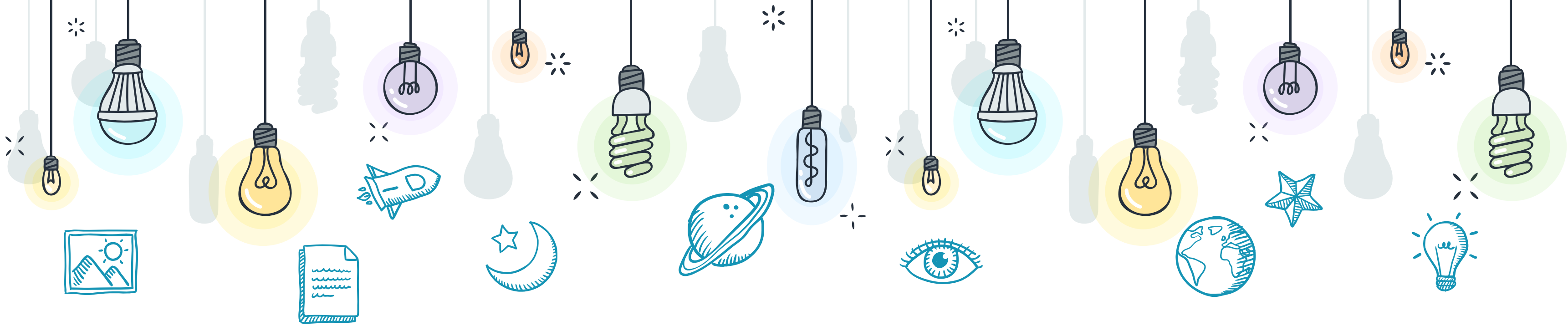


From Multimodal LLM to Human-level AI

Architecture, ***Modality***, ***Function***, ***Instruction***, ***Hallucination***, **Evaluation**, ***Reasoning*** **and Beyond**

https://mllm2024.github.io/ACM-MM2024/

ACM Multimedia 2024

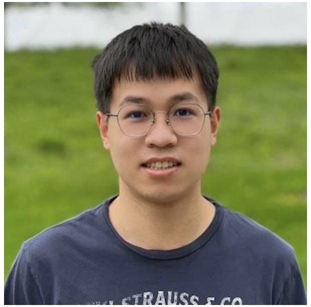
Melbourne, Australia

Hao Fei

National University of Singapore

Xiangtai Li

ByteDance/Tiktok

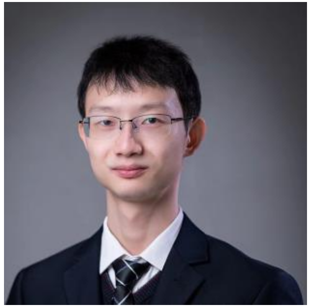


Haotian Liu

xAI

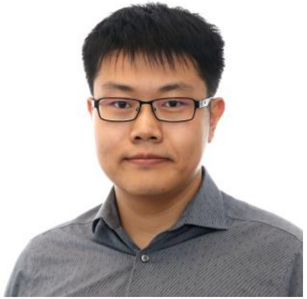
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

