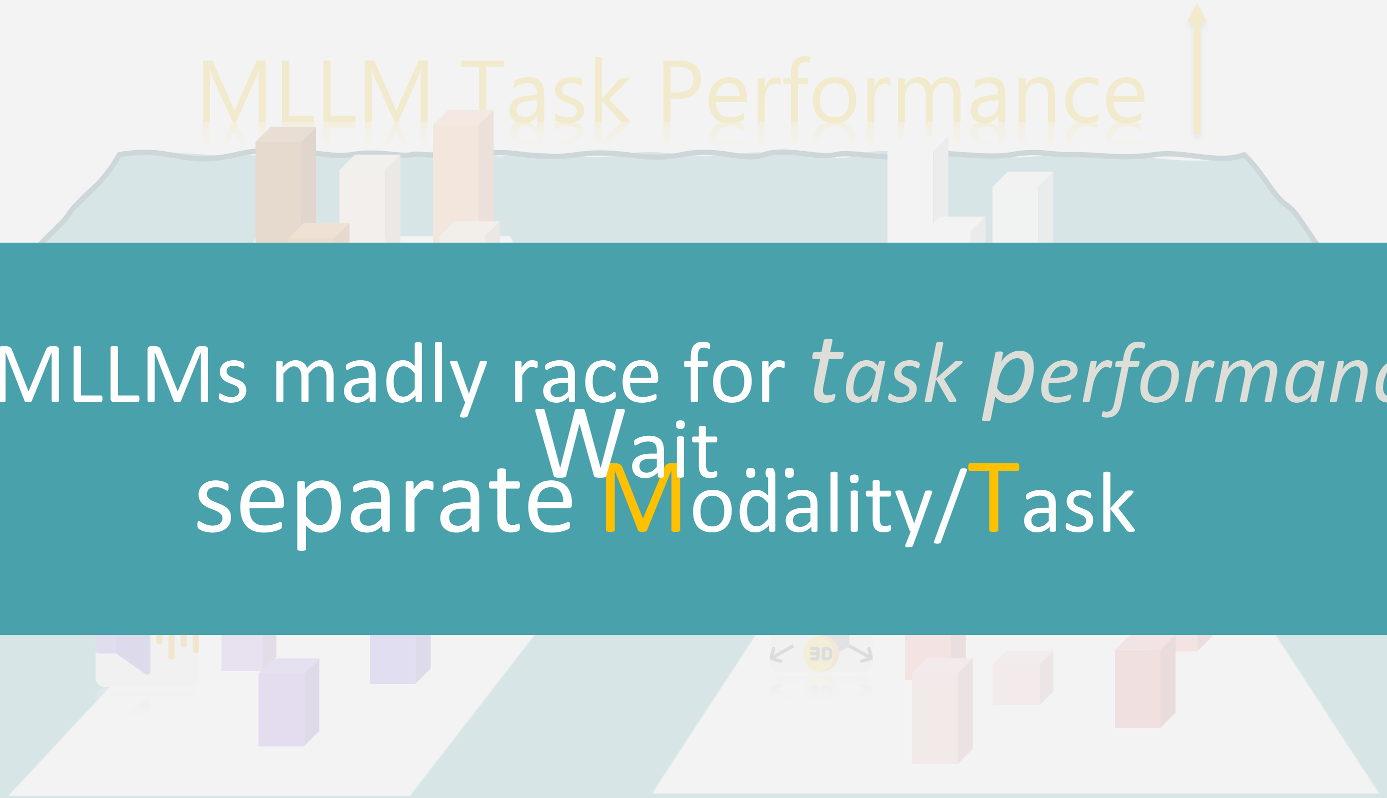
➢ Existing issue-II: *Rethinking MLLM evaluation beyond straightforward accuracy gains.*

Most existing MLLMs madly race for task performance of single modality/task.



MLLM Task Performance





Most MLLMs madly race for *task performance* of separate **M**odality/**T**ask

Wait ...

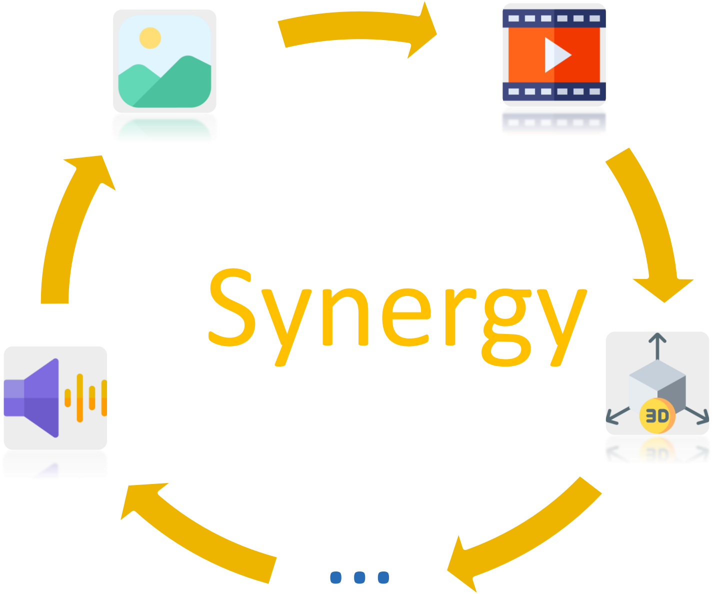
Does *higher results* simply mean *stronger intelligent multimodal AI* ?

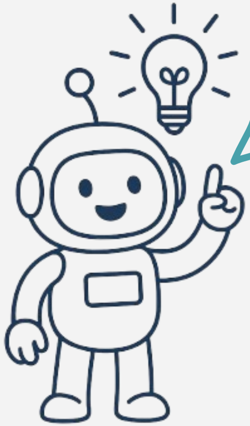
Synergy Drives Intelligence:

The Path Toward AGI

Synergy

The ability to generalize / transfer knowledge across Tasks, Modalities and Paradigm...


Synergy
3D



General-Level

Positioning and assessing the capabilities of current MLLM generalists

Level-5
Full Automation

Level-4
High Automation

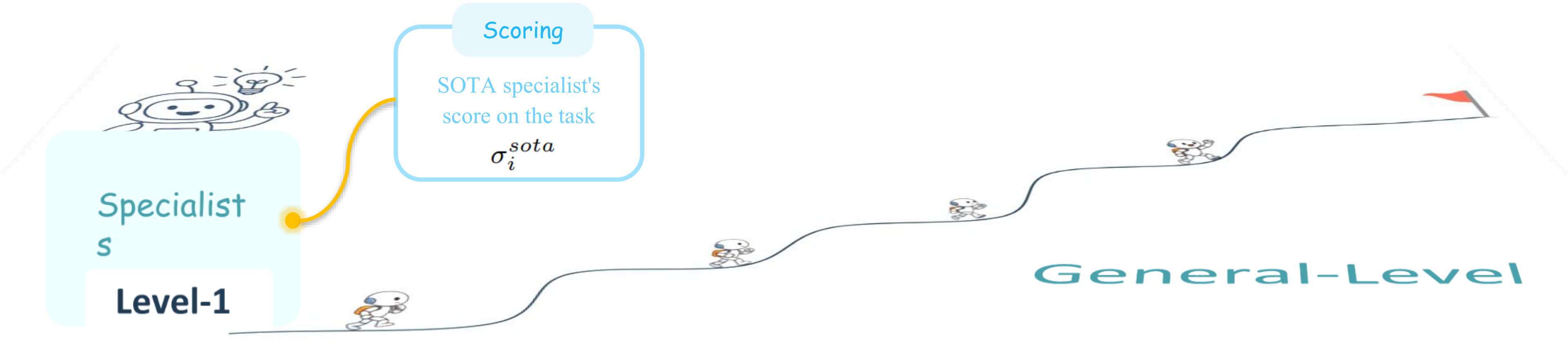
Level-3
Conditional Automation

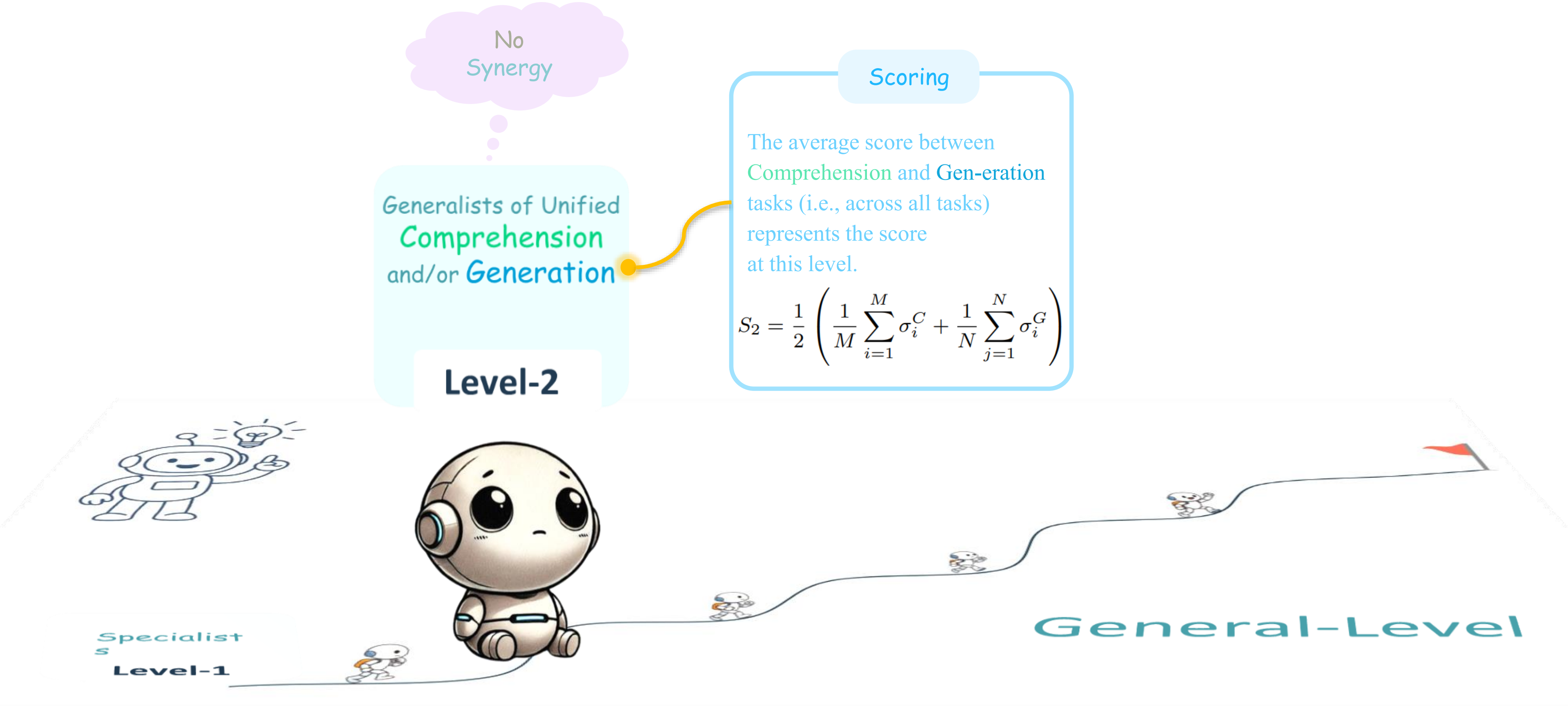
Level-2
Partial Automation

Level-1
Driver Assistance

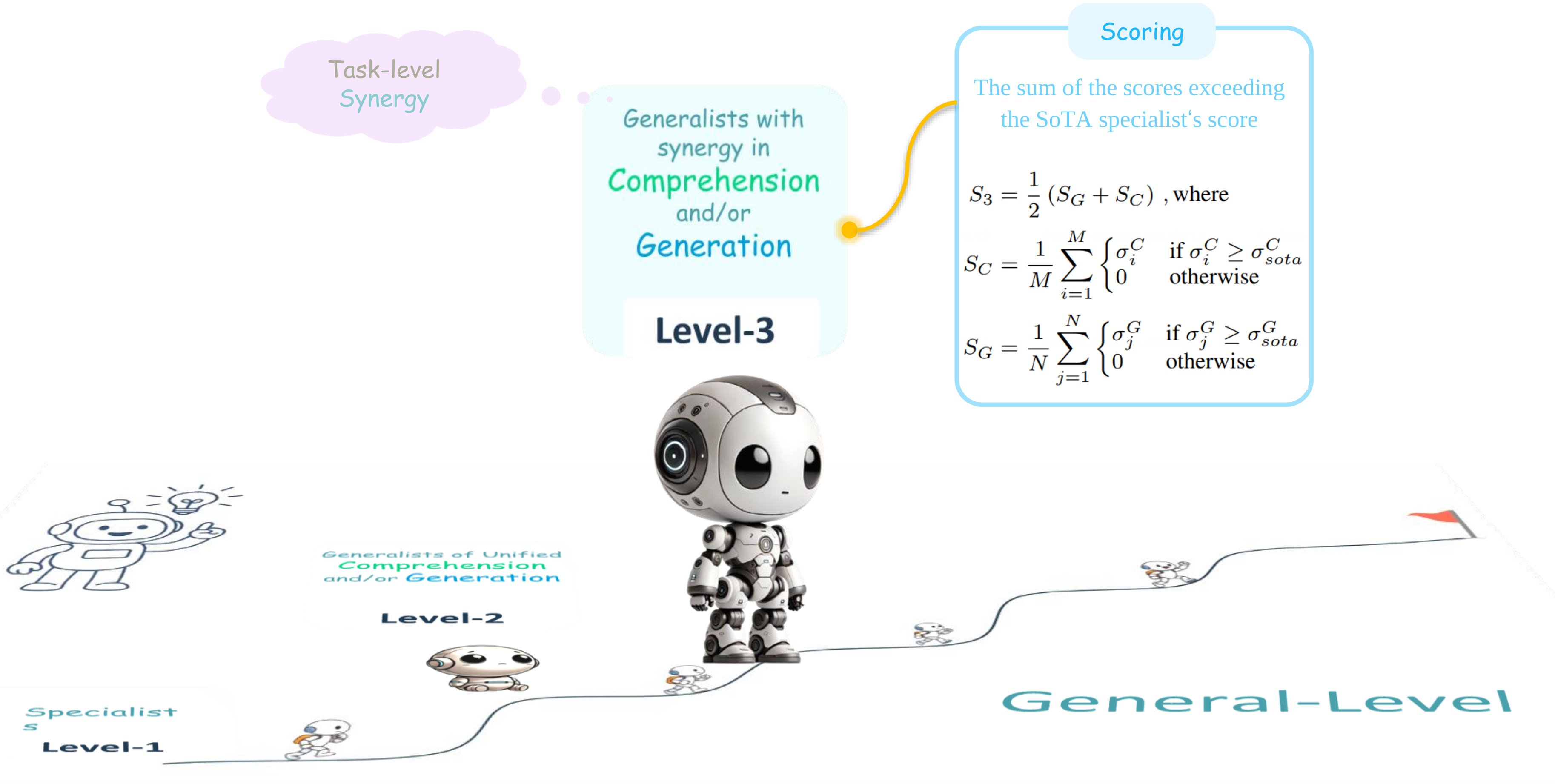
Level-0
No automation

Levels of Autonomous Driving

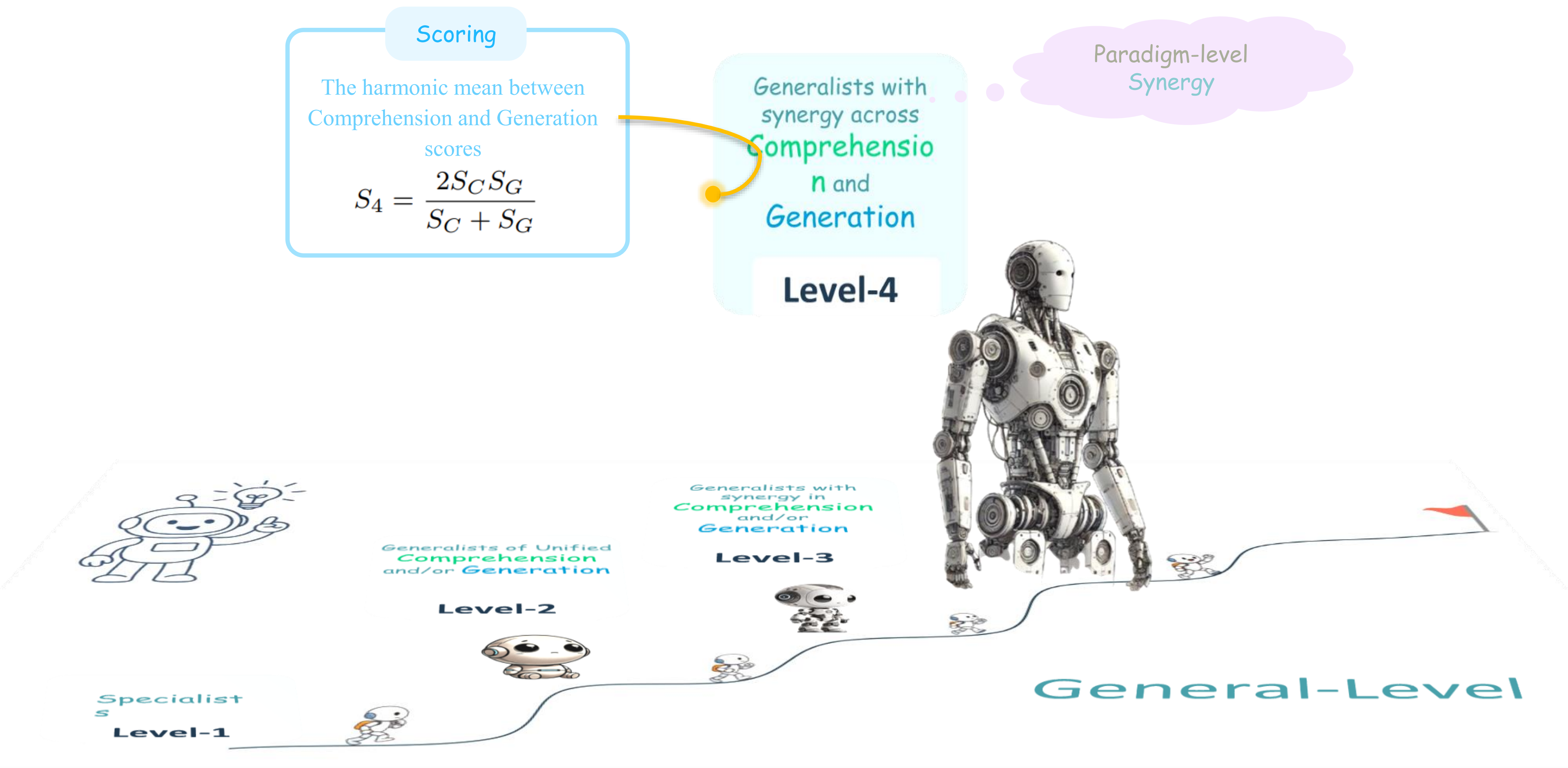

Scoring
SOTA specialist's score on the task
σ_i^{sota}
Specialists
Level-1
General-Level



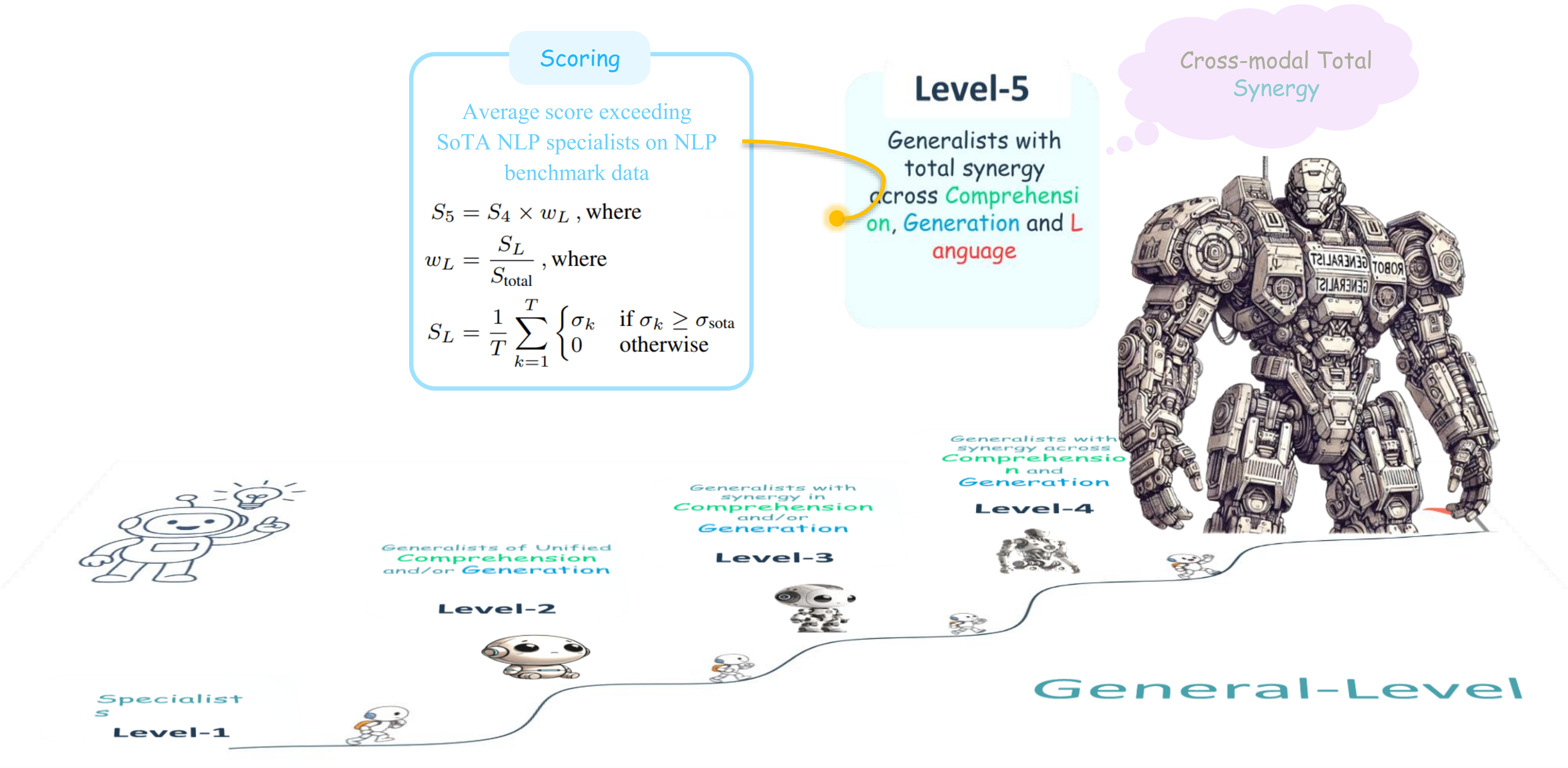
No Synergy
Scoring
The average score between Comprehension and Gen-eration tasks (i.e., across all tasks) represents the score at this level.
$$S_2 = \frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M}\sigma_i^C + \frac{1}{N}\sum_{j=1}^{N}\sigma_i^G\right)$$
Generalists of Unified Comprehension and/or Generation
Level-2
Specialists
Level-1
General-Level



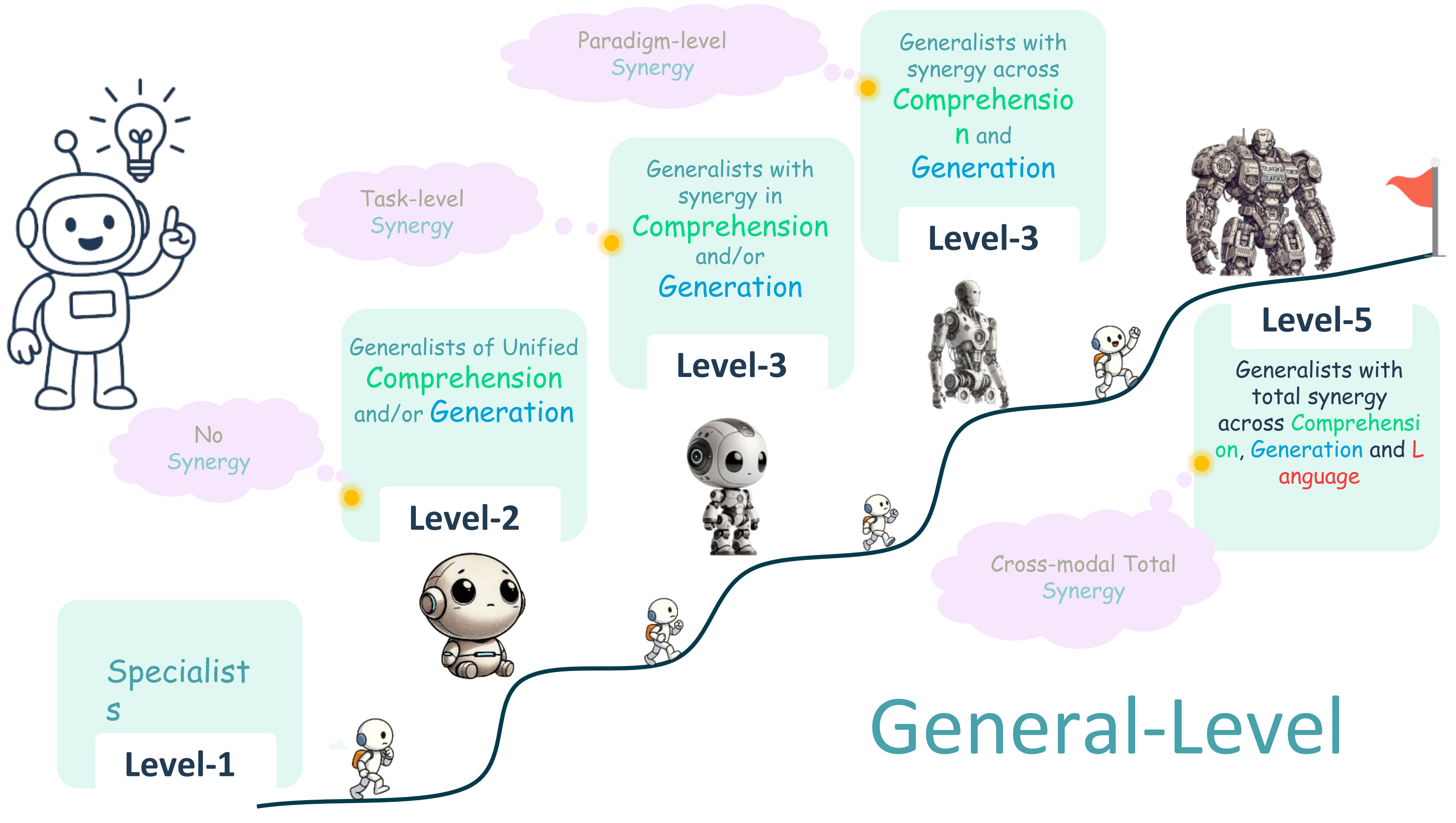
Task-level Synergy
Generalists with synergy in Comprehension and/or Generation
Level-3
Scoring
The sum of the scores exceeding the SoTA specialist's score
$S_3 = \frac{1}{2}(S_G + S_C)$, where
$S_C = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i^C & \text{if } \sigma_i^C \geq \sigma_{sota}^C \\ 0 & \text{otherwise}\end{cases}$
$S_G = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j^G & \text{if } \sigma_j^G \geq \sigma_{sota}^G \\ 0 & \text{otherwise}\end{cases}$
Generalists of Unified Comprehension and/or Generation
Level-2
Specialists
Level-1
General-Level



Scoring
The harmonic mean between Comprehension and Generation scores
$$S_4 = \frac{2S_C S_G}{S_C + S_G}$$
Generalists with synergy across Comprehension and Generation
Level-4
Paradigm-level Synergy
Generalists with synergy in Comprehension and/or Generation
Level-3
Generalists of Unified Comprehension and/or Generation
Level-2
Specialists
Level-1
General-Level



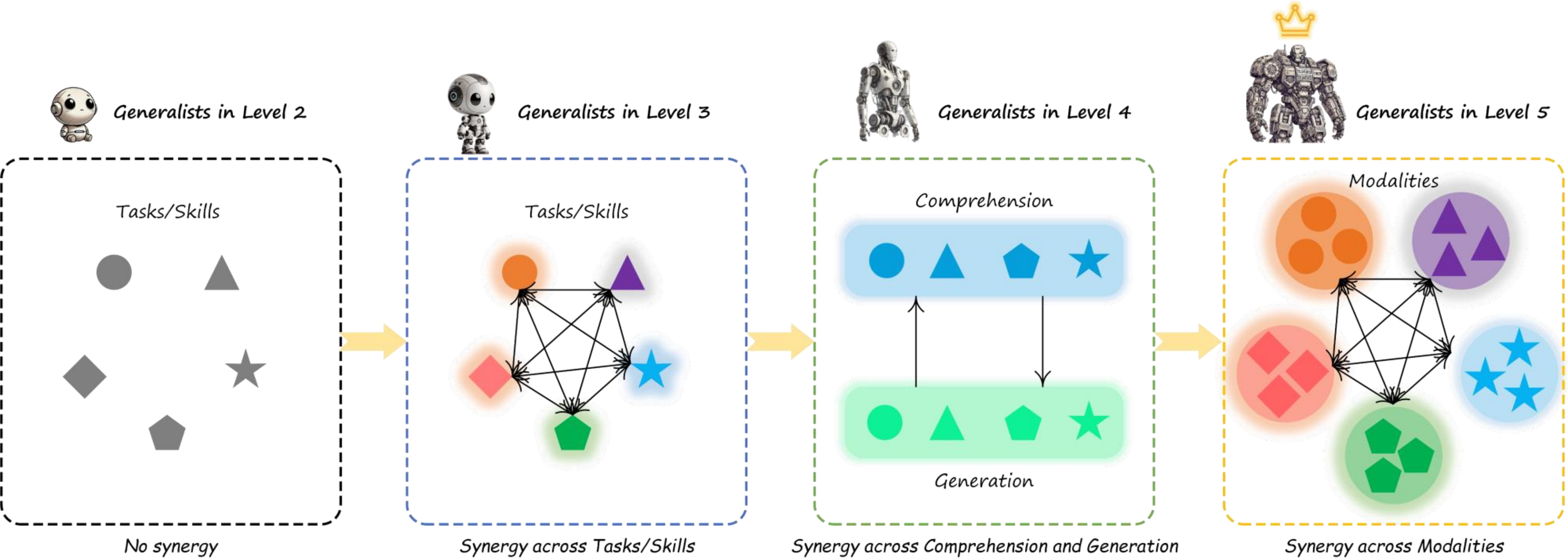
Scoring
Average score exceeding SoTA NLP specialists on NLP benchmark data
$S_5 = S_4 \times w_L$, where
$w_L = \frac{S_L}{S_{\text{total}}}$, where
$S_L = \frac{1}{T}\sum_{k=1}^{T}\begin{cases}\sigma_k & \text{if } \sigma_k \geq \sigma_{\text{sota}} \\ 0 & \text{otherwise}\end{cases}$
Level-5
Generalists with total synergy across Comprehension, Generation and Language
Cross-modal Total Synergy
Generalists with synergy across Comprehension and Generation
Level-4
Generalists with synergy in Comprehension and/or Generation
Level-3
Generalists of Unified Comprehension and/or Generation
Level-2
Specialists
Level-1
General-Level



Paradigm-level Synergy
Generalists with synergy across Comprehension and Generation
Level-3
Task-level Synergy
Generalists with synergy in Comprehension and/or Generation
Level-3
Generalists of Unified Comprehension and/or Generation
Level-2
No Synergy
Level-5
Generalists with total synergy across Comprehension, Generation and Language
Cross-modal Total Synergy
Specialists
Level-1
General-Level

Path to Multimodal Generalist: General-Level

General-Level: Synergy-centered evaluation framework



Path to Multimodal Generalist: General-Level

What is a (Multimodal) Generalist?

- One single model is capable of handling multiple tasks
- In most cases, an LLM serves as the core intelligence component
- At the very least, can be prompted using natural language to express user intentions
- e.g., MLLMs, or large multimodal foundational models, as well as multimodal agents

GPT-4o | LLaVA | Gemini

Blip | NExT-GPT | …

Path to Multimodal Generalist: General-Level

What is a (Multimodal) Specialist?

- In most cases, a specialist model can — and only can — achieve SoTA performance on a specific task
- It is typically fine-tuned on the training set of that task
- In most cases, the model often has a smaller parameter size compared to generalist models
- It mostly does not incorporate an LLM as the core reasoning or intelligence engine

DINO | SDXL | SAM

WaveNet | Nerf | ...

Path to Multimodal Generalist: General-Level

Relaxation of Scoring

- How to measure the **synergy effect** between on task-A & on task-B ?

the performance of a generalist on joint modeling of tasks A and B $P_\theta(y|A, B)$ *should exceed its performance when modeling task A alone* $P_\theta(y|A)$ *or task B alone* $P_\theta(y|B)$.

$$P_\theta(y|A, B) > P_\theta(y|A) \quad \& \quad P_\theta(y|A, B) > P_\theta(y|B)$$

Path to Multimodal Generalist: General-Level

Relaxation of Scoring

- How to measure the **synergy effect** between on task-A & on task-B ?

$$P_\theta(y|A, B)$$

$$\cancel{P_\theta(y|A)}$$

$$\cancel{P_\theta(y|B)}$$

Path to Multimodal Generalist: General-Level

Relaxation of Scoring

- How to measure the **synergy effect** between on task-A & on task-B ?

the stronger a model's synergy capability, the more likely it is to surpass the task performance of SoTA specialists when there is a synergy.

Let's simplify the rule:

if a generalist outperforms a SoTA specialist in a specific task, we consider it as evidence of a synergy effect, i.e., leveraging the knowledge learned from other tasks or modalities to enhance its performance in the targeted task.

Path to Multimodal Generalist: General-Level

One more notice

- There’s never a fair comparisons for generalist with specialist

Specialist

Generalis t

Fine-tuned on training set

No task-specific fine-tuning

Hard!

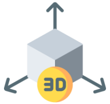
Unfair!

But Necessary!

Path to Multimodal Generalist: General-Level

Modality-specific Scoring

☞ calculate the specific score component S_k^i of a generalist in the i-th modality (assuming there are N modalities in total) for the score S_k .

	S_2^{img}	S_2^{vid}	S_2^{aud}	S_2^{3d}	$\dots$	S_2
	S_3^{img}	S_3^{vid}	S_3^{aud}	S_3^{3d}	$\dots$	S_3
	S_4^{img}	S_4^{vid}	S_4^{aud}	S_4^{3d}	$\dots$	S_4
	S_5^{img}	S_5^{vid}	S_5^{aud}	S_5^{3d}	$\dots$	S_5
						

$$S_k = \sum_{i}^{N} \frac{1}{N} S_k^i$$

Path to Multimodal Generalist: General-Level

Independence from Peer Generalists

The scores of any generalist:

√ • *depend solely on the data of the task and the reference score of SoTA specialist*

✖ • *without relying on the scores of other tested generalists*

Path to Multimodal Generalist: General-Level

Monotonicity Across Levels

Key Attribute:

- *If a generalist is rated at the highest level* ***k****, it should achieve valid scores at all levels from 2 to* ***k****.*
- *As the level increases, the expected scores should decrease:* $S_{k-1} > S_k$

- *The monotonicity reflects increasing task difficulty and stricter capability demands at higher levels.*
- *The property ensures that <u>stronger generalists maintain consistent performance across multiple difficulty levels</u>.*
- *It provides a <u>realistic</u> and <u>interpretable</u> evaluation standard for generalist models.*

Path to Multimodal Generalist: General-Level

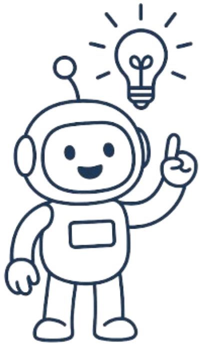
Encouraging Rich and Balanced Multimodal Task Support

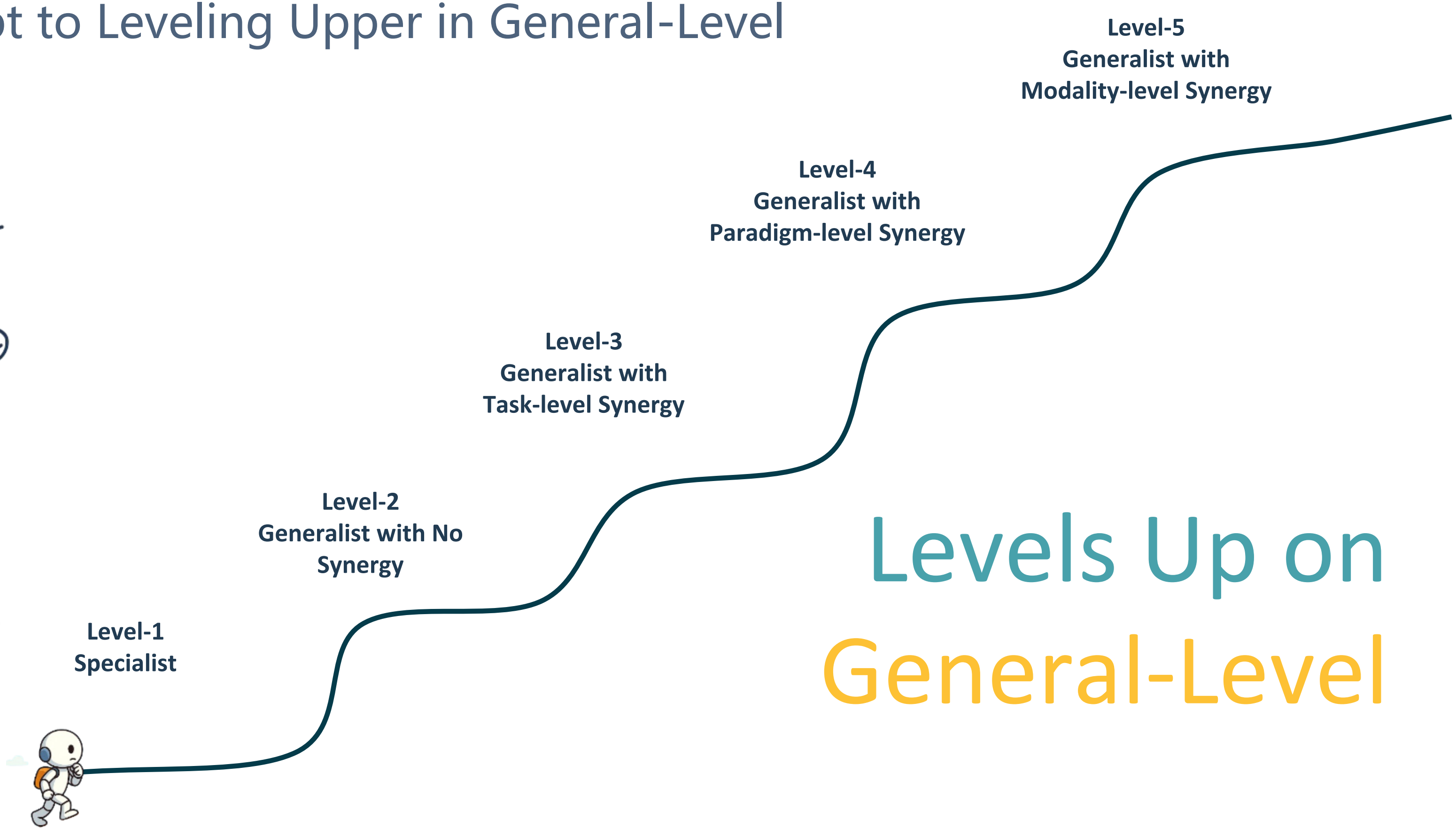
Key Attribute:

- *More task, the better*
- *More balance, the better*

Path to Multimodal Generalist: General-Level

Receipt to Leveling Upper in General-Level





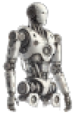
 **Level-2**: Generalists of Unified Comprehension and/or Generation	Models are task-unified players, e.g., MLLMs, capable of supporting different modalities and tasks. Such MLLMs can integrate various models through existing encoding and decoding technologies to achieve aggregation and unification of various modalities and tasks (such as comprehension and generation tasks).	The average score between Comprehension and Generation tasks (i.e., across all tasks) represents the score at this level. A model that can score non-zero on the data is considered capable of supporting that task. The more supported tasks and the higher the scores, the higher its overall score: $$S_2 = \frac{1}{2}\left(\frac{1}{M}\sum_{i=1}^{M}\sigma_i^C + \frac{1}{N}\sum_{j=1}^{N}\sigma_i^G\right)$$	Unified-io-2 (Lu et al., 2024a), AnyGPT (Zhan et al., 2024), NExT-GPT (Wu et al., 2024a), SEED-LLaMA (Ge et al., 2023), GPT-4V (OpenAI, 2022b), $\cdots$

Supporting as many tasks and functionalities as possible

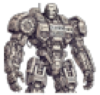
Level-1: Specialists	Various current models, each fine-tuned on a specific task or dataset of specific modalities, are task-specific players (i.e., SoTA specialists). This includes various learning tasks, such as linguistic/visual recognition, classification, generation, segmentation, grounding, inpainting, and more.	For each task in the benchmark (i-th task), the current SoTA specialist's score is recorded as: $$\sigma_i^{sota}$$	CLIP (Li et al., 2022), FLUX (Labs, 2023), FastSpeech2 (Ren et al., 2021), $\cdots$

 **Level-3**: Generalists with **synergy** in Comprehension and/or Generation	Models are task-unified players, and synergy is in Comprehension and/or Generation. MLLMs enhance several tasks' performance beyond corresponding SoTA scores through joint learning across multiple tasks due to the synergy effect.	Assign a mask weight of 0 or 1 to each task; mask=1 only if the corresponding score (σ_i^C or σ_i^G) exceeds the SoTA specialist's score, otherwise mask=0. Then, calculate the average score between S_C and S_G. The more tasks to surpass the SoTA specialist, the higher the S_3: $S_3 = \frac{1}{2}\left(S_G + S_C\right)$, where $S_C = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i^C & \text{if } \sigma_i^C \geq \sigma_{sota}^C \\ 0 & \text{otherwise}\end{cases}$ $S_G = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j^G & \text{if } \sigma_j^G \geq \sigma_{sota}^G \\ 0 & \text{otherwise}\end{cases}$	GPT-4o (OpenAI, 2022b), Gemini-1.5 (Team et al., 2024a), Claude-3.5 (Team, 2024), DeepSeek-VL (Lu et al., 2024b), LLaVA-One-Vision (Li et al., 2024d), Qwen2-VL (Wang et al., 2024a), InternVL2.5 (Chen et al., 2024c), Phi-3.5-Vision (Abdin et al., 2024), $\cdots$

Generalists achieving as stronger synergy and cross as many tasks as possible

 **Level-3**:	Models are task-unified players, and syn-	Assign a mask weight of 0 or 1 to each task;	GPT-4o (OpenAI,
 **Level-4**: Generalists with **synergy** across Comprehension and Generation	Models are task-unified players, and synergy is across Comprehension and Generation.	Calculate the harmonic mean between Comprehension and Generation scores. The stronger synergy a model has between Comprehension and Generation tasks, the higher the score: $$S_4 = \frac{2S_C S_G}{S_C + S_G}$$	Mini-Gemini (Li et al., 2024c), Vitron-V1 (Fei et al., 2024a), Emu2-37B (Sun et al., 2024), $\cdots$
		$$S_C = \frac{1}{M}\sum_{i=1}^{M}\begin{cases}\sigma_i^C & \text{if } \sigma_i^C \geq \sigma_{sota}^C \\ 0 & \text{otherwise}\end{cases}$$ $$S_G = \frac{1}{N}\sum_{j=1}^{N}\begin{cases}\sigma_j^G & \text{if } \sigma_j^G \geq \sigma_{sota}^G \\ 0 & \text{otherwise}\end{cases}$$	2024d), Qwen2-VL (Wang et al., 2024a), InternVL2.5 (Chen et al., 2024c), Phi-3.5-Vision (Abdin et al., 2024), $\cdots$

Generalists in unified comprehension and generation capability with synergy in between

 **Level-5**: Generalists with **total synergy** across Comprehension, Generation and Language	Models are task-unified players, preserving the synergy effect across Comprehension, Generation, and Language. In other words, the model not only achieves cross-modality synergy between Comprehension and Generation groups but also further realizes synergy with language. The Language intelligence can enhance multimodal intelligence and vice versa; understanding multimodal information can also aid in understanding language.	Calculate the model's average score exceeding SoTA NLP specialists on NLP benchmark data; normalize it to a [0,1] weight, and multiply it by the score from level-4 as the level-5 score: $S_5 = S_4 \times w_L$, where $w_L = \frac{S_L}{S_{\text{total}}}$, where $S_L = \frac{1}{T} \sum_{k=1}^{T} \begin{cases} \sigma_k & \text{if } \sigma_k \geq \sigma_{\text{sota}} \\ 0 & \text{otherwise} \end{cases}$	*None found yet (Let's wait for multimodal ChatGPT moment!)*

Generalists achieving cross-modal synergy with abductive reasoning ability

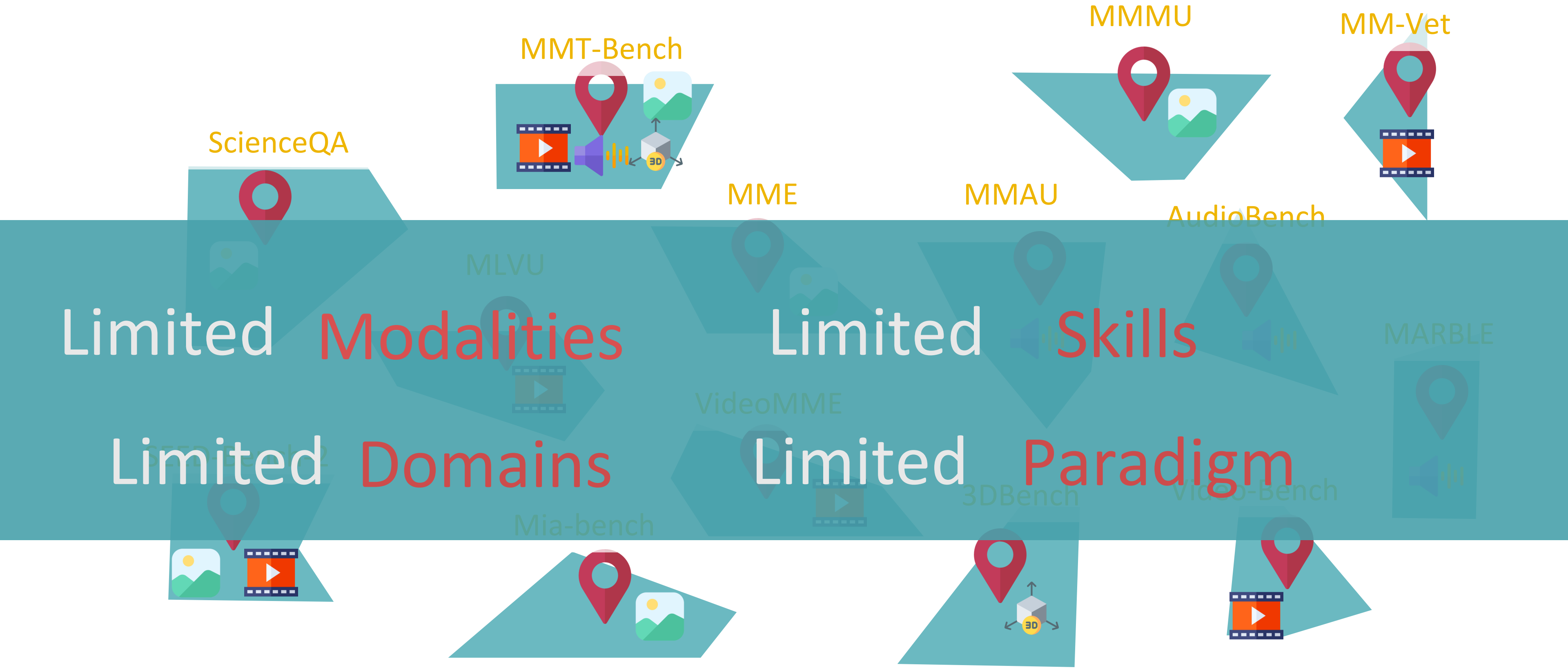
Path to Multimodal Generalist: General-Bench

Why General-Bench?