Nanyang Technological University

Kaipeng Zhang

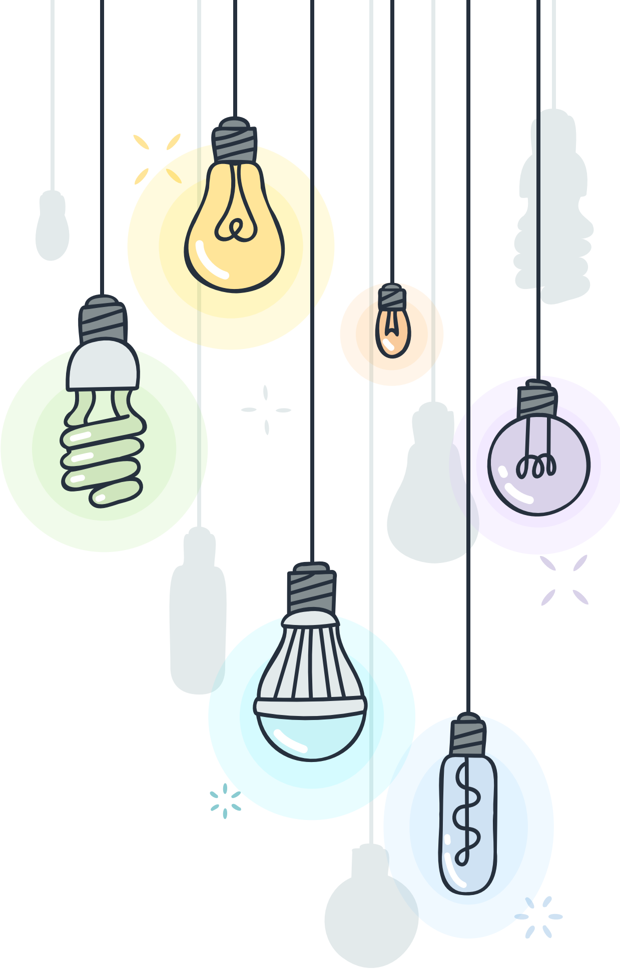
Shanghai AI Lab

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-VI

Evaluation & Generalist
Road to L5 MM Generalist



Hanwang Zhang
Assoc. Prof
Nanyang Technological University
hanwangzhang@ntu.edu.sg
https://mreallab.github.io/

Road to L5 MM Generalist

https://path2generalist.github.io

Hanwang Zhang 张含望

Hao Fei
NUS

Yuan Zhou
NTU

Juncheng Li
ZJU

Kaihang Pan
ZJU

Zhongqi Yue
NTU

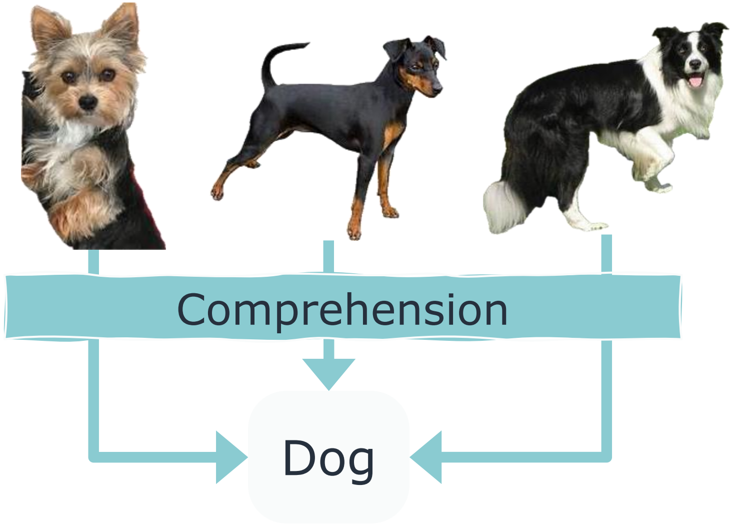
Shuicheng Yan
Skywork AI

Comprehension & Generation

(1) Comprehension is a "**many-to-one**” paradigm

Example:

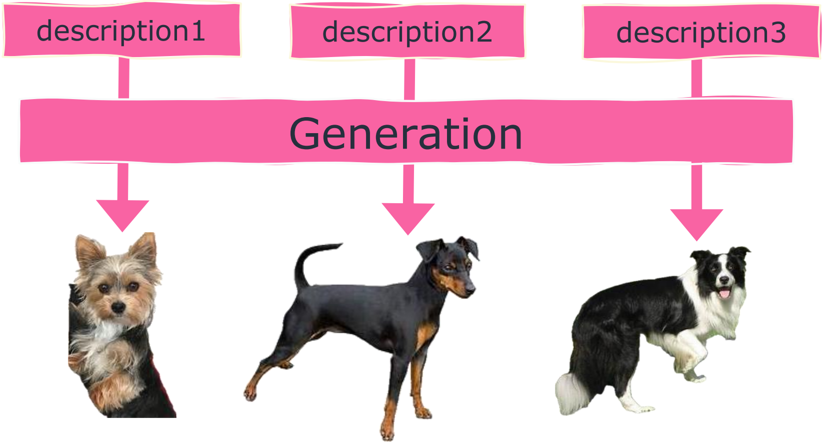
What is it?



(2) Generation is an ”**One-to-one**” paradigm

Example:

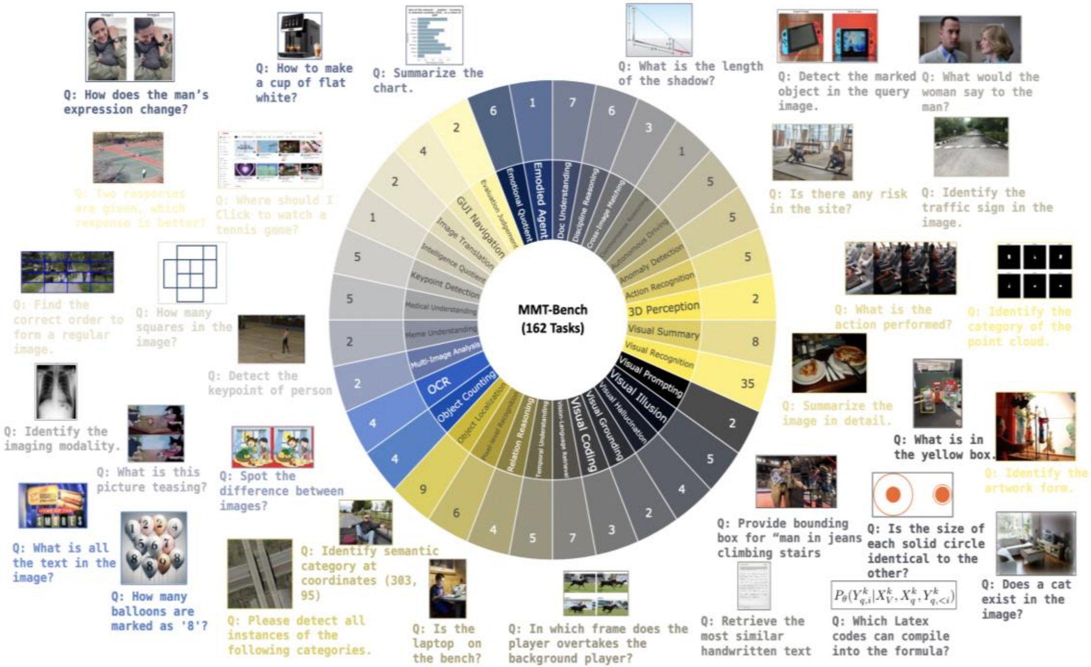
Generate a dog according to descriptions!



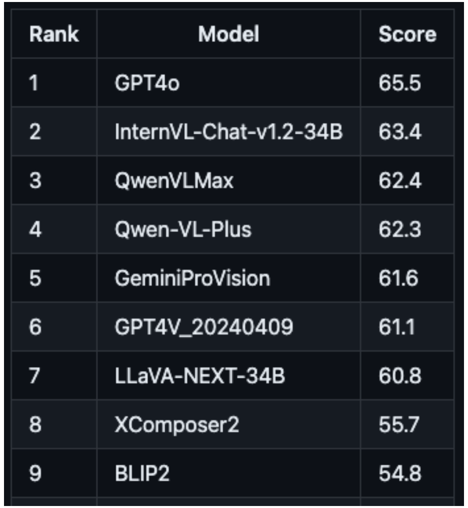
*Inputs and outputs are matched one-by-one.

Differences between Comprehension and Generation

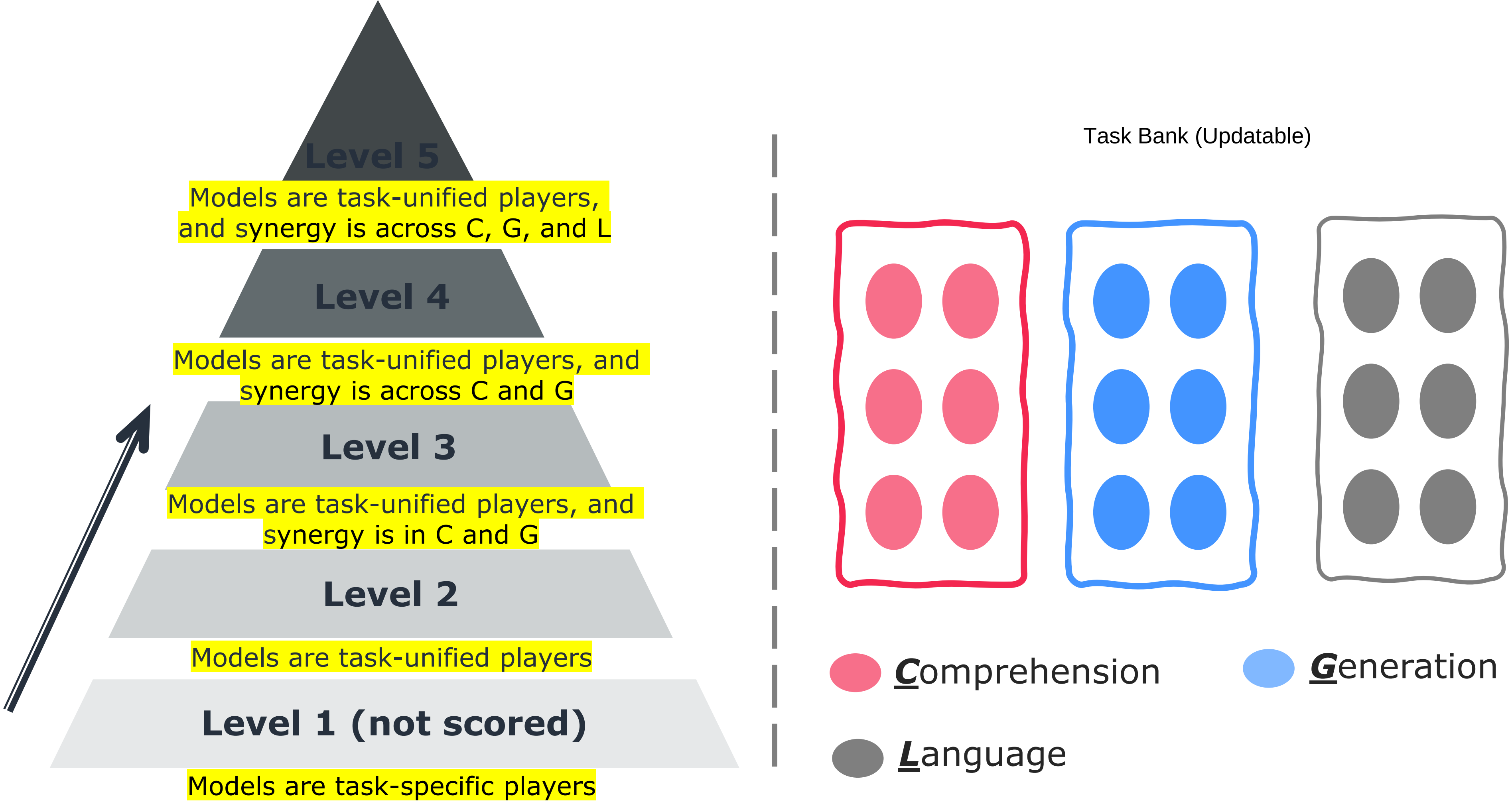
Today’s benchmarks are challenging but still fail to systematically reflect MLLMs’ synergy in/across comprehension and generation.