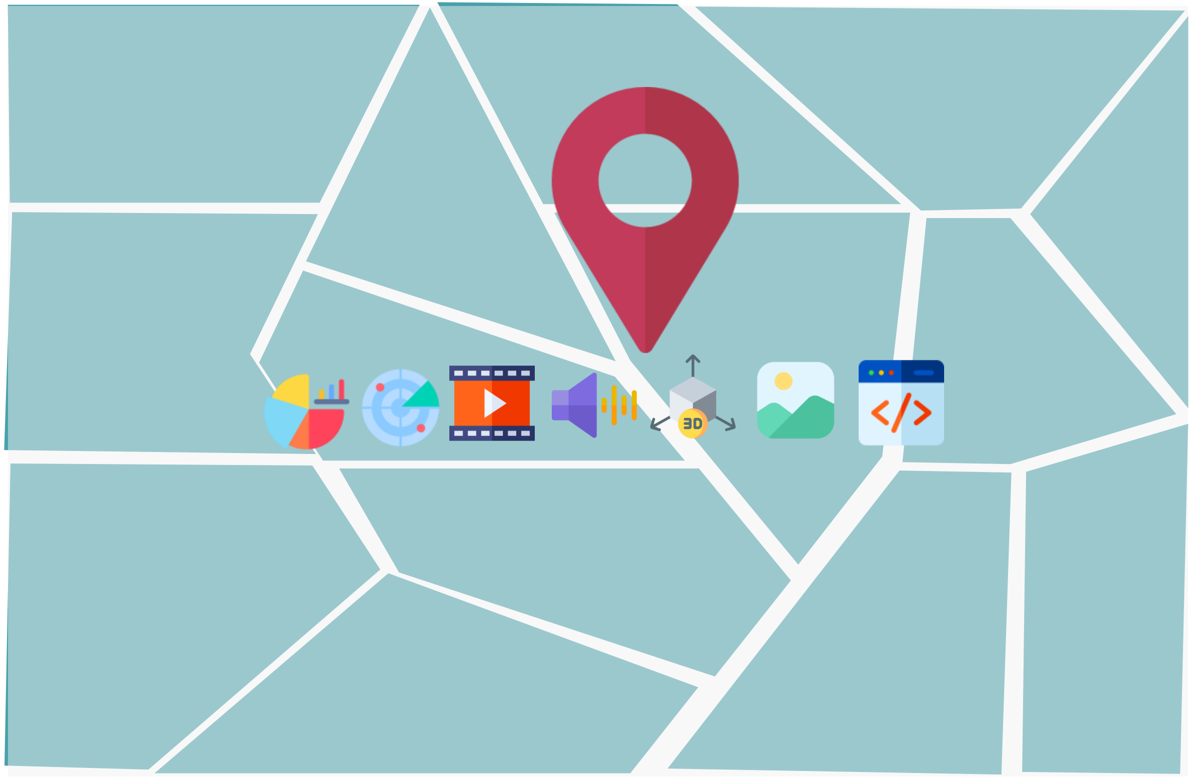
So, where to evaluate generalist models across these five levels?

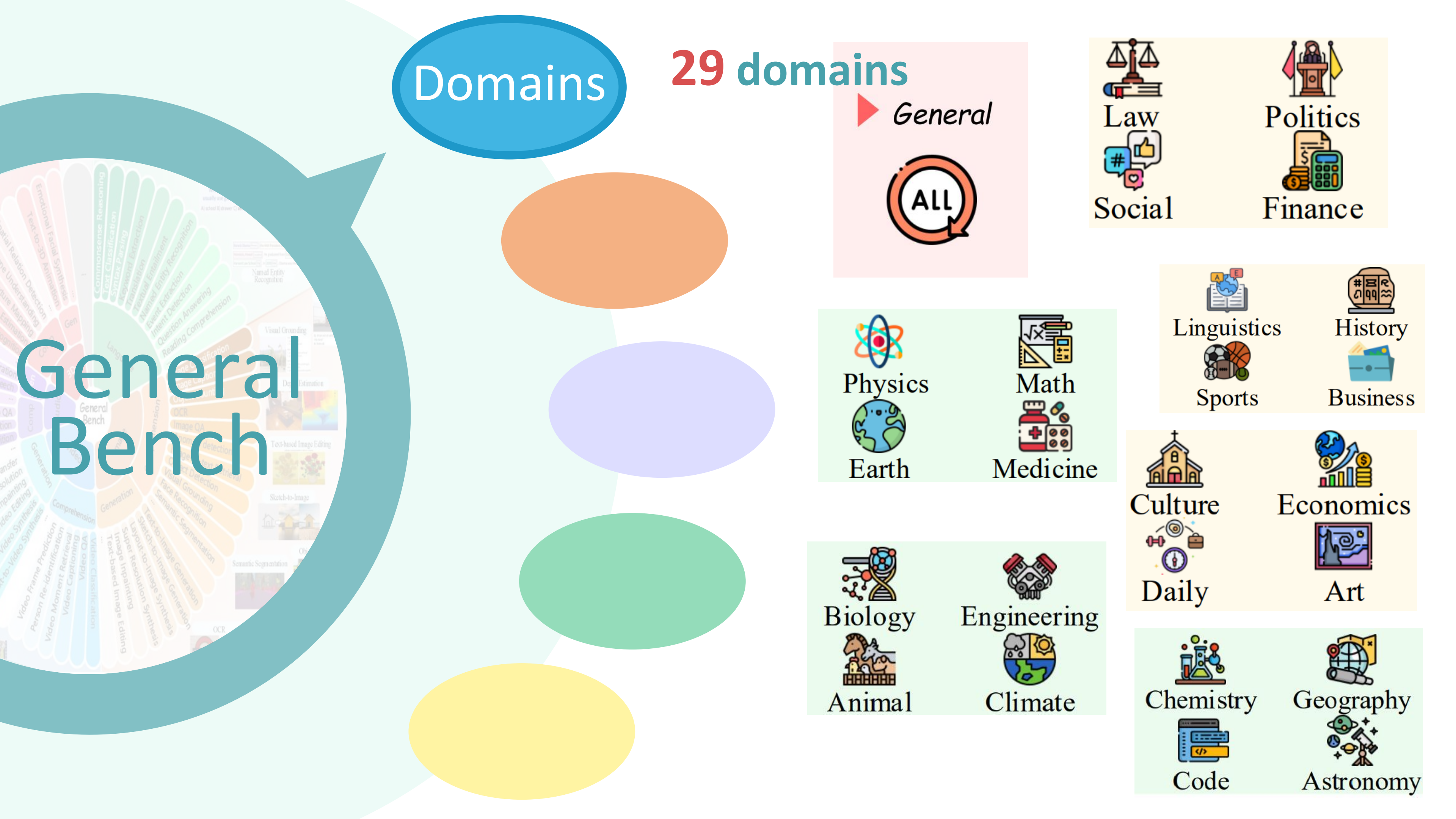
Using Existing Benchmark

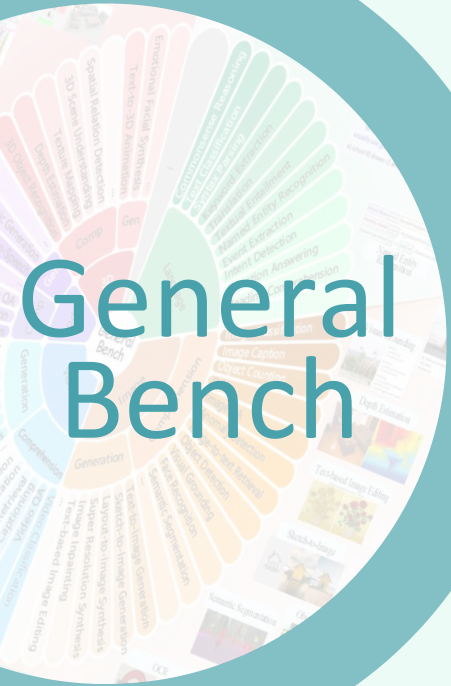
Existing MLLM Benchmark



We Propose General-Bench




General Bench
Domains
29 domains
General
ALL
Law
Politics
Social
Finance
Linguistics
History
Sports
Business
Physics
Math
Earth
Medicine
Culture
Economics
Daily
Art
Biology
Engineering
Animal
Climate
Chemistry
Geography
Code
Astronomy

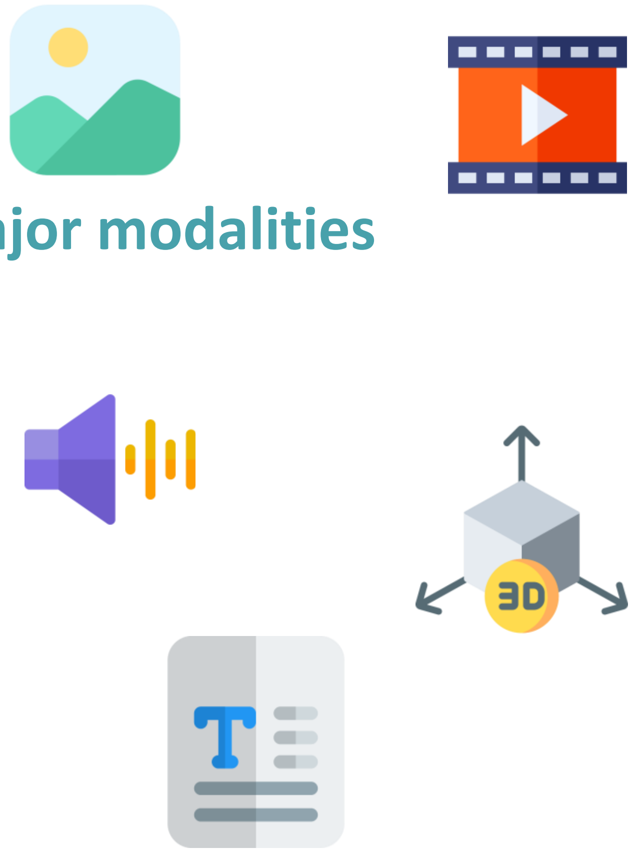


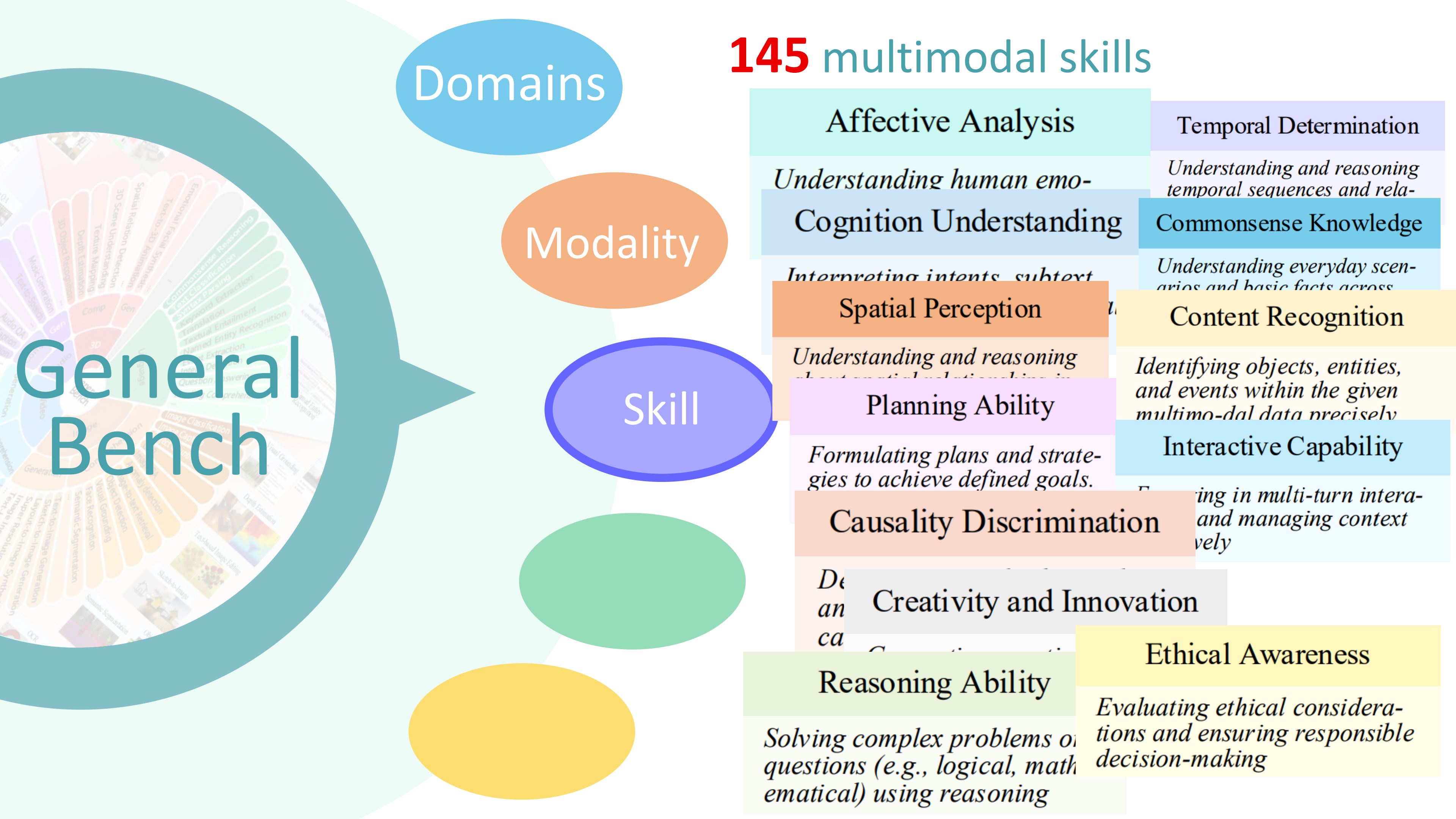
General Bench

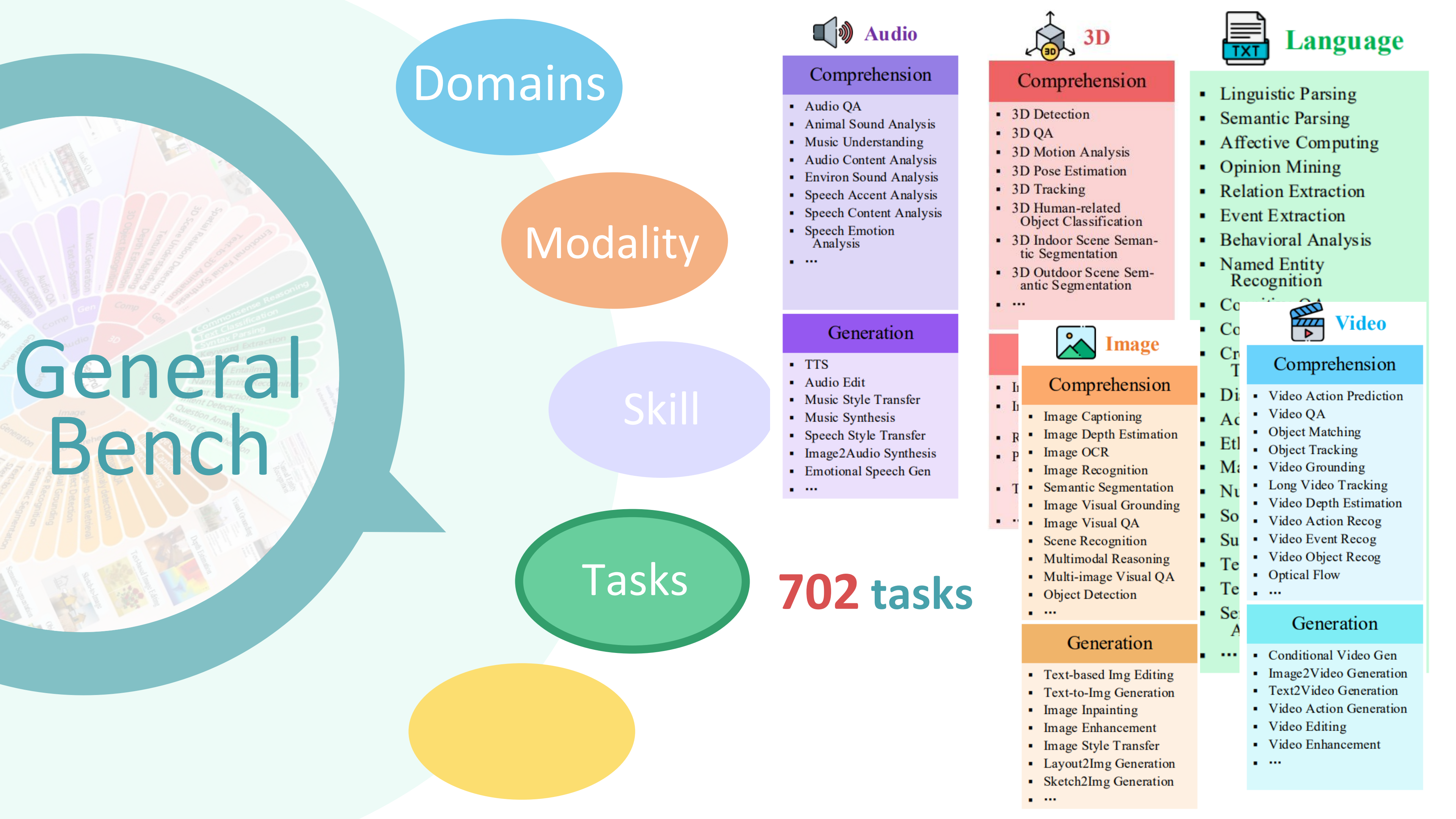
Domains

Modality

5 major modalities




General Bench
Domains
Modality
Skill
145 multimodal skills
Affective Analysis
Understanding human emo-
Temporal Determination
Understanding and reasoning
temporal sequences and rela-
Cognition Understanding
Interpreting intents, subtext
Commonsense Knowledge
Understanding everyday scen-
arios and basic facts across
Spatial Perception
Understanding and reasoning
Content Recognition
Identifying objects, entities,
and events within the given
multimo-dal data precisely
Planning Ability
Formulating plans and strate-
gies to achieve defined goals.
Interactive Capability
ing in multi-turn intera-
and managing context
vely
Causality Discrimination
Creativity and Innovation
Ethical Awareness
Evaluating ethical considera-
tions and ensuring responsible
decision-making
Reasoning Ability
Solving complex problems o
questions (e.g., logical, math
ematical) using reasoning


General Bench
Domains
Modality
Skill
Tasks
702 tasks
Audio
Comprehension
Audio QA
Animal Sound Analysis
Music Understanding
Audio Content Analysis
Environ Sound Analysis
Speech Accent Analysis
Speech Content Analysis
Speech Emotion Analysis
...
Generation
TTS
Audio Edit
Music Style Transfer
Music Synthesis
Speech Style Transfer
Image2Audio Synthesis
Emotional Speech Gen
...
3D
Comprehension
3D Detection
3D QA
3D Motion Analysis
3D Pose Estimation
3D Tracking
3D Human-related Object Classification
3D Indoor Scene Semantic Segmentation
3D Outdoor Scene Semantic Segmentation
...
Image
Comprehension
Image Captioning
Image Depth Estimation
Image OCR
Image Recognition
Semantic Segmentation
Image Visual Grounding
Image Visual QA
Scene Recognition
Multimodal Reasoning
Multi-image Visual QA
Object Detection
...
Generation
Text-based Img Editing
Text-to-Img Generation
Image Inpainting
Image Enhancement
Image Style Transfer
Layout2Img Generation
Sketch2Img Generation
...
Language
Linguistic Parsing
Semantic Parsing
Affective Computing
Opinion Mining
Relation Extraction
Event Extraction
Behavioral Analysis
Named Entity Recognition
...
Video
Comprehension
Video Action Prediction
Video QA
Object Matching
Object Tracking
Video Grounding
Long Video Tracking
Video Depth Estimation
Video Action Recog
Video Event Recog
Video Object Recog
Optical Flow
...
Generation
Conditional Video Gen
Image2Video Generation
Text2Video Generation
Video Action Generation
Video Editing
Video Enhancement
...