MMT-Bench
(162 Tasks)
Emodied Agent
Emotional Quotient
Evaluation Judgement
GUI Navigation
Image Translation
Intelligence Quotient
Keypoint Detection
Medical Understanding
Meme Understanding
Multi-Image Analysis
OCR
Object Counting
Object Localization
Pixel-level Recognition
Relation Reasoning
Temporal Understanding
Vision-Language Retrieval
Visual Coding
Visual Grounding
Visual Hallucination
Visual Illusion
Visual Prompting
Visual Recognition
Visual Summary
3D Perception
Action Recognition
Anomaly Detection
Autonomous Driving
Cross-Image Matching
Discipline Reasoning
Doc Understanding
6
1
7
6
3
1
5
5
5
2
8
35
2
5
4
2
3
7
5
4
6
9
4
4
2
2
5
5
1
2
4
2
Q: How does the man’s expression change?
Q: How to make a cup of flat white?
Q: Summarize the chart.
Q: What is the length of the shadow?
Q: Detect the marked object in the query image.
Q: What would the woman say to the man?
Q: Two responses are given, which response is better?
Q: Where should I Click to watch a tennis game?
Q: Is there any risk in the site?
Q: Identify the traffic sign in the image.
Q: Find the correct order to form a regular image.
Q: How many squares in the image?
Q: Detect the keypoint of person
Q: What is the action performed?
Q: Identify the category of the point cloud.
Q: Identify the imaging modality.
Q: What is this picture teasing?
Q: Spot the difference between images?
Q: Summarize the image in detail.
Q: What is in the yellow box.
Q: Identify the artwork form.
Q: What is all the text in the image?
Q: How many balloons are marked as '8'?
Q: Identify semantic category at coordinates (303, 95)
Q: Please detect all instances of the following categories.
Q: Is the laptop on the bench?
Q: In which frame does the player overtakes the background player?
Q: Provide bounding box for “man in jeans climbing stairs
Q: Is the size of each solid circle identical to the other?
Q: Retrieve the most similar handwritten text
$P_\theta(Y^k_{q,i}|X^k_V, X^k_q, Y^k_{q,<i})$
Q: Which Latex codes can compile into the formula?
Q: Does a cat exist in the image?