General Bench

Domains

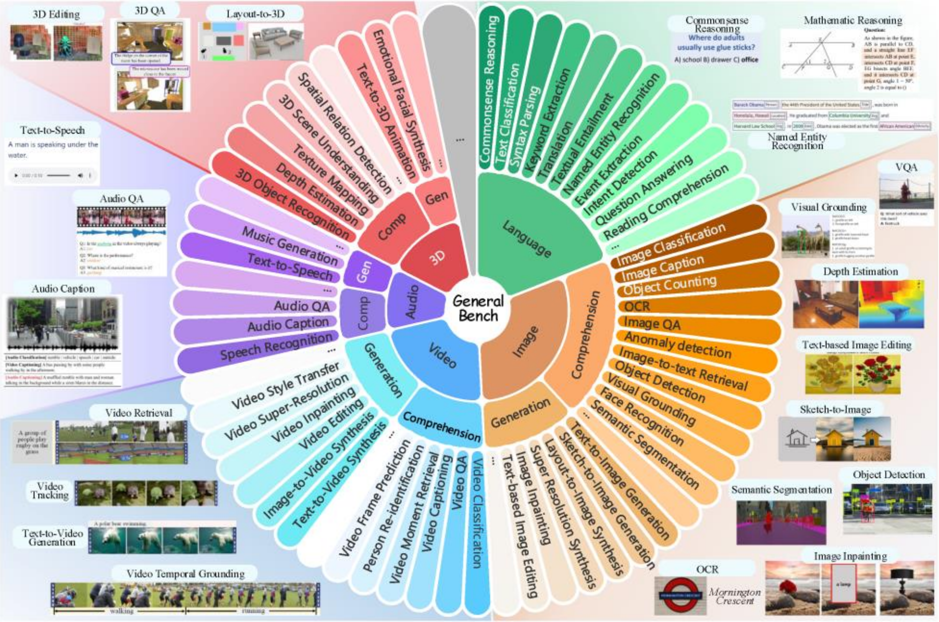
Modality

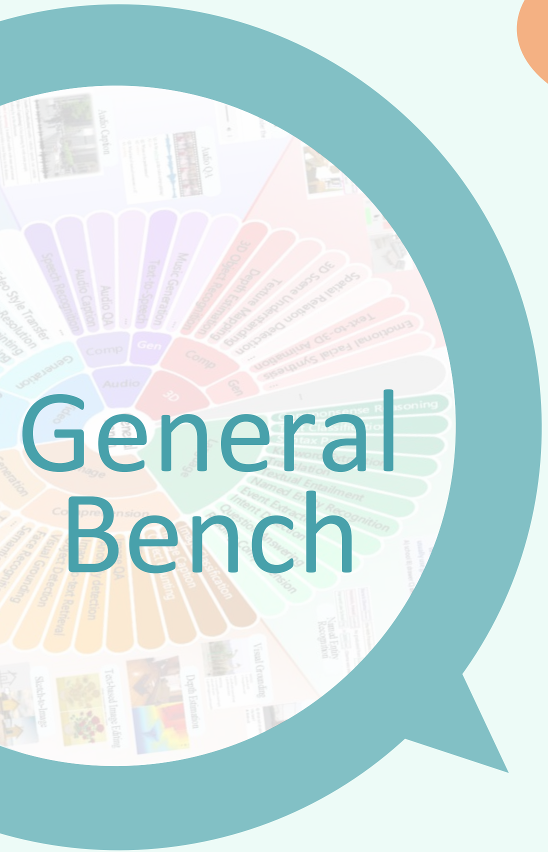
Skill

Tasks

Sample

325,876 samples



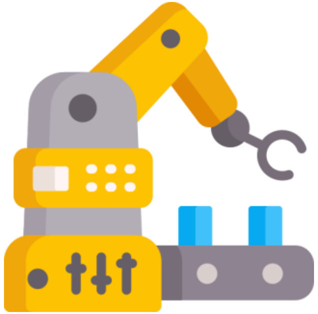


General Bench

- Modality
- Skill
- Tasks
- Sample
- Paradigm

Comprehension

Generation



Path to Multimodal Generalist: General-Bench

Statistics of General-Bench

		Image		Video		Audio		3D		Language	*TOTAL*
		Comp	Gen	Comp	Gen	Comp	Gen	Comp	Gen		
#Skill	Single	40	15	20	6	9	11	13	9	22	145
	Sum	55		26		20		22			
#Task	Single	271	45	126	46	24	20	30	22	118	702
	Sum	316		170		44		52			
#Instance	Single	124,880	26,610	44,442	16,430	11,247	9,516	23,705	10,614	58,432	325,876
	Sum	151,490		60,872		20,763		34,319			

Path to Multimodal Generalist: General-Bench

Statistics of General-Bench

Benchmark	SEED-Bench	MMBench	MMMU	LVLM-eHub	MMIU	MMT-Bench	MEGA-Bench	**General-Bench**
Modality	Txt,Img,Vid	Txt,Img	Txt,Img	Txt,Img	Txt,Img,Vid, Point-Cloud,Depth	Txt,Img,Vid, Point-Cloud	Txt,Img,Vid	Txt,Img,Vid,Aud, Time,Depth,3D-RGB, Point-Cloud,Infrared, Spectrogram,Radar, Code,Doc,Graph,· · ·
Task Scheme	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.	Comp.+Gen.
# Domain	1	1	6	1	1	4	5	**29**
# Skill	12	2	6	6	7	32	10	**145**
# Task	12	20	30	47	52	162	505	**702**
# Sample	19K	3K	11.5K	2.1K	11.7K	31K	8K	**325.8K**
Answer Form	MC-QA	MC-QA	MC-QA	MC-QA	MC-QA	MC-QA	Free-Form	Free-Form
# Metric	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Origin (45)	Origin (**58**)
Annotation	Manual	Repurposed	Manual	Repurposed	Repurposed	Repurposed	Manual	Manual
# Tested Models	12	21	24	8	22	30	22	**172+102**

Path t

How a

#	Metric	Range	Calculation	Representative Tasks
• General				
1	Acc↑	[0,1]	Accuracy is defined as the ratio of correctly classified instances to the total number of instances.	Classification
2	Macro-Acc↑	[0,1]	Macro-Acc evaluates how well a model performs on average across all classes, regardless of class imbalance.	Event Relation Prediction
3	EM-Acc↑	[0,1]	Exact Match Accuracy evaluates the percentage of predictions that are exactly the same as their corresponding references.	QA, machine translation, or summarization
4	AP↑	[0,1]	AP, Average Precision, is a metric used to evaluate the performance of object detection tasks, reflecting the overall precision-recall trade-off across multiple thresholds.	Anomaly Detection
5	mAP ↑	[0,1]	mAP, Mean Average Precision, is the mean of Average Precision values across all queries or instances:	2D/3D Detection
6	F1↑	[0,1]	F1 score is the harmonic mean of Precision and Recall.	QA
7	Micro-F1↑	[0,1]	Micro-F1 score is the harmonic mean of the Micro-averaged precision and recall.	Classification
8	AUC↑	[0,1]	AUC is used in binary classification tasks and measures the area under the ROC curve. It represents the model's ability to distinguish between classes.	Image Generation
• Ranking-related				
9	R@k↑	[0,1]	R@k measures the Recall rate at the top k results in tasks like image retrieval, where the true positive must appear within the top k predicted results.	Image Scene Graph Parsing
10	AP@k↑	[0,1]	AP@k is the Average Precision calculated at an IoU threshold of k (k¡1). This metric is typically used when higher overlap between retrieved items and ground truth items is required.	Object Detection
11	mAP@k↑	[0,1]	mAP@k refers to the mean Average Precision where the Intersection over Union (IoU) threshold is set to k (k¡1).	Object Detection
12	EM@1↑	[0,1]	Exact Match at 1 evaluates the proportion of instances for which the model's top prediction exactly matches the correct answer.	3D Question Answering
13	ANLS↑	[0,1]	ANLS, Average Normalized Levenshtein Similarity, measures how well a model ranks items in a list based on their relevance to a query.	OCR
• Regression-related				
14	MAE ↓	$[0,\infty)$	MAE, Mean Absolute Error, measures the average of the absolute differences between the predicted values and the actual values. It's typically used in regression tasks.	Object Counting
15	RMS ↓	$[0,\infty)$	RMS, Root Mean Square, is a metric for regression tasks that measures the square root of the average squared differences between the predicted values and true values.	Image Depth Estimation
16	MSE ↓	$[0,\infty)$	MSE, Mean Squared Error, is commonly used for regression tasks and measures the average squared differences between predicted values and actual values.	Object Matting
17	RMSE ↓	$[0,\infty)$	RMSE, Root Mean Squared Error.	Time Series Prediction
• Text Generation-related				

Pa

#	Metric	Range	Calculation	Representative Tasks
50	CLAP↑	[0,1]	CLAP (Contrastive Language-Audio Pretraining) evaluates the alignment between generated audio and text. It is derived from a contrastive learning framework where embeddings of audio and text are trained to be close in a shared latent space if they are semantically related.	Audio Editing
51	Style-CLAP ↑	[0,1]	Style-CLAP calculates the CLAP cosine similarity between the generated Mel spectrograms and the corresponding textual description of the style to evaluate style fit.	Music Style Transfer
52	MCD ↓	$[0,\infty)$	Mel-cepstral distortion (MCD) measures the spectral distance between the mel-cepstral coefficients (MCCs) of generated speech and reference speech, providing an indication of how closely the generated speech resembles the reference in terms of acoustic characteristics.	Speech Synthesis
53	WER ↓	[0,1]	WER (Word Error Rate) measures the percentage of errors in the transcribed output compared to the reference transcription.	TTS
54	FAD ↓	$[0,\infty)$	Frechet audio distance (FAD) evaluates the quality and realism of generated audio, and measures the similarity between the distribution of features obtained by VGGish in generated audio and those in a set of real (reference) audio samples.	Video-to-Audio
55	PCC ↑	[0,1]	Pitch-Class Consistency (PCC) is a metric used in the evaluation of generated music to assess how consistent the pitch classes (e.g., notes) are across pairs of bars in a piece of music. It measures the overlapping area between the pitch-class histograms of different bars, ensuring that the generated music maintains harmonic coherence.	Music Generation
• Human-aware Evaluation				
56	UPR ↑	[0,1]	UPR, User Preference Rates, UPR measures the proportion of times a particular system or model is preferred over alternatives in a set of user evaluations. It reflects the subjective preferences of users and is often derived from pairwise comparisons or ranking experiments.	Video Style Transfer
57	MOS ↑	[1,5]	Mean Opinion Score (MOS), in which human raters listen to synthesized speech and assess its naturalness, quality, and intelligibility using a 5-point Likert scale.	Speech Generation
58	GPT-Score ↑	[0,1]	GPT-Score evaluates the instruction following rate with GPT assistance, as an alternative to human evaluation.	Audio Question Answering

Path to Multimodal Generalist: General-Bench

How are the evaluation metrics?

- Mapping Functions of Scoring Metric

- Normalizing **MAE**:

$$y = 2 \times \text{sigmoid}\left(\frac{50}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **RMS**:

$$y = 2 \times \text{sigmoid}\left(\frac{50}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **MSE**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **RMSE**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **absRel**:

$$y = 2 \times \text{sigmoid}\left(\frac{0.1}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

Path to Multimodal Generalist: General-Bench

How are the evaluation metrics?

- Mapping Functions of Scoring Metric

- Normalizing **RTE**:

$$y = 2 \times \text{sigmoid}\left(\frac{0.5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **CD**:

$$y = 2 \times \text{sigmoid}\left(\frac{1}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **MCD**:

$$y = 2 \times \text{sigmoid}\left(\frac{5}{x}\right) - 1, \quad \text{where } x \in [0, +\infty), \quad y \in (0, 1).$$

- Normalizing **WER**:

$$y = 1 - x, \quad \text{where } x \in [0, 1], \quad y \in [0, 1].$$

- Normalizing **MS-SSIM**:

$$y = \frac{(x + 1)}{2}, \quad \text{where } x \in [-1, 1], \quad y \in [0, 1].$$

- Normalizing **MOS**:

$$y = \frac{x - 1}{4}, \quad \text{where } x \in [1, 5], \quad y \in [0, 1].$$

Path to Multimodal Generalist: General-Bench

How many multimodal generalist are included?

#	Model	Backbone	Size	Modality Support	Paradigm
• Language-oriented (Closed/Open-sourced) Models					
1	Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023)	Llama	8B	Language	/
2	Gemma-2-9b-it (Team et al., 2024b)	Gemma	9B	Language	/
3	GPT-J (Wang and Komatsuzaki, 2021)	GPT-J	6B	Language	/
4	ChatGLM-6B (GLM et al., 2024)	ChatGLM	6B	Language	/
5	Qwen2.5-7B-Instruct (Yang et al., 2024a)	Qwen2.5	7B	Language	/

Path to Multimodal Generalist: General-Bench

How many multimodal generalist are included?

6	InternLM2-Chat-7B (Cai et al., 2024)	InternLM2	7B	Language	/
7	Baichuan2-7B-Chat (Yang et al., 2023)	Baichuan2	7B	Language	/
8	Vicuna-7b-V1.5 (Chiang et al., 2023)	Vicuna	7B	Language	/
9	Falcon3-7B-Instruct (Almazrouei et al., 2023)	Falcon3	7B	Language	/
10	Ministral-8B-Instruct-2410 (Jiang et al., 2024a)	Ministral	8B	Language	/
11	Yi-lightning (Young et al., 2024)	Llama	6B	Language	/
12	GPT-3.5-turbo (OpenAI, 2022a)	GPT3.5	/	Language	/

Path to Multimodal Generalist: General-Bench

How many multimodal generalist are included?

• Multimodal Close-sourced Models					
1	GPT4-V (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
2	GPT4-o-mini (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
3	GPT4-o (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
4	GPT4-o-4096 (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
5	ChatGPT-o-latest (OpenAI, 2022b)	GPT4	/	Language, Image	Comprehension
6	Claude-3.5-Sonnet (Team, 2024)	Claude-3.5-Sonnet	/	Language, Image	Comprehension
7	Claude-3.5-Opus (Team, 2024)	Claude-3.5-Opus	/	Language, Image	Comprehension
8	Gemini-1.5-Pro (Team et al., 2024a)	Gemini	/	Language, Image	Comprehension
9	Gemini-1.5-Flash (Team et al., 2024a)	Gemini	/	Language, Image	Comprehension

Path to Multimodal Generalist: General-Bench

• Multimodal Open-sourced Models					
1	Yi-vision-v2 (Young et al., 2024)	LLaVa	6B	Language, Image	Comprehension
2	Emu2-37B (Sun et al., 2024)	LLaMA-33B	37B	Language, Image	Comprehension+Generation
3	InternVL2.5-2B (Chen et al., 2024c)	internlm2_5-1_8b-chat	2B	Language, Image	Comprehension
4	InternVL2.5-4B (Chen et al., 2024c)	Qwen2.5-3B-Instruct	4B	Language, Image	Comprehension
5	InternVL2.5-8B (Chen et al., 2024c)	internlm2_5-7b-chat	8B	Language, Image	Comprehension
6	Mini-InternVL-Chat-2B-V1-5 (Gao et al., 2024)	InternLM2-Chat-1.8B	2B	Language, Image	Comprehension
7	Mini-InternVL-Chat-4B-V1-5 (Gao et al., 2024)	Phi-3-mini-128k-instruct	4B	Language, Image	Comprehension
8	InternLM-XComposer2-VL-1.8B (Dong et al., 2024)	InternLM2-Chat-1.8B	1.8B	Language, Image	Comprehension
9	MoE-LLAVA-Phi2-2.7B-4e-384 (Lin et al., 2024a)	Phi2	2.7B	Language, Image	Comprehension
10	Monkey-10B-chat (Li et al., 2024e)	Qwev-7B	10B	Language, Image	Comprehension

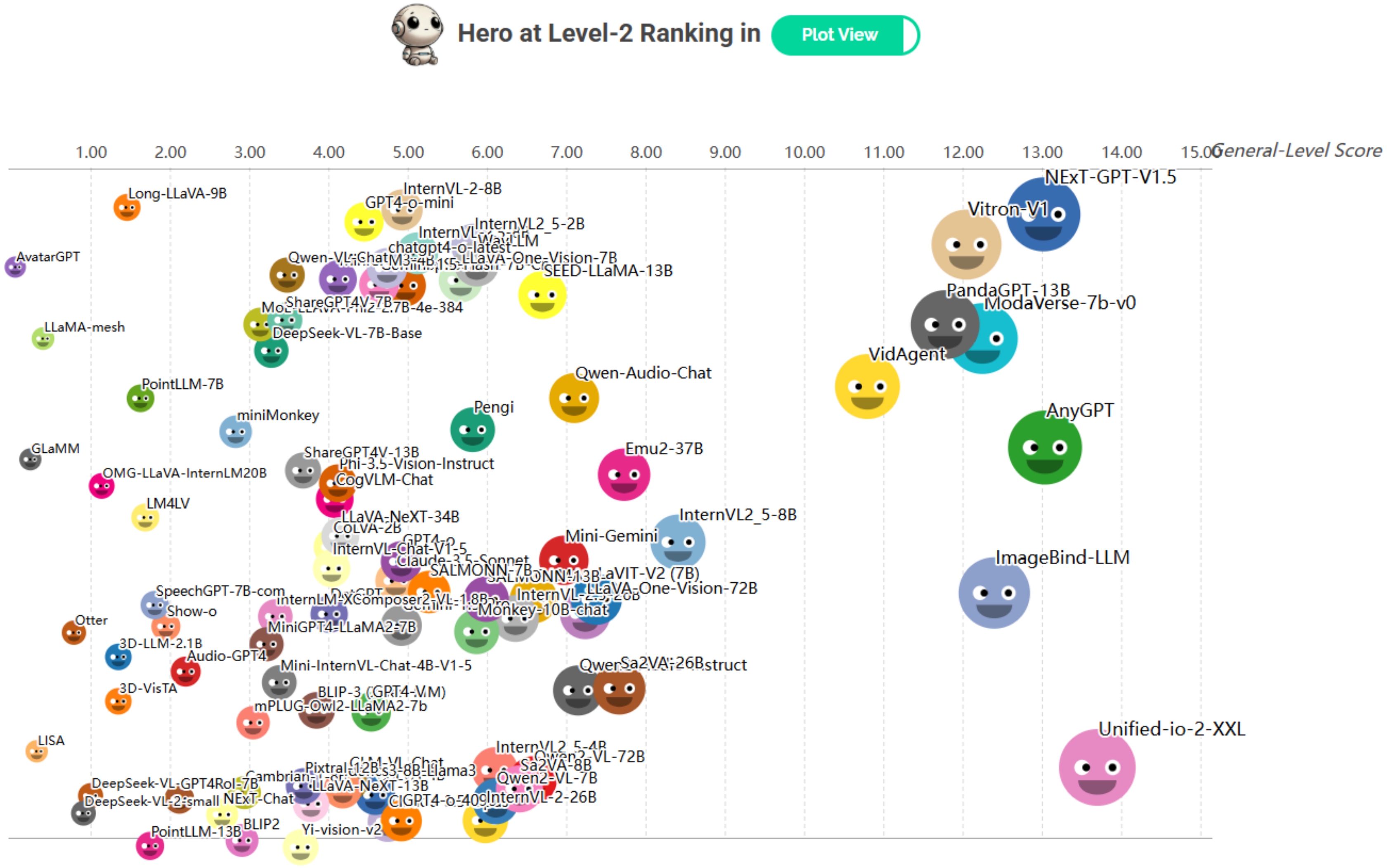
Path to Multimodal Generalist: General-Bench

How many multimodal generalist are included?