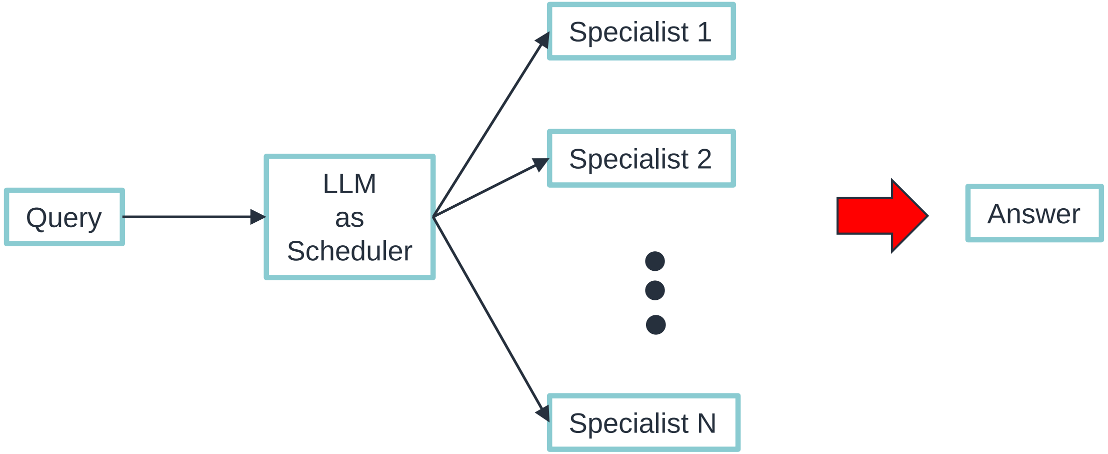
MMT-Bench Benchmark



Rank	Model	Score
1	GPT4o	65.5
2	InternVL-Chat-v1.2-34B	63.4
3	QwenVLMax	62.4
4	Qwen-VL-Plus	62.3
5	GeminiProVision	61.6
6	GPT4V_20240409	61.1
7	LLaVA-NEXT-34B	60.8
8	XComposer2	55.7
9	BLIP2	54.8

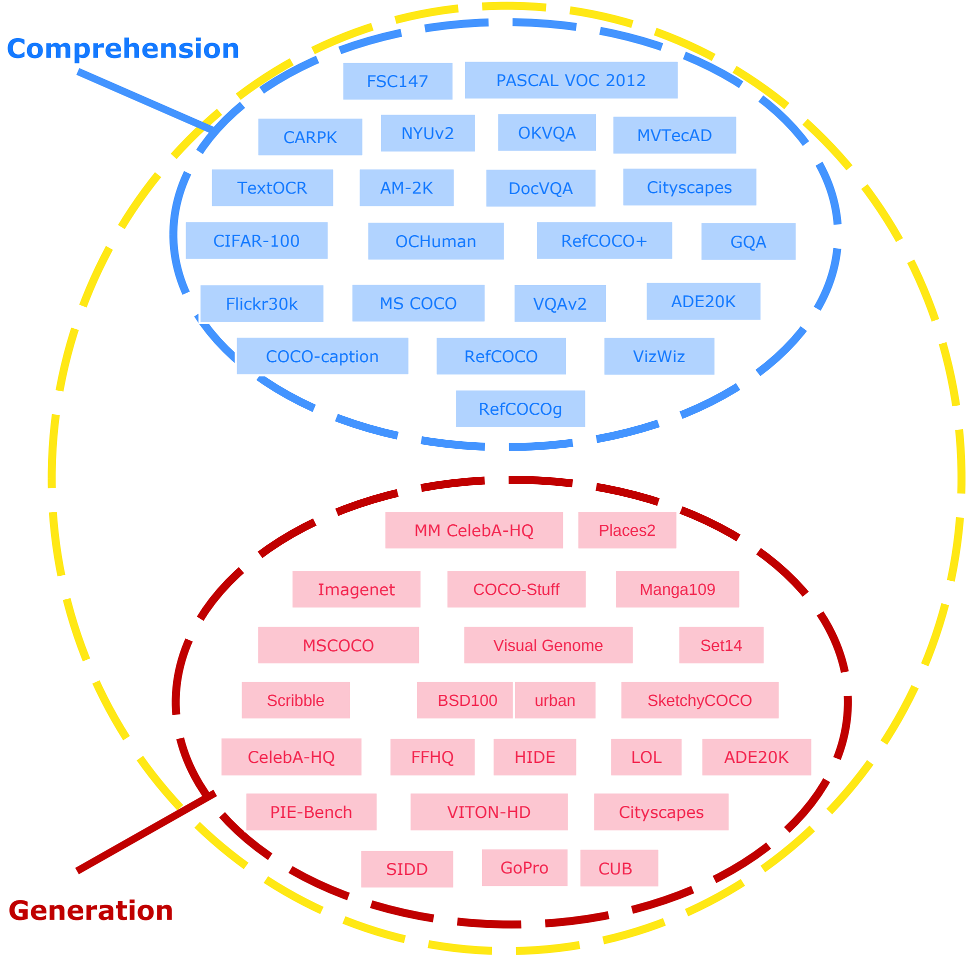

Level 5
Models are task-unified players, and synergy is across C, G, and L
Level 4
Models are task-unified players, and synergy is across C and G
Level 3
Models are task-unified players, and synergy is in C and G
Level 2
Models are task-unified players
Level 1 (not scored)
Models are task-specific players
Task Bank (Updatable)
Comprehension
Generation
Language