62	PointLLM-13B (Xu et al., 2025)	LLaMA	13B	Language, 3D	Comprehension
63	3D-VisTA (Zhu et al., 2023b)	BERT	1.3B	Language, 3D	Comprehension
64	AvatarGPT (Zhou et al., 2024a)	T5-large	770M	Language, 3D	Comprehension
65	MotionGPT-T5 (Jiang et al., 2024b)	T5	220M	Language, 3D	Generation
66	MotionGPT-LLaMA (Zhang et al., 2023e)	LLaMA	13B	Language, 3D	Generation
67	LLaMA-mesh (Zhang et al., 2023e)	LLaMA	7B	Language, 3D	Generation
68	GAMA (Ghosh et al., 2024)	Llama-2-7b-chat	7B	Language, Audio	Comprehension
69	Pengi (Deshmukh et al., 2023)	GPT2-base	124M	Language, Audio	Comprehension
70	WavLLM (Hu et al., 2024b)	LLaMA-2-7B-chat	7B	Language, Audio	Comprehension
71	SALMONN-7B (Tang et al., 2023)	Vicuna-7B	7B	Language, Audio (Speech)	Comprehension

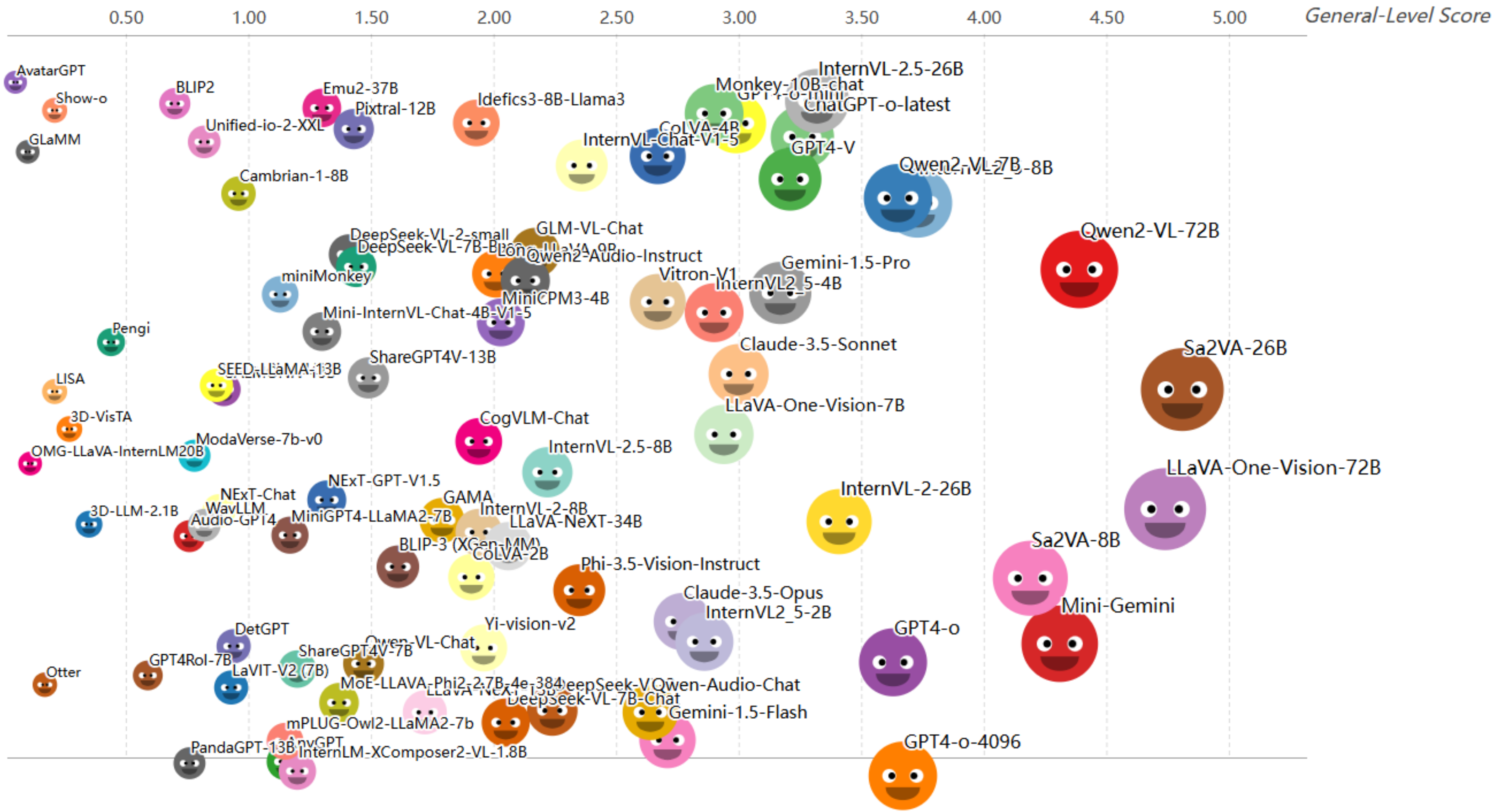
Path to Multimodal Generalist: General-Bench

What General-Bench Unveils? —— General-Level Leaderboards

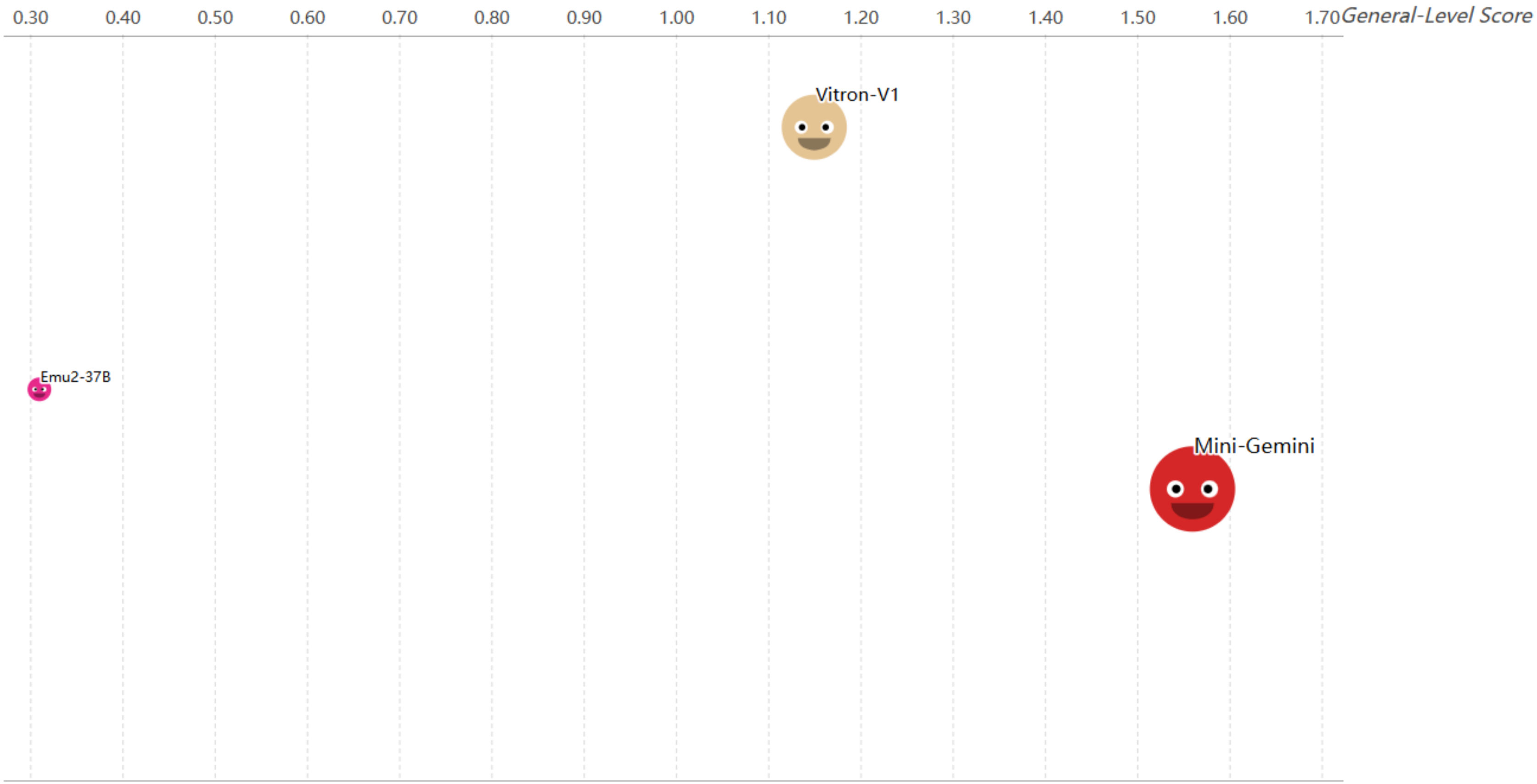


Hero at Level-2 Ranking in
Plot View
General-Level Score
1.00
2.00
3.00
4.00
5.00
6.00
7.00
8.00
9.00
10.00
11.00
12.00
13.00
14.00
15.00
NExT-GPT-V1.5
Vitron-V1
PandaGPT-13B
ModaVerse-7b-v0
VidAgent
AnyGPT
ImageBind-LLM
Unified-io-2-XXL
Long-LLaVA-9B
InternVL-2-8B
GPT4-o-mini
InternVL2_5-2B
chatgpt4-o-latest
LLaVA-One-Vision-7B
SEED-LLaMA-13B
AvatarGPT
ShareGPT4V-7B
DeepSeek-VL-7B-Base
LLaMA-mesh
PointLLM-7B
Qwen-Audio-Chat
Pengi
miniMonkey
GLaMM
ShareGPT4V-13B
Phi-3.5-Vision-Instruct
CogVLM-Chat
Emu2-37B
OMG-LLaVA-InternLM20B
LM4LV
InternVL2_5-8B
LLaVA-NeXT-34B
Mini-Gemini
InternVL-Chat-V1-5
Claude-3.5-Sonnet
SALMONN-7B
LLaVA-One-Vision-72B
SpeechGPT-7B-com
Show-o
Monkey-10B-chat
Otter
MiniGPT4-LLaMA2-7B
3D-LLM-2.1B
Audio-GPT4
Mini-InternVL-Chat-4B-V1-5
3D-VisTA
BLIP-3 (GPT4-V.M)
mPLUG-Owl2-LLaMA2-7b
LISA
Qwen2-VL-72B
Sa2VA-8B
Qwen2-VL-7B
InternVL-2-26B
DeepSeek-VL-2-small
NExT-Chat
LLaVA-NeXT-13B
PointLLM-13B
BLIP2
Yi-vision-v2

Hero at Level-3 Ranking in Plot View

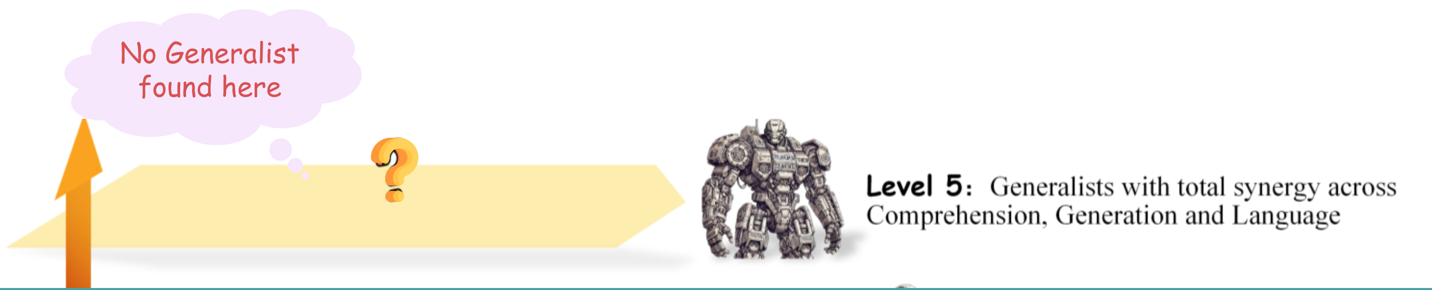


Hero at Level-4 Ranking in Plot View



Hero at Level-5 Ranking in Plot View

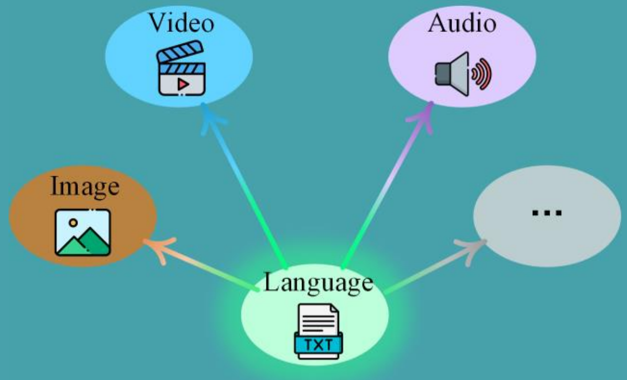
Submit your multimodal generalist to the leaderboard!

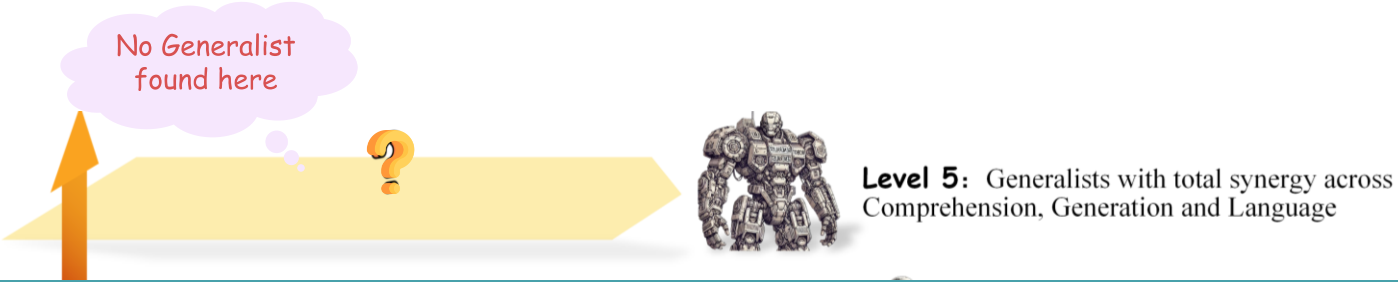


General Bench

Current Multimodal Generalists (MLLMs) mostly has Language-centric Architectures with LLMs as backbones

Language-centric Multimodal Generalists hinge on Language Intelligence, rather than Native Multimodal Intelligence

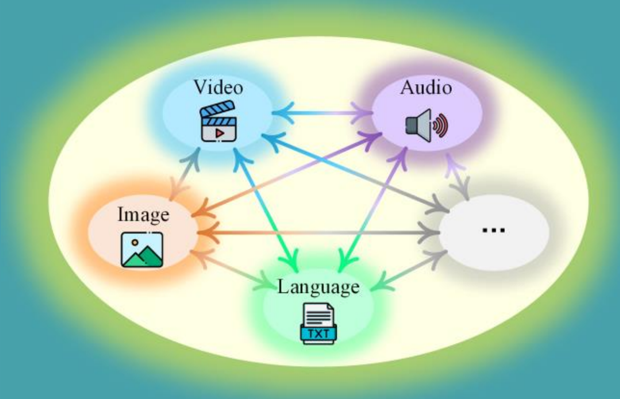




General Bench

Toward Level-5:

Achieving **Total Synergy** Across Modalities, Tasks, Paradigms for Native Multimodal intelligence



Path to Multimodal Generalist: General-Bench

What General-Bench Unveils? —— Quantitative Performances

Model	#1 #11 #21 #31	#2 #12 #22 #32	#3 #13 #23 #33	#4 #14 #24 #34	#5 #15 #25 #35	#6 #16 #26 #36	#7 #17 #27 #37	#8 #18 #28 #38	#9 #19 #29 #39	#10 #20 #30 #40	Task Completion: #Supported Task	Task Completion: #Win-over-Specialist	Level Score on Image: Level-2	Level Score on Image: Level-3	Level Score on Image: Level-4
	Image Comprehension Skill (Avg within each #I-C Group)														
SoTA Specialist	51.27	53.32	42.04	22.30	39.02	22.42	46.02	15.67	51.20	28.01	/	/	/	/	/
	36.40	65.15	43.78	58.90	63.73	87.84	58.66	72.25	34.51	95.70					
	70.00	50.40	65.97	16.60	78.00	50.48	19.90	53.55	64.10	35.90					
	39.80	57.20	54.60	63.27	29.60	87.10	98.00	39.60	36.42	82.02					
GPT-4V	69.42	58.64	39.54	0.00	66.18	36.08	61.74	0.00	16.90	20.88	177 (65.1%)	105 (38.6%)	18.16	12.85	0.00
	0.00	0.00	51.04	63.52	0.00	70.90	51.60	0.00	0.00	0.00					
	71.90	37.12	50.30	16.06	72.20	0.00	0.00	72.51	0.00	97.98					
	40.05	0.00	90.40	0.00	31.64	89.10	22.22	22.54	18.08	84.84					
GPT-4o	73.87	63.42	43.23	0.00	71.56	39.65	68.83	0.00	67.80	23.24	177 (65.1%)	112 (41.2%)	19.67	14.51	0.00
	0.00	0.00	71.23	61.54	0.00	79.38	55.25	0.00	0.00	0.00					
	81.30	39.61	48.63	15.12	93.00	0.00	0.00	77.53	0.00	98.79					
	44.30	0.00	90.40	0.00	33.47	91.20	35.56	24.80	21.12	87.88					
Gemini-1.5-Pro	72.33	23.41	39.39	0.00	62.38	34.30	66.25	0.00	59.20	23.79	177 (65.1%)	101 (37.1%)	19.67	12.66	0.00
	0.00	0.00	60.86	40.10	0.00	0.00	58.09	0.00	0.00	0.00					
	84.57	31.55	60.87	15.20	86.40	0.00	0.00	76.72	0.00	96.76					
	36.41	0.00	98.00	0.00	38.45	92.00	30.37	22.18	21.20	83.23					
Gemini-1.5-Flash	67.00	25.79	37.85	0.00	59.45	29.91	63.61	0.00	56.50	22.19	177 (65.1%)	94 (34.6%)	18.54	10.85	0.00
	0.00	0.00	55.22	32.92	0.00	0.00	54.57	0.00	0.00	0.00					
	80.63	28.97	56.91	16.57	82.60	0.00	0.00	73.57	0.00	93.42					
	28.53	0.00	96.40	0.00	29.97	90.20	27.96	20.64	18.22	80.40					
Claude-3.5-Opus	65.38	57.69	39.95	0.00	63.35	34.50	63.43	0.00	45.62	20.44	178 (65.4%)	93 (34.2%)	19.00	11.08	0.00
	0.00	0.00	60.21	58.15	0.00	66.57	51.23	0.00	0.00	0.00					
	70.39	41.19	54.75	13.87	77.80	0.00	0.00	73.04	0.00	94.65					
	38.28	0.00	91.38	0.00	0.00	87.31	23.87	28.71	25.75	84.65					
Emu2-32B	53.76	7.31	36.62	0.00	41.31	22.22	41.89	0.00	21.20	12.83	178 (65.4%)	52 (19.1%)	30.90	5.18	1.25
	0.00	0.00	39.47	12.20	0.00	0.00	44.51	5.28	0.00	0.00					
	56.33	29.43	45.46	21.45	64.20	0.00	0.00	54.59	0.00	70.34					
	17.73	0.00	72.80	0.00	0.00	73.40	31.72	14.09	18.73	56.97					
Phi-3.5-Vision-Instruct	55.32	3.44	34.16	0.00	42.61	42.04	51.34	0.00	0.00	24.35	179 (65.8%)	85 (31.3%)	16.46	9.39	0.00
	0.00	0.00	41.00	21.77	0.00	0.00	52.13	11.89	0.00	0.00					
	67.56	32.32	51.51	23.70	90.10	0.00	0.00	57.68	0.00	52.02					
	19.31	0.00	83.40	0.00	15.02	80.00	3.98	23.06	25.41	71.31					
Qwen2-VL-72B	66.98	5.74	35.64	0.00	56.58	40.50	48.79	0.00	43.18	25.32	177 (65.1%)	99 (36.4%)	19.41	12.34	0.00
	0.00	0.00	45.66	29.44	0.00	0.00	59.87	10.89	0.00	0.00					
	81.86	38.59	58.99	16.17	97.43	0.00	0.00	72.47	0.00	92.41					
	4.33	0.00	77.64	0.00	16.83	79.34	11.65	29.62	32.22	62.83					

Model	Image Generation Skill (Avg within each #I-G Group)								Task Completion		Level Score on Image		
	#1 **#9**	**#2** **#10**	**#3** **#11**	**#4** **#12**	**#5** **#13**	**#6** **#14**	**#7** **#15**	**#8**	#Supported Task	#Winning-Specialist	Level-2	Level-3	Level-4
SoTA Specialist	18.70 53.16	45.40 16.47	33.77 25.33	16.30 43.93	4.86 20.35	24.00 67.44	99.29 36.11	15.06	/	/	/	/	/
SEED-LLaMA-14B	127.10 30.18	0.00 87.90	37.10 14.58	7.51 175.33	127.42 0.00	98.33 51.82	0.00 62.60	0.00	35 (77.8%)	0 (0.0%)	26.81	3.49	0.00
Emu2-32B	93.52 40.51	0.00 118.55	34.85 15.43	8.53 154.26	101.80 0.00	81.95 57.09	0.00 58.17	0.00	34 (75.6%)	2 (4.4%)	30.90	5.18	1.25
AnyGPT	158.21 28.88	0.00 108.06	40.47 14.91	10.30 193.39	117.21 0.00	115.91 53.02	0.00 64.21	0.00	36 (80.0%)	0 (0.0%)	23.10	1.29	0.00
LaVIT-V2 (7B)	79.79 46.40	0.00 89.78	31.35 15.79	11.87 161.54	149.78 0.00	59.23 50.18	0.00 51.68	0.00	36 (80.0%)	0 (0.0%)	29.50	3.71	0.00
NExT-GPT-V1.5	49.71 28.19	0.00 86.45	6.00 6.53	3.91 53.42	75.71 12.45	41.20 38.98	0.00 72.72	47.30	41 (91.1%)	0 (0.0%)	18.69	3.24	0.00
Vitron-V1	19.78 37.88	0.00 24.89	21.17 17.95	7.45 31.04	32.15 0.00	35.33 48.30	86.53 58.87	23.47	42 (93.3%)	3 (6.7%)	30.13	7.65	4.59

Model	Video Comprehension Skill (Avg within each #V-C Group)										Task Completion		Level Score on Video		
	#1 #11	#2 #12	#3 #13	#4 #14	#5 #15	#6 #16	#7 #17	#8 #18	#9 #19	#10 #20	#Supported Task	#Win-over-Specialist	Level-2	Level-3	Level-4
SoTA Specialist	37.43 45.84	49.64 13.92	21.31 0.14	23.06 48.06	81.85 68.96	85.43 63.62	54.53 77.02	64.83 75.08	40.65 37.20	30.80 44.00	/	/	/	/	/
InternVL-2.5-8B	33.15 0.00	27.54 0.00	14.51 0.00	18.83 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.85	55 (43.7%)	5 (4.0%)	5.76	1.24	0.00
InternVL-2.5-26B	37.03 0.00	32.01 0.00	18.71 0.00	21.57 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 5.30	55 (43.7%)	26 (20.6%)	6.70	3.76	0.00
Qwen2-VL-72B	38.22 0.00	32.32 0.00	19.35 0.00	22.70 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 5.70	55 (43.7%)	22 (17.5%)	6.89	5.22	0.00
DeepSeek-VL-2	21.50 0.00	18.90 0.00	12.10 0.00	12.10 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 3.20	55 (43.7%)	5 (4.0%)	3.98	0.64	0.00
LLaVA-One-Vision-72B	31.20 0.00	31.30 0.00	19.10 0.00	10.60 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 1.70	56 (44.4%)	21 (16.7%)	5.83	3.75	0.00
Sa2VA-8B	33.19 0.00	25.11 60.28	16.75 0.00	8.67 0.00	0.00 19.85	0.00 37.83	0.00 46.36	71.03 42.58	50.95 48.02	0.00 1.48	91 (72.2%)	32 (25.4%)	8.31	4.38	0.00
Sa2VA-26B	35.33 0.00	26.33 0.00	17.58 0.00	10.39 0.00	0.00 28.41	0.00 38.91	0.00 47.10	0.00 43.12	0.00 48.42	0.00 1.70	81 (64.3%)	27 (21.4%)	8.81	4.58	0.00
CoLVA-4B	32.68 0.00	26.45 0.00	13.55 0.00	17.62 0.00	0.00 45.81	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.23	63 (50.0%)	8 (6.3%)	4.78	1.24	0.00
InternVL-2-8B	32.69 0.00	27.09 0.00	14.24 0.00	17.61 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.85	55 (43.7%)	0 (0.0%)	5.64	0.46	0.00
Long-LLaVA-9B	36.14 0.00	26.25 0.00	15.89 0.00	15.53 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 4.20	54 (42.9%)	22 (17.5%)	5.84	3.81	0.00

Model	Video Generation Skill (Avg within each #V-G Group)						Task Completion		Level Score on Video		
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	69.09	55.79	88.94	62.90	37.79	51.46	/	/	/	/	/
VidAgent	52.42	47.73	88.84	63.61	0.00	0.00	30 (65.2%)	0 (0.0%)	25.00	0.00	0.00
LM4LV	0.00	0.00	0.00	0.00	25.90	5.93	8 (17.4%)	0 (0.0%)	6.74	0.00	0.00
NExT-GPT-V1.5	26.78	6.72	130.22	16.03	0.08	0.06	40 (87.0%)	0 (0.0%)	8.34	0.71	0.00
Vitron-V1	36.74	19.32	116.31	25.09	0.08	0.06	40 (87.0%)	0 (0.0%)	18.72	3.04	0.00