Task dispatcher is NOT synergy



ViperGPT: Visual Inference via Python Execution for Reasoning. ICCV'23
Visual Programming: Compositional visual reasoning without training. CVPR'23

Comprehension & Generation



Comprehension

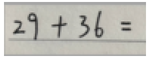
FSC147
PASCAL VOC 2012
CARPK
NYUv2
OKVQA
MVTecAD
TextOCR
AM-2K
DocVQA
Cityscapes
CIFAR-100
OCHuman
RefCOCO+
GQA
Flickr30k
MS COCO
VQAv2
ADE20K
COCO-caption
RefCOCO
VizWiz
RefCOCOg

MM CelebA-HQ
Places2
Imagenet
COCO-Stuff
Manga109
MSCOCO
Visual Genome
Set14
Scribble
BSD100
urban
SketchyCOCO
CelebA-HQ
FFHQ
HIDE
LOL
ADE20K
PIE-Bench
VITON-HD
Cityscapes
SIDD
GoPro
CUB

Generation

Example:

# Are there elephants in the image?
# Yes



# Is the answer to the above question 65?
# Yes

# Segment out the dog from the image.

# Describe the photo.
# A child in a purple outfit is seated on a chair.

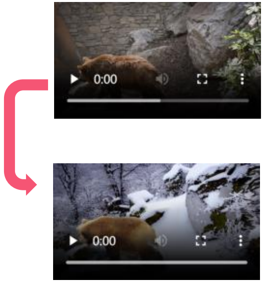
Comprehension

Generation

Example:

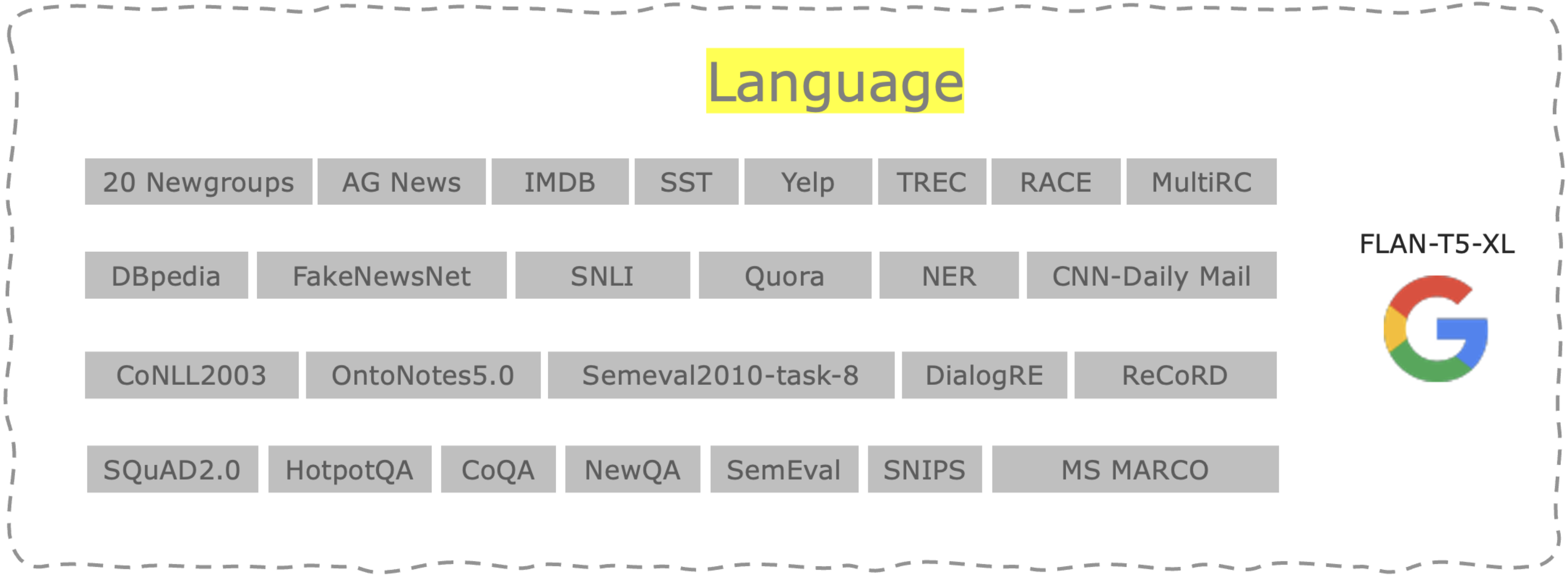
# Please generate a video where a dog run past a car on the street in the snow.