Model	Audio Comprehension Skill (Avg within each #A-C Group)									Task Completion		Level Score on Audio		
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	87.27	79.08	70.62	79.00	71.87	62.90	58.70	77.90	78.07	/	/	/	/	/
Qwen-Audio-Chat	56.93	68.77	76.80	37.70	47.71	19.79	56.44	85.15	78.50	30 (100.0%)	6 (25.0%)	28.39	10.57	0.00
Qwen2-Audio-Instru	72.65	74.80	61.40	36.80	45.82	13.45	61.68	78.95	67.99	24 (100.0%)	6 (25.0%)	28.61	8.53	0.00
GAMA	57.00	64.20	68.00	53.20	18.43	26.95	48.85	85.55	61.80	23 (95.8%)	4 (16.7%)	26.35	7.15	0.00
Pengi	52.88	60.07	56.70	36.78	19.77	19.55	42.95	77.40	61.17	23 (95.8%)	1 (4.2%)	23.29	1.74	0.00
SALMONN-13B	67.89	56.33	67.80	29.45	24.67	19.36	43.95	76.55	56.67	23 (95.8%)	2 (8.3%)	23.95	3.61	0.00
WavLLM	64.45	41.07	71.20	30.08	31.30	26.55	45.75	61.40	64.57	24 (100.0%)	2 (8.3%)	23.49	3.28	0.00
NExT-GPT-V1.5	43.23	29.13	65.80	26.70	14.47	25.65	47.95	70.20	69.43	24 (100.0%)	0 (0.0%)	25.05	1.34	0.00
PandaGPT (13B)	41.80	20.23	45.20	20.98	8.47	20.50	42.25	54.80	65.83	24 (100.0%)	0 (0.0%)	16.98	0.65	0.00
ModaVerse-7b-v0	34.10	16.37	32.80	15.20	6.60	8.90	35.05	49.20	60.13	23 (95.8%)	0 (0.0%)	26.10	1.14	0.00
Any-GPT	44.50	32.13	63.40	48.08	16.27	36.40	52.65	67.95	44.63	23 (95.8%)	1 (4.2%)	29.06	3.29	0.00
Unified-io-2-XXL	30.15	27.60	56.10	28.58	15.47	38.35	38.70	63.50	60.63	24 (100.0%)	0 (0.0%)	25.63	1.01	0.00

Model	Audio Generation Skill (Avg within each #A-G Group)											Task Completion		Level Score on Audio		
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	31.50	3.82	3.64	4.68	41.54	51.40	11.52	6.80	8.33	22.88	20.33	/	/	/	/	/
Unified-io-2-XXL	18.36	2.03	5.11	40.52	16.41	24.31	16.97	86.23	94.52	0.25	2.24	17 (85.0%)	0 (0.0%)	25.63	1.01	0.00
Any-GPT	23.50	3.24	4.57	33.58	13.38	14.05	27.49	45.36	83.89	0.25	2.47	17 (85.0%)	1 (5.0%)	29.06	3.29	0.00
NExT-GPT-V1.5	13.60	1.15	4.07	50.51	34.51	1.35	12.36	96.70	99.23	0.25	7.77	17 (85.0%)	1 (5.0%)	25.05	1.34	0.00
AudioGPT	0.50	1.32	4.61	23.10	29.48	0.00	0.00	46.30	79.98	0.25	0.00	13 (65.0%)	1 (5.0%)	8.80	3.02	0.00
SpeechGPT	0.10	2.79	4.44	32.35	0.00	0.00	0.00	30.24	85.54	0.25	0.00	11 (55.0%)	0 (0.0%)	7.22	0.00	0.00
ModaVerse	12.30	1.15	4.29	50.50	28.99	1.05	16.45	100.00	100.00	0.25	4.17	17 (85.0%)	2 (10.0%)	26.10	1.14	0.00

Model	3D Comprehension Skill (Avg within each #D-C Group)													Task Completion		Level Score on 3D		
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	**#10**	**#11**	**#12**	**#13**	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	96.24	98.35	97.78	78.50	70.02	81.20	55.00	88.28	75.20	9.96	68.52	47.14	22.30	/	/	/	/	/
3D-VisTA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.37	0.00	7 (23.3%)	2 (6.7%)	5.41	1.07	0.00
PointLLM-7B	46.16	7.50	72.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8 (26.7%)	0 (0.0%)	6.53	0.00	0.00
PointLLM-13B	48.79	10.00	78.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9 (30.0%)	0 (0.0%)	7.00	0.00	0.00
3D-LLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.34	0.00	7 (23.3%)	1 (3.3%)	5.41	1.38	0.00
AvatarGPT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.70	1 (3.3%)	0 (0.0%)	0.21	0.21	0.00

Model	3D Generation Skill (Avg within each #D-G Group)									Task Completion		Level Score on 3D		
	#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**	**#9**	#Task-Supprt	#Win-Spclst	Level-2	Level-3	Level-4
SoTA Specialist	0.22	7.12E-5	24.42	25.69	78.06	83.64	6540.02	6540.02	0.23	/	/	/	/	/
MotionGPT-T5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.51	1 (4.5%)	0 (0.0%)	0.00	0.00	0.00
MotionGPT-LLaMA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	1 (4.5%)	0 (0.0%)	0.00	0.00	0.00
LLaMA-Mesh	0.00	0.00	0.00	17.55	0.00	0.00	0.00	0.00	0.00	1 (4.5%)	0 (0.0%)	1.60	0.00	0.00

Model	Language Skill (Avg within each #L Group) #1 #12	#2 #13	#3 #14	#4 #15	#5 #16	#6 #17	#7 #18	#8 #19	#9 #20	#10 #21	#11 #22	Task Completion #Supported Task	#Win-over-Specialist	Level Score Level-5
SoTA Specialist	62.62 86.95	86.23 0.31	76.78 94.40	71.00 91.41	58.02 86.05	62.80 86.03	75.11 84.72	77.84 83.67	79.70 58.61	71.91 77.73	28.27 92.38	/	/	/
Meta-Llama-3.1-8B-Instruct	39.75 45.34	56.76 7.95	54.21 76.40	60.52 51.80	20.01 65.90	37.17 41.10	36.23 24.49	29.12 30.70	53.23 8.08	44.49 32.40	14.80 54.35	113 (98.3%)	0 (0.0%)	0.00
ChatGLM-6b	28.97 42.84	33.24 10.91	37.24 41.80	46.10 45.81	19.39 24.50	27.84 16.45	18.85 0.12	35.88 8.41	27.85 2.70	38.51 23.80	13.93 45.37	96 (83.5%)	0 (0.0%)	0.00
Vicuna-7b-v1.5	24.78 43.98	11.18 11.41	33.44 0.00	41.19 0.00	4.51 0.00	13.25 0.96	19.94 0.07	35.27 0.47	54.81 0.00	40.58 23.13	5.06 15.40	72 (62.6%)	0 (0.0%)	0.00
Falcon3-7B-Instruct	36.79 48.15	58.36 5.15	49.91 88.80	56.80 85.89	21.38 45.65	37.12 42.86	32.03 27.64	42.11 34.22	55.79 11.19	42.07 39.80	15.56 58.75	112 (97.4%)	0 (0.0%)	0.00
Ministral-8B-Instruct-2410	41.74 23.39	54.21 11.08	49.53 84.80	51.92 72.60	39.32 56.70	40.49 37.14	13.00 6.28	22.86 31.38	56.87 9.37	43.46 25.53	13.73 40.44	112 (97.4%)	0 (0.0%)	0.00
Yi-Lightning	41.73 52.68	60.54 5.37	55.39 72.60	60.51 56.24	20.53 64.75	39.83 43.59	22.45 28.27	43.57 42.84	62.52 25.34	42.03 29.27	15.29 60.49	113 (98.3%)	0 (0.0%)	0.00
GPT-4V	27.55 44.56	62.40 3.16	34.57 86.20	32.55 83.23	14.43 65.10	27.84 53.82	27.79 54.14	36.07 45.45	65.36 33.86	42.11 26.46	13.96 24.24	113 (98.3%)	0 (0.0%)	0.00
GPT-4o	26.25 46.41	62.57 2.58	33.98 85.40	31.50 86.30	16.20 67.50	26.26 56.10	27.14 57.42	36.64 46.97	66.86 39.52	42.69 32.07	14.49 28.50	113 (98.3%)	0 (0.0%)	0.00
Emu2-32B	32.91 50.15	45.43 9.53	47.04 57.54	39.56 48.78	27.74 43.76	31.24 36.67	39.04 19.84	41.72 24.01	45.48 13.78	46.35 26.47	13.05 31.72	113 (98.3%)	0 (0.0%)	0.00
DeepSeek-VL-7B	29.97 79.68	44.39 83.00	55.55 62.20	20.36 50.60	40.49 62.30	57.93 46.87	49.85 4.12	48.73 28.46	27.03 8.11	56.76 31.80	10.37 40.97	114 (99.1%)	0 (0.0%)	0.00
Qwen2-VL-7B	23.91 37.23	27.51 6.48	37.68 64.00	46.40 37.00	17.84 3.50	20.96 20.50	36.25 0.24	29.29 4.87	35.42 6.00	35.58 20.87	12.62 21.79	94 (81.7%)	0 (0.0%)	0.00
LLaVA-One-Vision-72B	50.44 43.81	41.98 3.55	54.55 84.80	61.13 10.43	29.87 59.35	56.99 34.91	35.24 42.94	43.27 28.63	55.23 19.26	41.49 52.20	17.73 71.95	110 (95.7%)	0 (0.0%)	0.00
InternVL2.5-8B	42.93 71.96	47.76 75.20	59.54 55.40	31.17 68.40	42.86 56.75	32.72 55.60	50.98 22.12	43.02 36.48	30.85 9.80	51.23 32.13	9.07 53.67	114 (99.1%)	0 (0.0%)	0.00

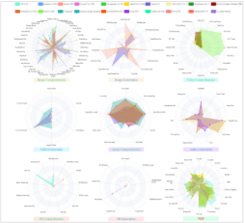
Path to Multimodal Generalist: General-Bench

What General-Bench Unveils? —— Quantitative Performances

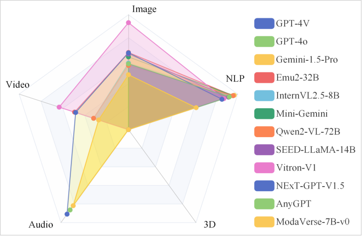
- *Observation-1: Lack of task support.*
- *Observation-2: Few generalists surpass the SoTA specialist.*
- *Observation-3: Focus more on content comprehension than supporting generation.*
- *Observation-4: Insufficient support for all modalities.*
- *Observation-5: Multimodality does NOT really enhance language.*

Path to Multimodal Generalist: General-Bench

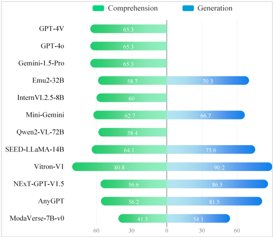
What General-Bench Unveils? —— In-depth Analysis



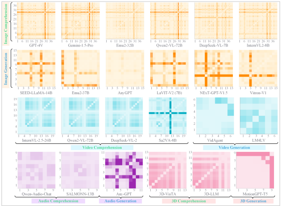
Task Supporting



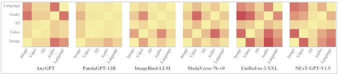
Modality Supporting



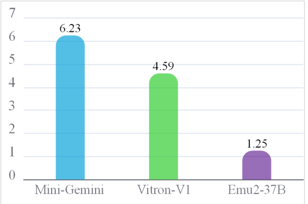
Capabilities on Comprehension vs. Generation



Synergy Across Skills



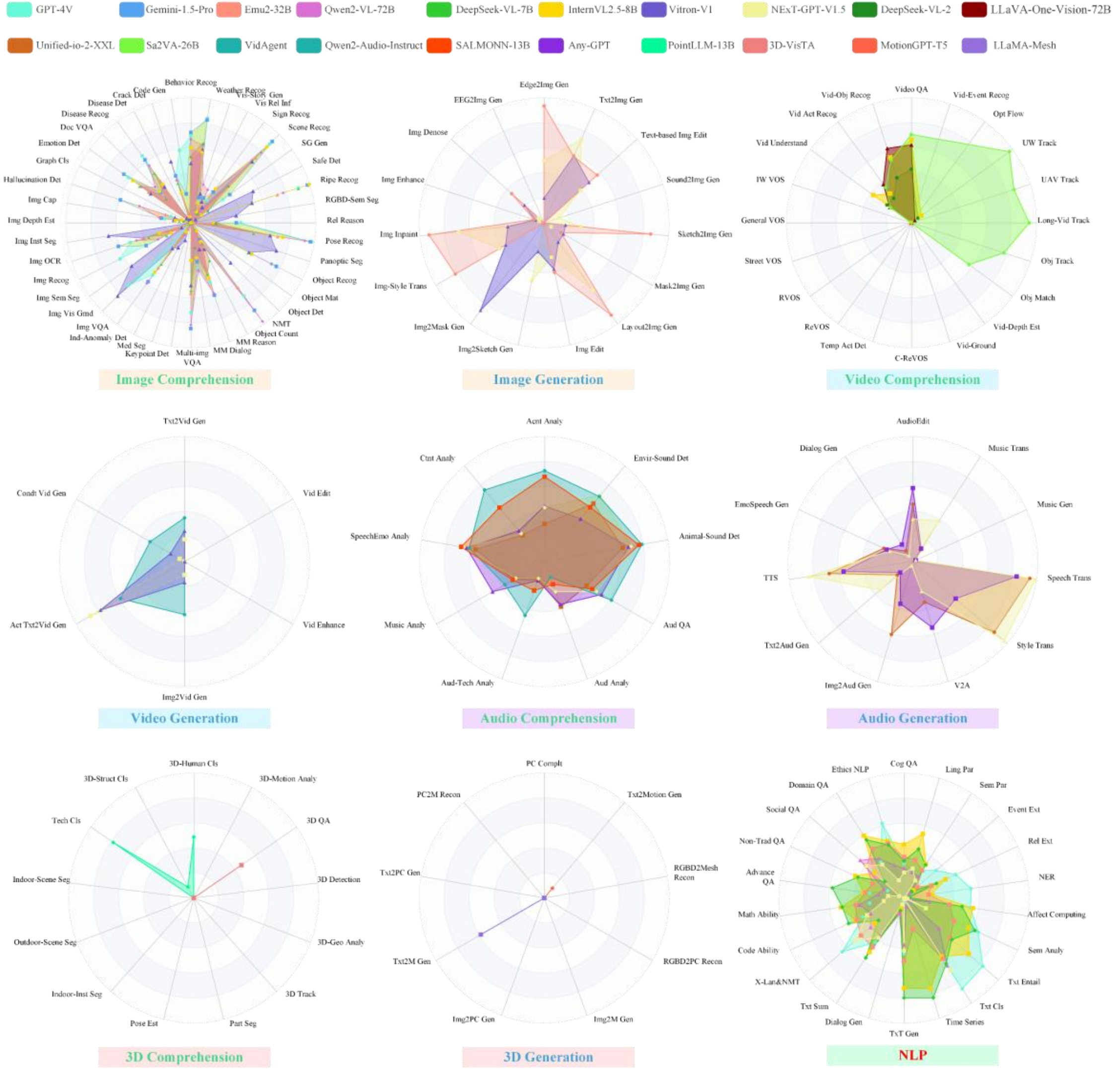
Synergy Across Modalities



Synergy Across Comprehension and Generation

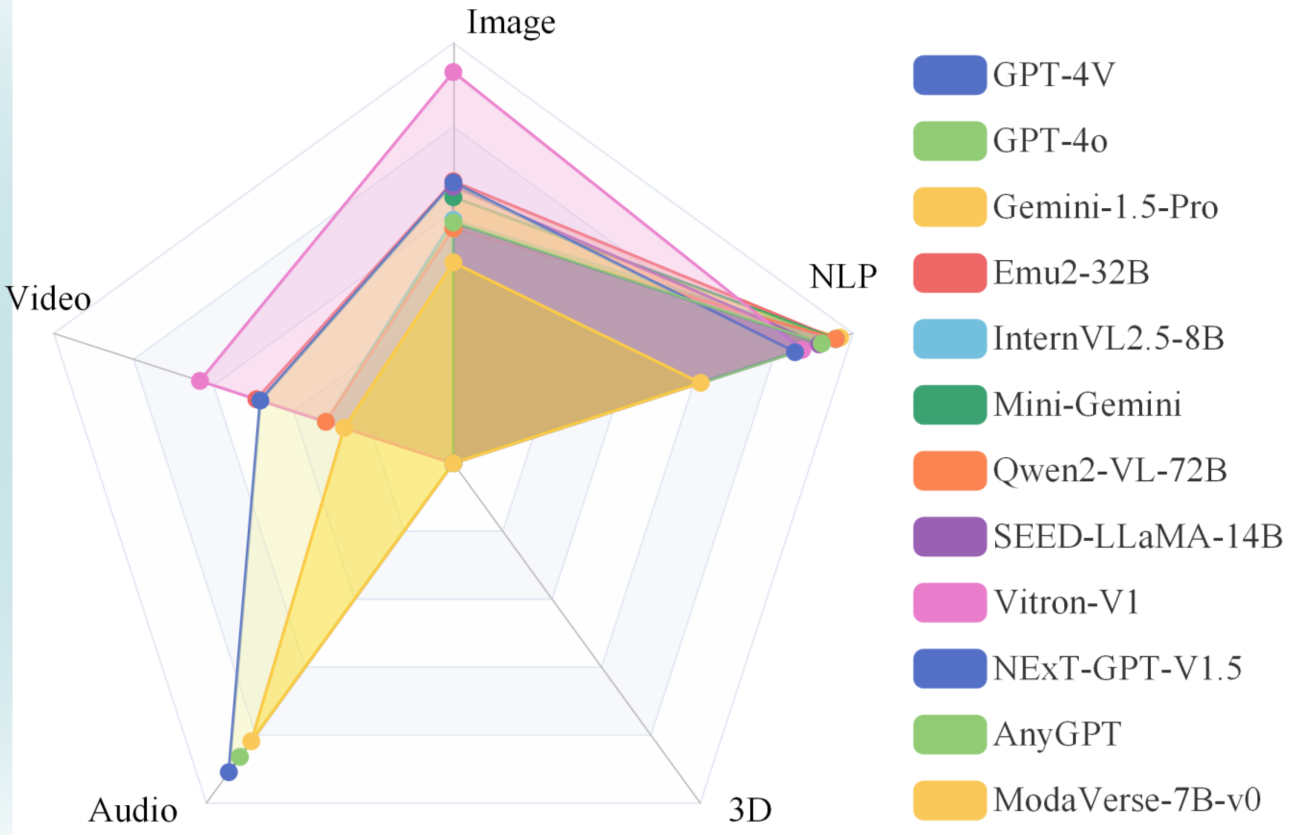
Task Supporting

Current MLLMs generally exhibit limited task Support, with a Strong bias toward Simpler Comprehension tasks and significant challenges in covering diverse and complex generation skills across modalities.



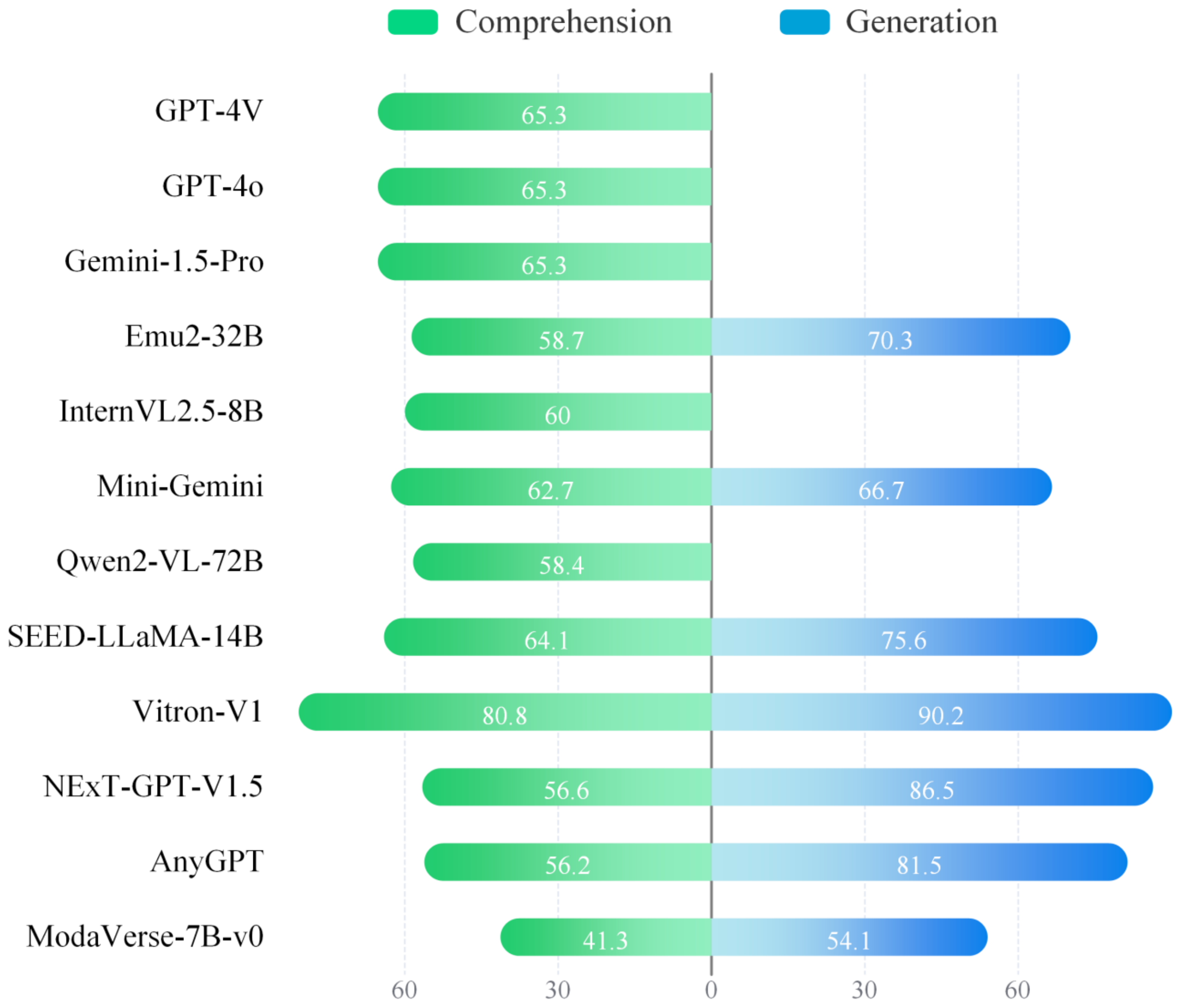
Modality Supporting

Most MLLMs support Only a Single Non-language Modality, while only a few-like NExT-GPT-1.5 or Unified-IO-2 demonstrate truly broad, all-modality capabilities.



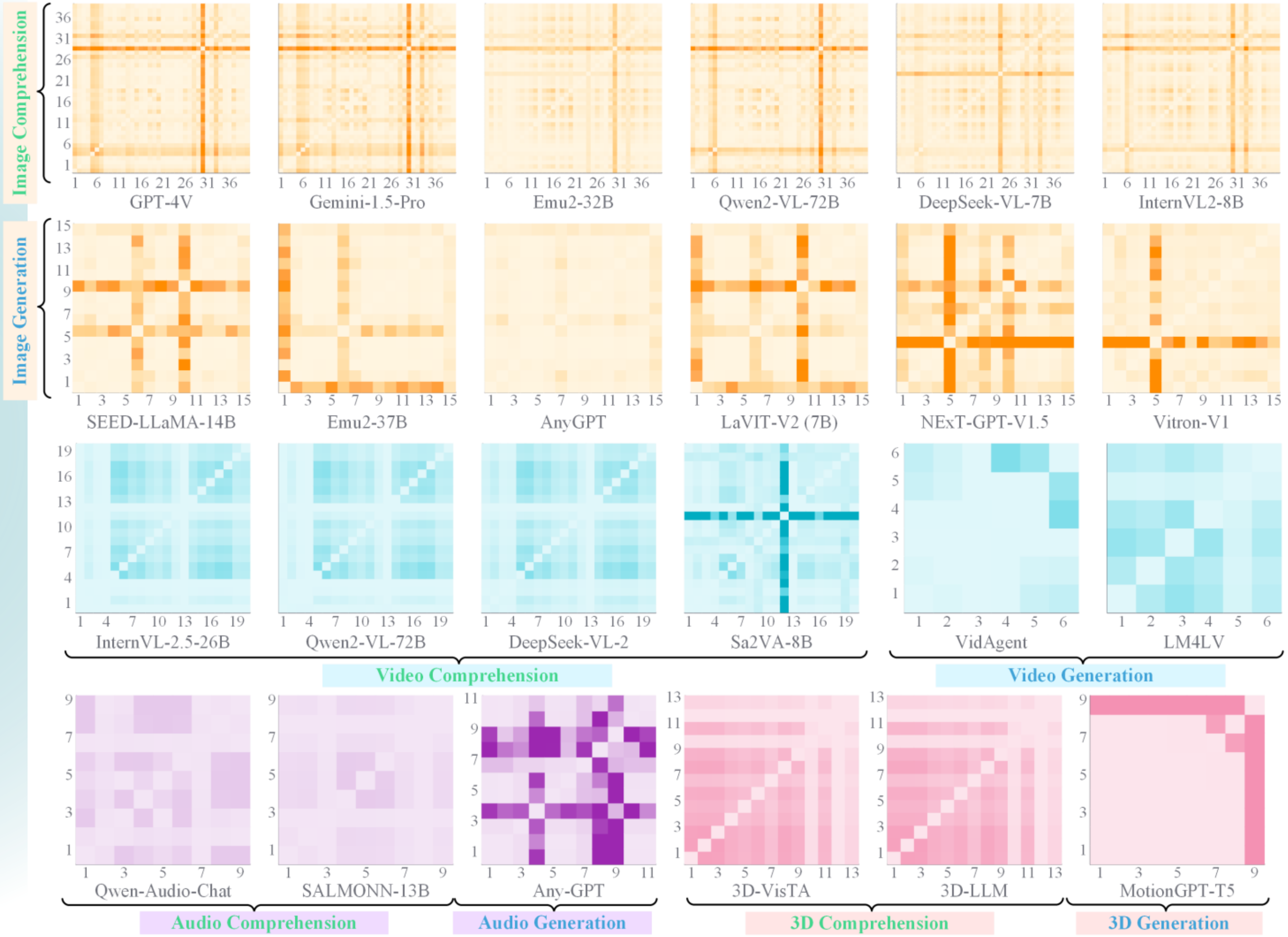
Comprehension vs. Generation

Most MLLMs are Stronger at Comprehension than generation, due to the greater complexity and training cost of generation; only a few models, like Vitron-V1, demonstrate balanced capabilities across both paradigms.



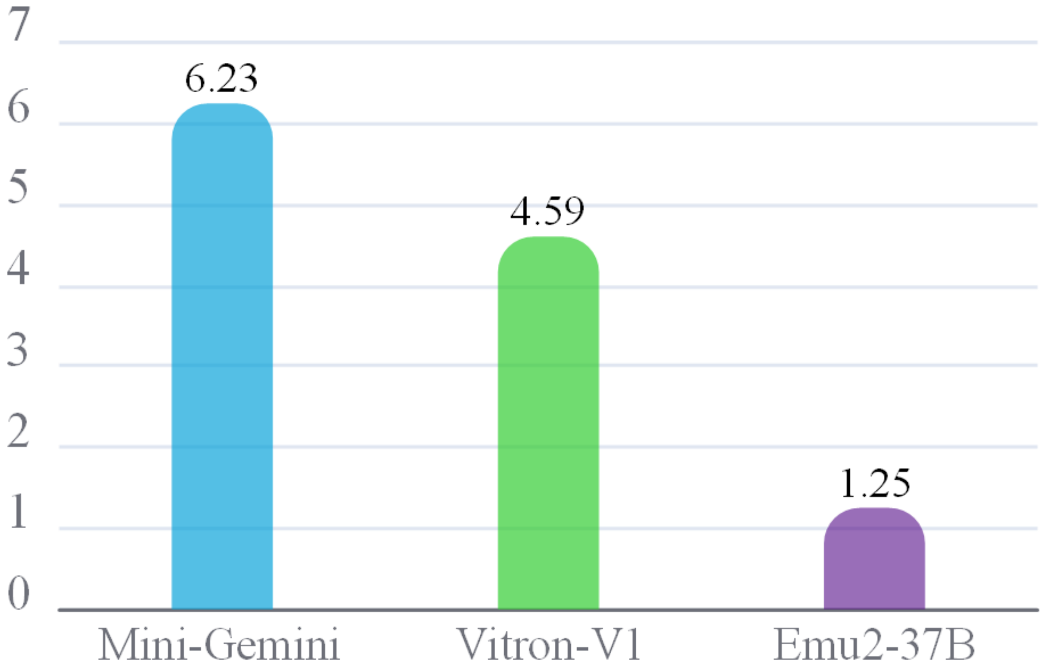
Synergy Across Skills

Synergy effects in MLLMs are **uneven across skills**, with stronger synergy observed ingeneration tasks and among closely related skills, particularly in models with higher Level3 scores.



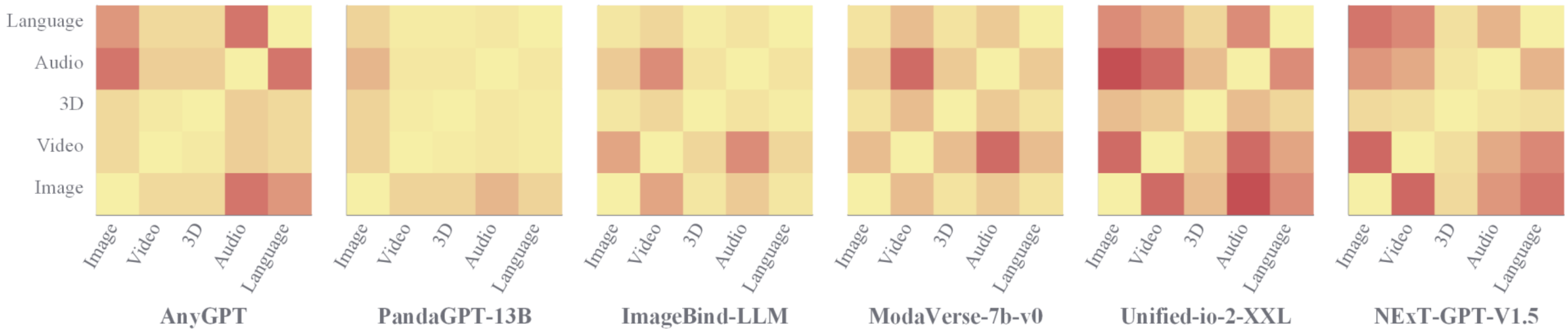
Synergy Across Comprehension & Generation

Only a few MLLMs exhibit synergy between comprehension and generation,
with Mini-Gemini showing the strongest effect-mainly within the image modality.



Synergy Across Modalities

Synergy is **strongest between image and video modalities**, while language shows only one-way synergy toward other modalities; no modalities really-significantly enhance language tasks-highlighting a key limitation of current MLLMs.



Path to Multimodal Generalist: General-Bench

How to use General-Bench?

General-Level Open Set

With inputs and labels of samples all publicly open, for open-world use (e.g. academic experiment).

General-Level Close Set

With only sample inputs available, which participants can use for ranking in our leaderboard.

Path to Multimodal Generalist: General-Bench

- How to participate the Leaderboard?

General Bench

Four-scoped leaderboard

Scope-A: *Full-spectrum Hero*

Difficulty: ☆☆☆☆☆☆

Number of leaderboards: ☆

Details:

- ✔ Covers all General-Level tasks and modalities.
- ✔ Most challenging track; requires high model capacity and resource commitment.

Highlights:

- ✔ Evaluates holistic generalization and cross-modal synergy.
- ✔ Suitable for near-AGI or foundation-level multimodal generalists.

General Bench

Four-scoped leaderboard

Scope-A: *Full-spectrum Hero*

Scope-B: *Modality-specific Unified Hero*

Difficulty: ☆☆☆☆

Number of leaderboards: ☆☆☆

Details:

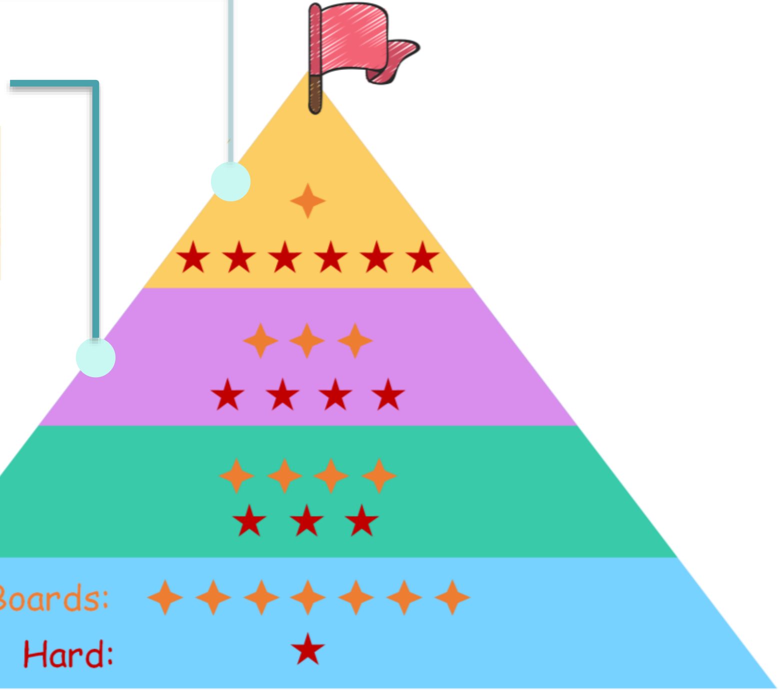
✔ 7 separate leaderboards (4 single modality + 3 combined modality).

✔ Focuses on mastering diverse tasks within a single modality.

Highlights:

✔ Measures within-modality generalization.

✔ Suited for intermediate-level models with cross-task transferability.



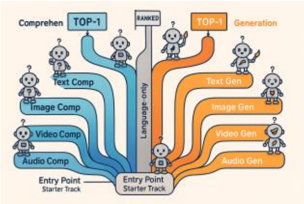
General Bench

Scope-A: *Full-spectrum Hero*

Scope-B: *Modality-specific Unified Hero*

Scope-C: *Comprehension/Generation Hero*

Four-scoped leaderboard



Difficulty: ☆☆☆

Number of leaderboards: ☆☆☆☆

Details:

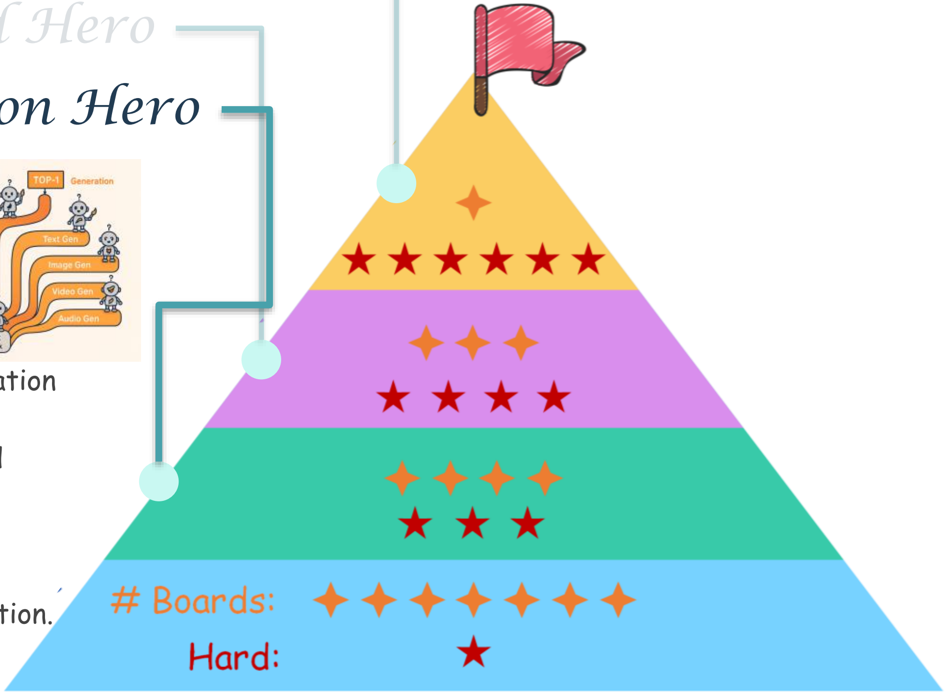
✔ 8 leaderboards: 2 × 4 for multimodal comprehension/generation under different modalities.

✔ Supports entry-level model evaluation or teams with limited resources.

Highlights:

✔ Assesses task-type specialization: understanding or generation.

✔ Reflects generalization across task types.



General Bench

Four-scoped leaderboard

Scope-A: *Full-spectrum Hero*

Scope-B: *Modality-specific Unified Hero*

Scope-C: *Comprehension/Generation Hero*

Scope-D: *Skill-specific Hero*

Difficulty: ☆☆☆

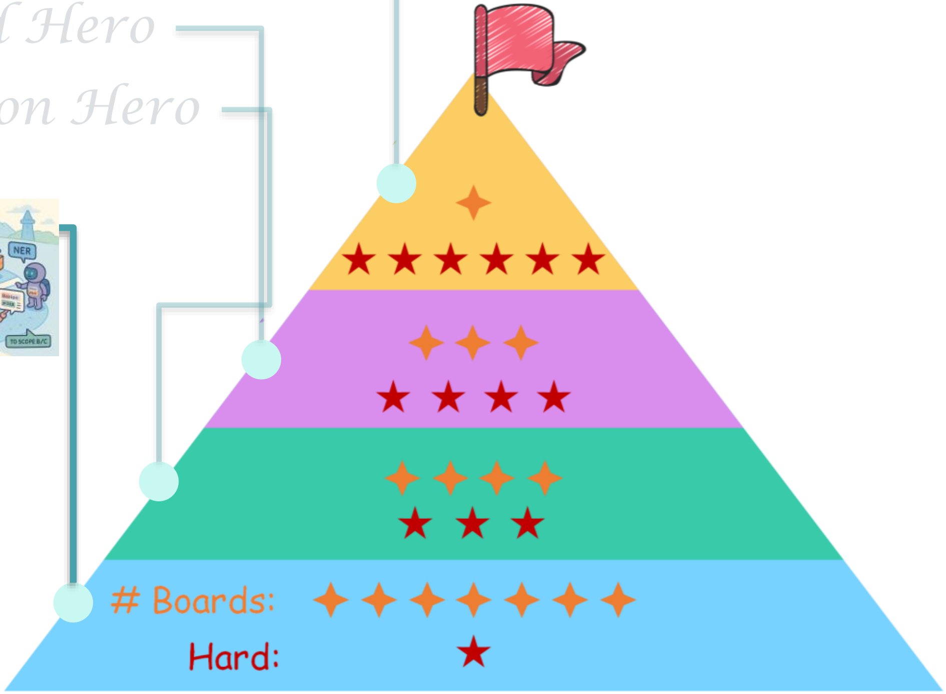
Number of leaderboards: ☆☆☆☆☆☆☆

Details:

- ✔ Large number of sub-leaderboards, each scoped to a skill set
- ✔ Easiest to participate; lowest cost.

Highlights:

- ✔ Evaluates fine-grained skill performance.
- ✔ Helps identify model strengths and specialization areas.



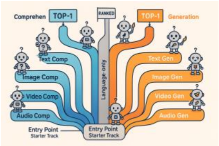
General Bench

Four-scoped leaderboard

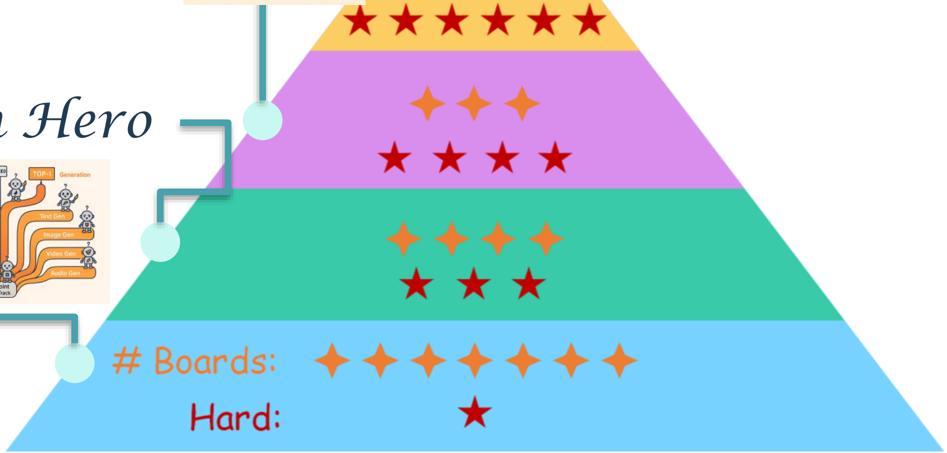
Scope-A: *Full-spectrum Hero*

Scope-B: *Modality-specific Unified Hero*

Scope-C: *Comprehension/Generation Hero*



Scope-D: *Skill-specific Hero*



Model Diagnostics

Model Diagnostics 🕵️

In this page, we present a comprehensive diagnostic analysis of multimodal generalist models that are included in our General-Bench 🧠 leaderboard. Built upon an exceptionally large-scale, multi-dimensional 📊 evaluation benchmark, General-Bench enables broad and in-depth assessment across diverse modalities, tasks, and paradigms 🔄.

While leaderboard rankings 🏅 offer a high-level view of overall performance, they often mask the nuanced strengths and weaknesses exhibited by each model across different dimensions. To bridge this gap, our Model Diagnostics aims to unpack these subtleties—identifying where each model excels 💪 and where it struggles ⚠️ across modalities, capabilities, and task types.

We believe such fine-grained diagnostics are essential for guiding the future development of stronger and more robust multimodal models 🚀. We believe this effort plays a critical role in advancing the field toward truly universal multimodal generalists—and ultimately, Artificial General Intelligence (AGI) 🤖.

Submission and Contribution

Welcome to submit your Multimodal Generalist to General-Level leaderboard, or contribute your dateset to General-Bench to maximize the visibility.

🛎 Guidelines for Submitting Model to Leaderboard

- Please first download the corresponding `close-set data` for your selected leaderboard (based on its unique identifier).
- You are also encouraged to download the `open-set data` for model debugging and development purposes.
- Based on the close-set data, conduct inference using your model, and save the output results into a single `[model]-[leaderboard-id].zip` file.
- In the following submission process, in addition to uploading the evaluation result file, please fill in the following `required information fields` to help us properly process your submission on the backend.
- Please refer to the `documentation` for more detailed instructions.
- To ensure fairness of the evaluations, General-Level have implemented the following restrictions:
    1. A maximum of submitting 2 results past 24 hours (excluding exceptions);
    2. A maximum of submitting 4 results past 7 days (excluding exceptions);
    3. Before the evaluation of the latest submission finished (evaluation results / error logs generated), users are not allowed to start a new submission.

Submit to Leaderboard | **Contribute to General-Bench**

Click or drag file here to upload an evaluation result file (.zip)

Current file section status: no file selected

Submission and Contribution

Welcome to submit your Multimodal Generalist to General-Level leaderboard, or contribute your dateset to General-Bench to maximize the visibility.

📍 Guidelines for Contributing Data to General-Bench

- **General-Bench is open and non-commercial**. A key feature of this project for evaluating multimodal generalist models is the need for broad coverage—including diverse modalities, tasks, paradigms, domains, and capabilities. We greatly appreciate your contributions of new data and tasks 😘, which will also benefit the whole community. Once your data is included in General-Bench, your contribution will be **acknowledged on the website homepage** to increase its visibility, and it will also be **cited** in our technical paper.
- We especially welcome datasets that feature 1) **highly challenging** tasks, or (2) task definitions involving **multiple modalities** simultaneously.
- Please fill in the required information fields. Refer to the `documentation` for detailed instructions. This includes:
  1. The name of the dataset (or task), the number of instances (including Open/Close set split);
  2. The task's modality, paradigm, domain, and targeted evaluation capabilities;
  3. A description of the evaluation methodology used for the task.
- Please submit your data as a single `[data-name].zip` file, together with an evaluation manual (might be `txt`, `doc`, `md` etc., all zipped in `[data-name]-[eval-instruction].zip`).