# Swap out the background of the video for a snowy scene.



# Generate a picture about burning fire.

NLP



Language

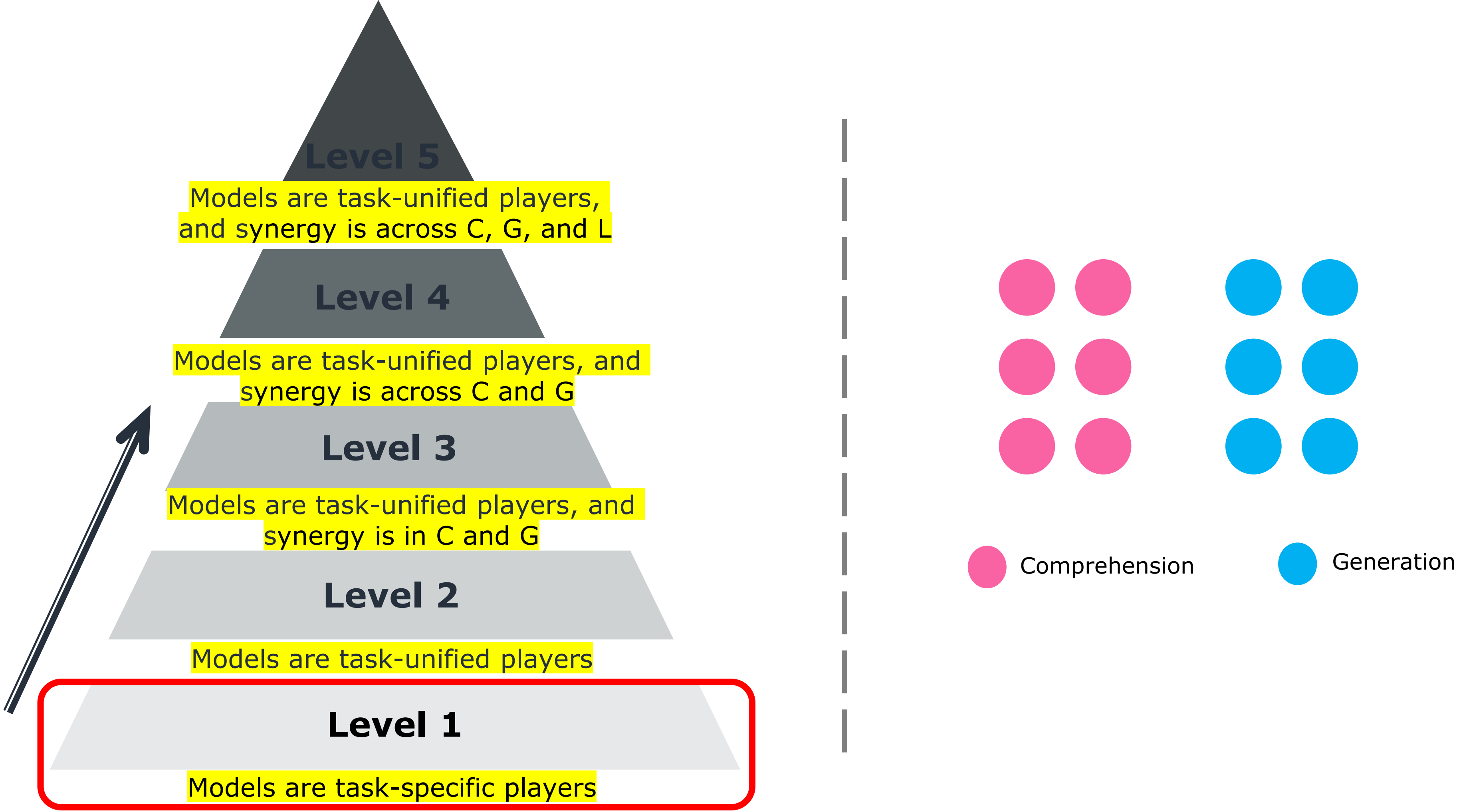
20 Newgroups | AG News | IMDB | SST | Yelp | TREC | RACE | MultiRC