Submit to Leaderboard | Contribute to General-Bench

Click or drag file to this area to upload the dataset file (.zip)

Current file section status: no file selected

Click or drag file to this area to upload the data instruction file (.zip)

Current file section status: no file selected

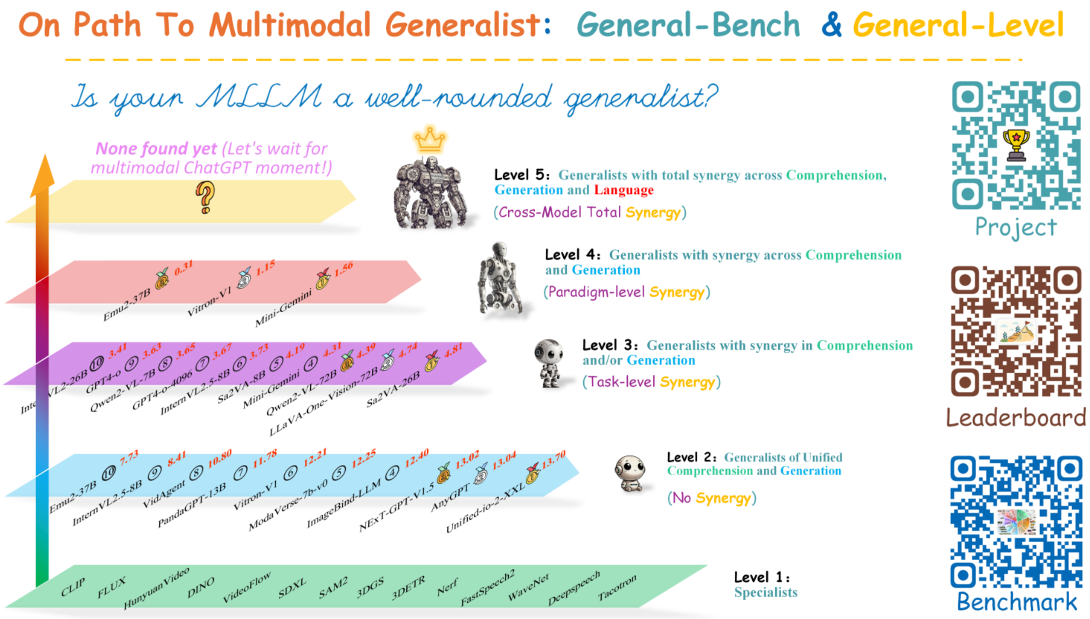
Path to Multimodal Generalist: General-Bench

On Path to Multimodal Generalist: General-Level and General-Bench

Project: https://generalist.top/

Paper: https://arxiv.org/abs/2505.04620

Benchmark: https://generalist.top/leaderboard



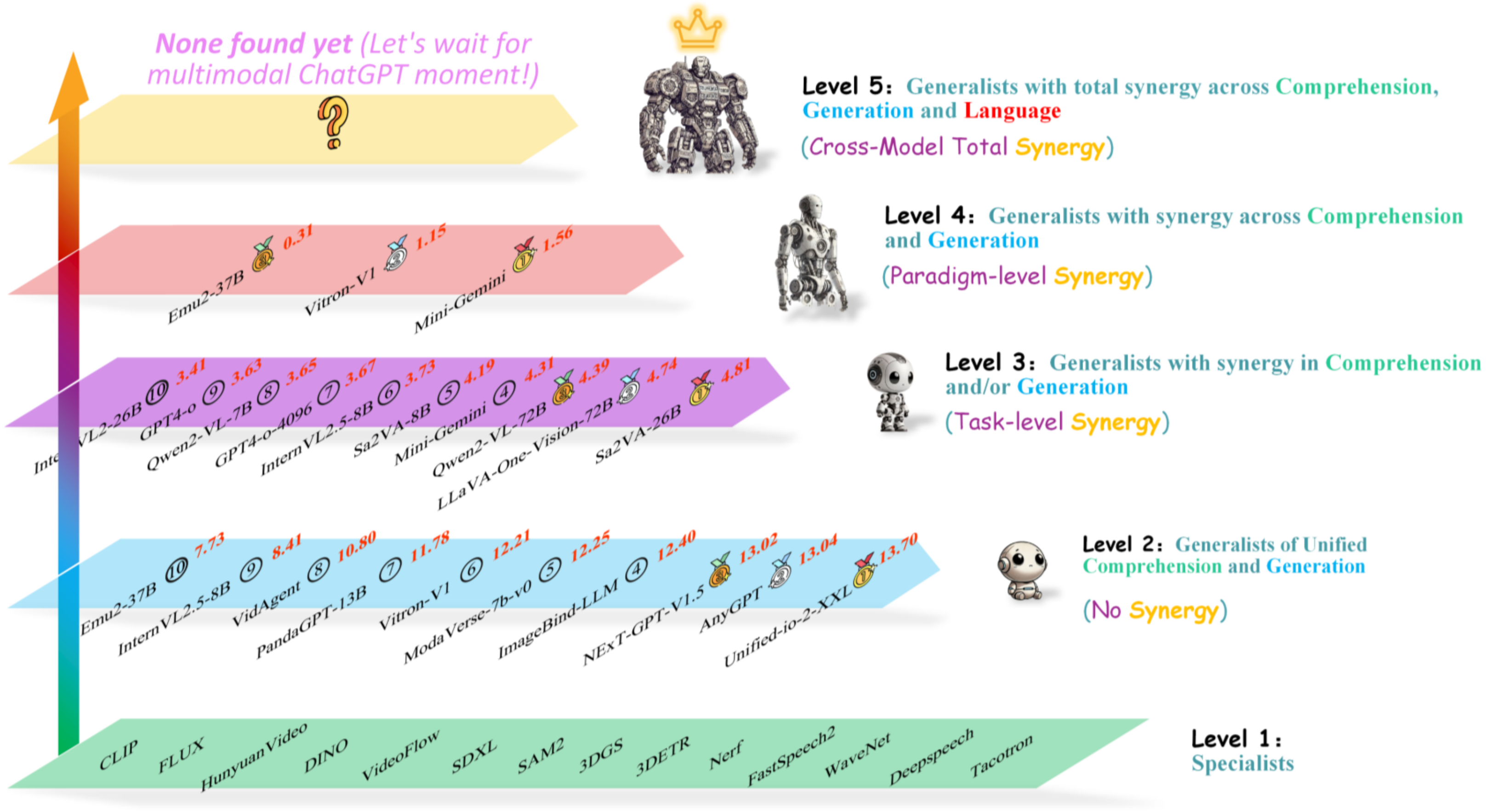
- Hao Fei, Yuan Zhou, ⋯, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, Hanwang Zhang. "On Path to Multimodal Generalist: General-Level and General-Bench". ICML. 2025

Table of Content

Path to Multimodal Generalist: What's Next

- Improving from **Generalist Model** perspective

Path to Multimodal Generalist: What's Next



Path to Multimodal Generalist: What's Next

Goals to Next-generation **Multimodal Generalist**

- **Multimodality** ★★★★★☆

  supporting diverse modalities and tasks, enabling models to seamlessly process and reason across language, vision, audio, and more—much like human cognition

- **Unification** ★★★☆☆☆

  integrating both perception and generation capabilities into a single architecture

- **Advancement** ★★★☆☆☆

  enabling higher-order functionalities with advanced capability, such as fine-grained advanced reasoning in complex contexts

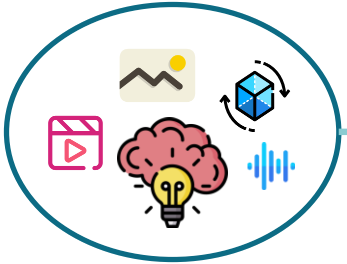
- **Generalizability** ☆☆☆☆☆☆

  achieving cross-modality and cross-task generalization, where knowledge learned in one modality or task can transfer to others

Path to Multimodal Generalist: What’s Next

Angle-I: Multimodal Generalists with in-depth **Modality**&**Task** Unification

- Enhance **breadth** capability.



MLLM/
Multimodal Generalist

Breadth: More Unified

- ✓ Supporting More Tasks
- ✓ Supporting More Modalities

Path to Multimodal Generalist: What's Next

<table>
<tr><th></th><th colspan="4">Modality (w/ Language)</th></tr>
<tr><th></th><th>Image</th><th>Video</th><th>Audio</th><th>3D</th></tr>
<tr><td rowspan="4">Input-side Perceiving</td><td>Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, …</td><td>VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, …</td><td>AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, …</td><td>3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, …</td></tr>
<tr><td>[Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, …</td><td>[Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, …</td><td>–</td><td>–</td></tr>
<tr><td colspan="2">Video-LLaVA, Chat-UniVi, LLaMA-VID</td><td>–</td><td>–</td></tr>
<tr><td colspan="3">Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, …</td><td>–</td></tr>
<tr><td rowspan="3">Perceiving + Generating</td><td>GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, …</td><td>GPT4Video, Video-LaVIT, VideoPoet, …</td><td>AudioGPT, SpeechGPT, VIOLA, AudioPaLM, …</td><td>–</td></tr>
<tr><td colspan="2">[Pixel-wise] Vitron</td><td>–</td><td>–</td></tr>
<tr><td colspan="3">NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, …</td><td>–</td></tr>
</table>

Path to Multimodal Generalist: What’s Next

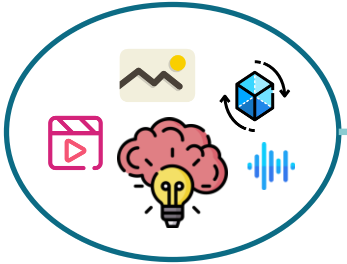
Angle-I: Multimodal Generalists with in-depth **Modality**&**Task** Unification

Image Comprehension	Classification & Recognition; Detection; Segmentation; VQA; Captioning; Spatial Reasoning; Image-to-X Generation; OCR; Matching; Multimodal Translation
Image Generation	Text-to-Image Generation; Conditioned Image Generation; Editing; Enhancement
Video Comprehension	Video Segmentation & Tracking & Grounding; Video QA; Video Recognition; Action Understanding; Depth Estimation; Flow Estimation; Object Matching
Video Generation	Text-to-Video Generation; Conditioned Video Generation; Video Enhancement; Video Editing
3D Comprehension	Classification; Segmentation; QA & Caption; 3D Estimation
3D Generation	Point Cloud Generation; Mesh Generation; Motion Generation
Audio Comprehension	Audio & Sound Understanding; Speech Understanding; Music Understanding
Audio Generation	Speech Generation; Conditioned Audio Generation; Audio Editing; Music Generation
Language	Reasoning Ability; Math Ability; Code Ability; Social QA; Science QA; Cross-lingual Ability; Text Generation; Text Classification; Affective Computing; Information Extraction; Linguistic & Semantic Parsing; Time Series

Path to Multimodal Generalist: What's Next

Angle-II: Unified Comprehension & Generation

- Further enhance **breadth** capability.



MLLM/
Multimodal Generalist

Breadth: More Unified

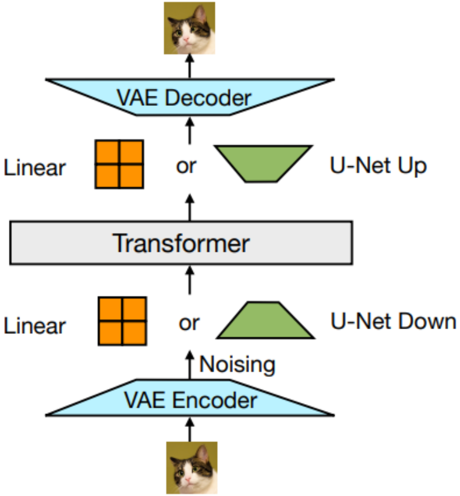
- Supporting More Tasks
- Supporting More Modalities
- Supporting More Paradigm

Path to Multimodal Generalist: What's Next

Angle-II: Unified Comprehension & Generation

What is the optimal model architecture under unified MLLM?

- Pipeline Agent
- Joint Encoder+LLM+Diffusion
- Joint LLM^{AR} Tokenization (VQ-VAE)
- Joint LLM^{AR}+Diffusion



- Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, Kaiming He. Autoregressive Image Generation without Vector Quantization. 2024.
- Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. 2024.
- Zhou, Chunting, et al. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. 2024.

Path to Multimodal Generalist: What’s Next

Angle-II: Unified Comprehension & Generation

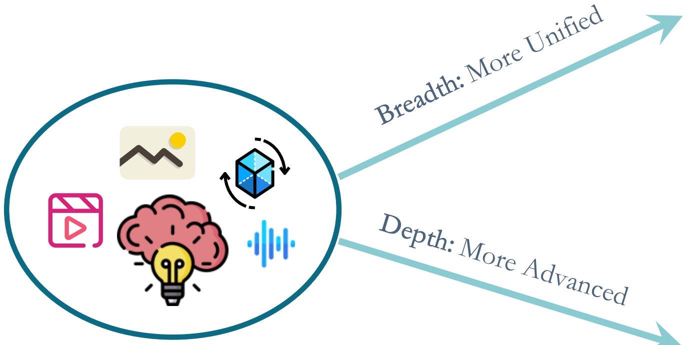
Still much room to explore

- *Generation hurt comprehension? Can both two enhance others?*
- *How to obtain better tokenizer? How to handle Video tokenizer?*
- *How far to beat SoTA specialist?*
- *What’s the best architecture for other modalities?*
- …

Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

- Further enhance capabilities both in **breadth** and **depth**.



MLLM/
Multimodal Generalist

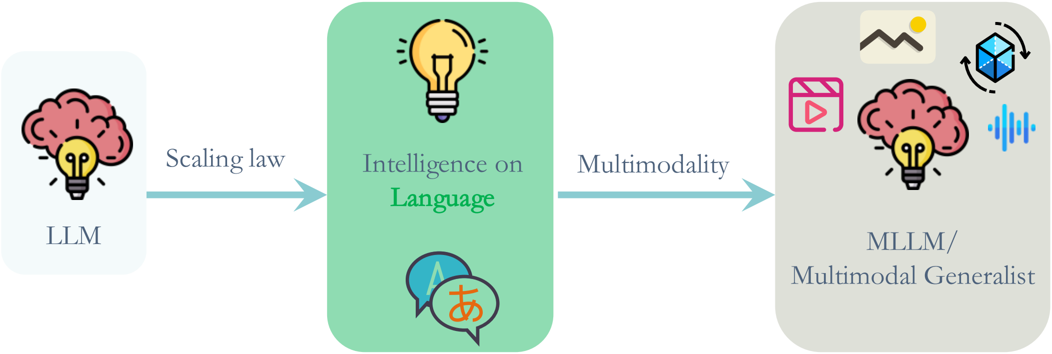
- Supporting More Tasks
- Supporting More Modalities
- Supporting More Paradigm

- Human-level Reasoning
- Synergy between Comp&Gen
- Cross-modal/Cross-task Generalizability

Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

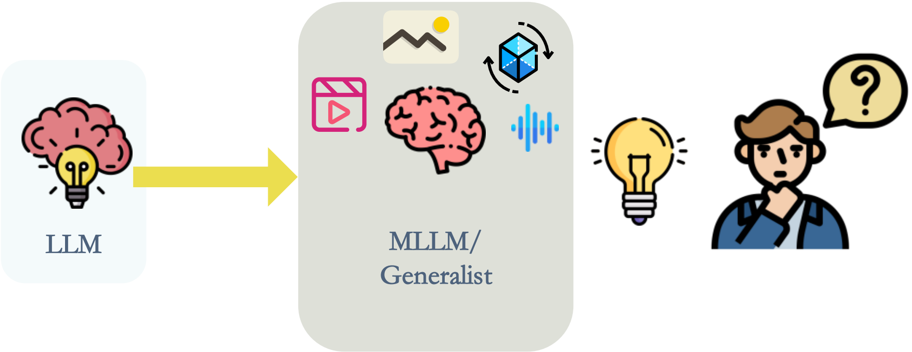
The language intelligence of LLMs empowers multimodal intelligence.



Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

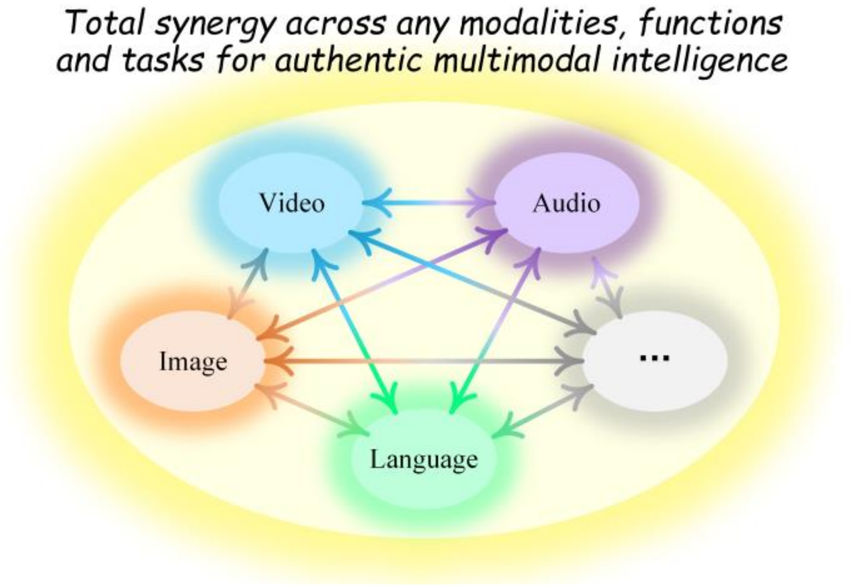
- Could the **scaling law** and **emergence** success of LLMs be replicated in multimodality to achieve the intelligence of native MLLMs?



Path to Multimodal Generalist: What’s Next

Angle-III: Native Multimodal Intelligence

- Ideal intelligent pattern in multimodal generalist



Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

Still much room to explore

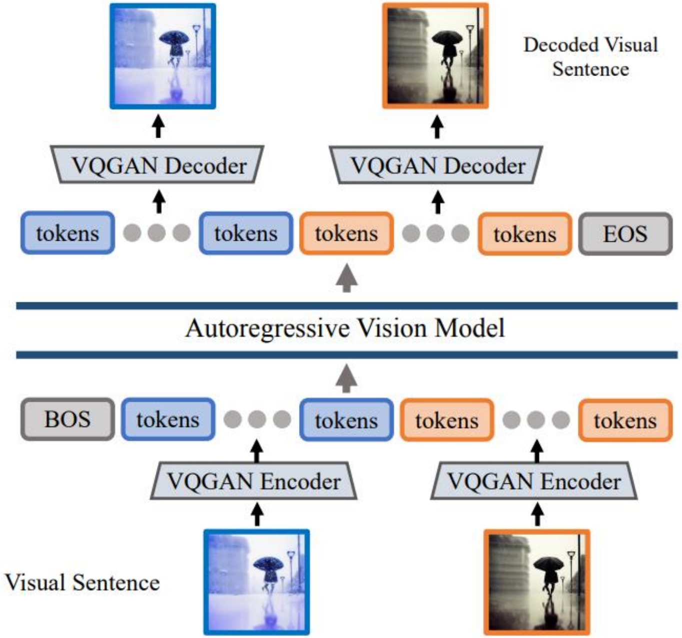
- *Architecture*
- *Data Scale*
- *Training/Learning*
- …

Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

➢ Large Vision Model (LVM)

- mimicking LLM pretraining
- next visual token prediction



- Yutong Bai, Xinyang Geng, Karttikeya Mangalam etc. Sequential Modeling Enables Scalable Learning for Large Vision Models. CVPR. 2024

Angle-III: Native Multimodal Intelligence

What scale of dataset is required for pre-training from scratch?

Modality	LLM/MLLM	Amount
Language	Chat-GPT4	13 Trillion text tokens
Vision	LVM	420 Billion visual tokens
Multimodalities	Unified-IO 2	1 Trillion text tokens, 1 Billion image-text pairs, 180 Million video clips, 130 Million interleaved image & text, 3 Million 3D assets, 1 Million agent trajectories

Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

- Training/Learning
  - *Synergistic Training*

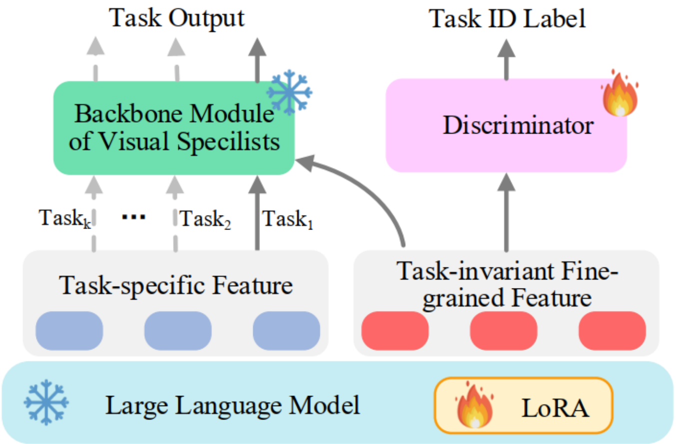


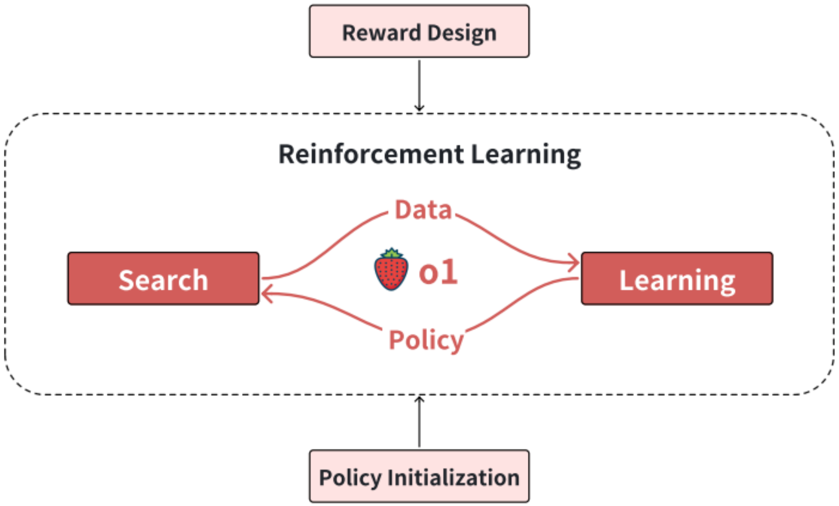
Figure 3: Illustration of the synergy module.

- **Hao Fei**, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, Shuicheng Yan. "VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing". **NeurIPS**. 2024

Path to Multimodal Generalist: What's Next

Angle-III: Native Multimodal Intelligence

- Training/Learning
  - *R1/O1 for interleaved multimodality?*
  - *RL Scaling*



- Zeng, etc. "Scaling of Search and Learning: A Roadmap to Reproduce o1 from Reinforcement Learning Perspective". Arixv. 2024

Path to Multimodal Generalist: What’s Next

- Improving from **Evaluation Framework** perspective

Path to Multimodal Generalist: What's Next

Angle-I: Further refinement of the General-Level framework

- *The synergy measurement is simplified by assuming performance beyond SoTA specialists implies synergy, avoiding direct modeling.*

- *There is room for improving algorithmic design to better reflect true multimodal coordination and synergy.*

Path to Multimodal Generalist: What's Next

Angle-II: Expanding the General-Bench

- *Expanding to cover more comprehensive tasks and modalities for fair and complete evaluation.*

- *Imbalance exists — image tasks dominate, while audio and 3D modalities are underrepresented.*

- *True multimodal generalists should handle modality-switching and interleaved reasoning.*

- *Incorporate tasks that involve multi-turn, cross-modal interactions for both comprehension and generation.*

Path to Multimodal Generalist: What's Next

Angle-III: Rethinking Evaluation Paradigm for Model Capabilities

- *Many current evaluation still follow traditional paradigms*
  - *work well for simple tasks (e.g., multiple-choice, classification)*
  - *but fail on format-free multimodal generation tasks, metrics like **FID/FVD** are increasingly viewed as inadequate for evaluating video or 3D generation quality.*

- *There is a growing reliance on human evaluation, but it lacks scalability.*
  - *use LLMs as judges, but face challenges in evaluation stability and reproducibility.*
  - *adopts a single metric per task, which may introduce bias; should incorporate multiple complementary metrics for more holistic assessment.*

- *Should also assess interpretability and reasoning traceability.*

Thanks!

Any questions?

Reaching out to me:
Haofei37@nus.edu.sg