DBpedia | FakeNewsNet | SNLI | Quora | NER | CNN-Daily Mail

CoNLL2003 | OntoNotes5.0 | Semeval2010-task-8 | DialogRE | ReCoRD

SQuAD2.0 | HotpotQA | CoQA | NewQA | SemEval | SNIPS | MS MARCO

FLAN-T5-XL



Level 5

Models are task-unified players, and synergy is across C, G, and L

Level 4

Models are task-unified players, and synergy is across C and G

Level 3

Models are task-unified players, and synergy is in C and G

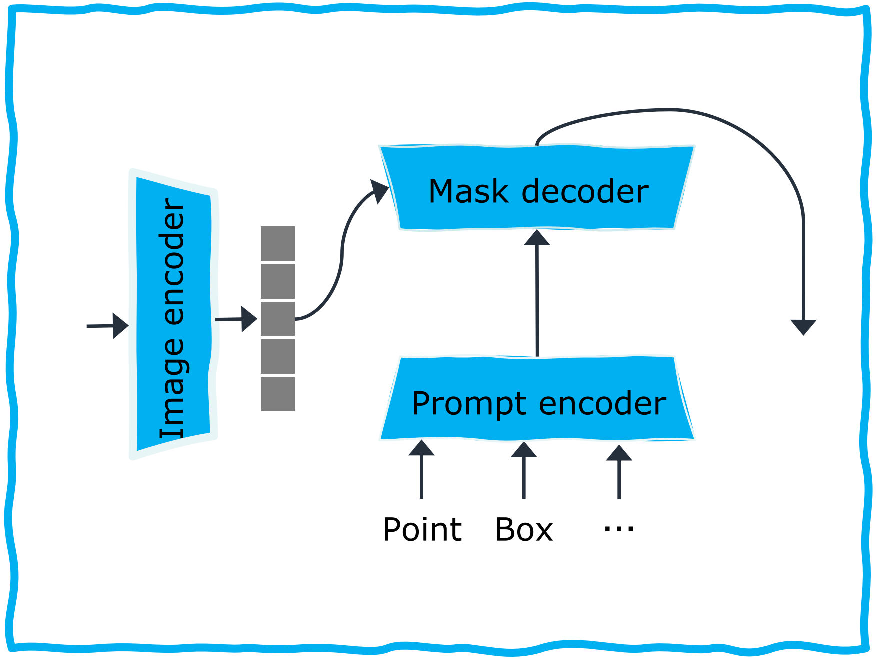
Level 2

Models are task-unified players

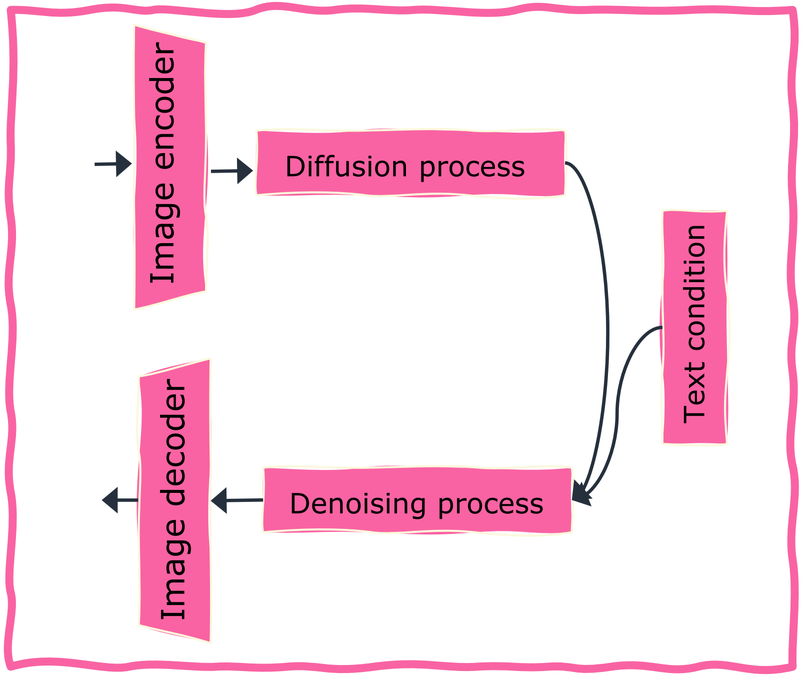
Level 1

Models are task-specific players

Level 1: Specialist



SAM (Segmentation)



Stable Diffusion (Text-to-image)

Examples for the framework of specialist models.
They are specially designed/fine-tuned for specific tasks.

Level 1: Specialist

Comprehension

FSC147	PASCAL VOC 2012	CARPK	NYUv2	OKVQA
counTR	SegCLIP	counTR	TransDepth	GIT

TextOCR	AM-2K	DocVQA	Cityscapes	CIFAR-100
Parseq	GFM	Donut	OneFormer	Astroformer

RefCOCO+	GQA	Flickr30k	MS COCO	VQAv2	ADE20K
polygon-former	GIT	CLIP	DINO	GIT	OneFormer

COCO-caption	RefCOCO	VizWiz	MVTecAD	OCHuman
GRIT	polygon-former	GIT	CPR	BUCTD

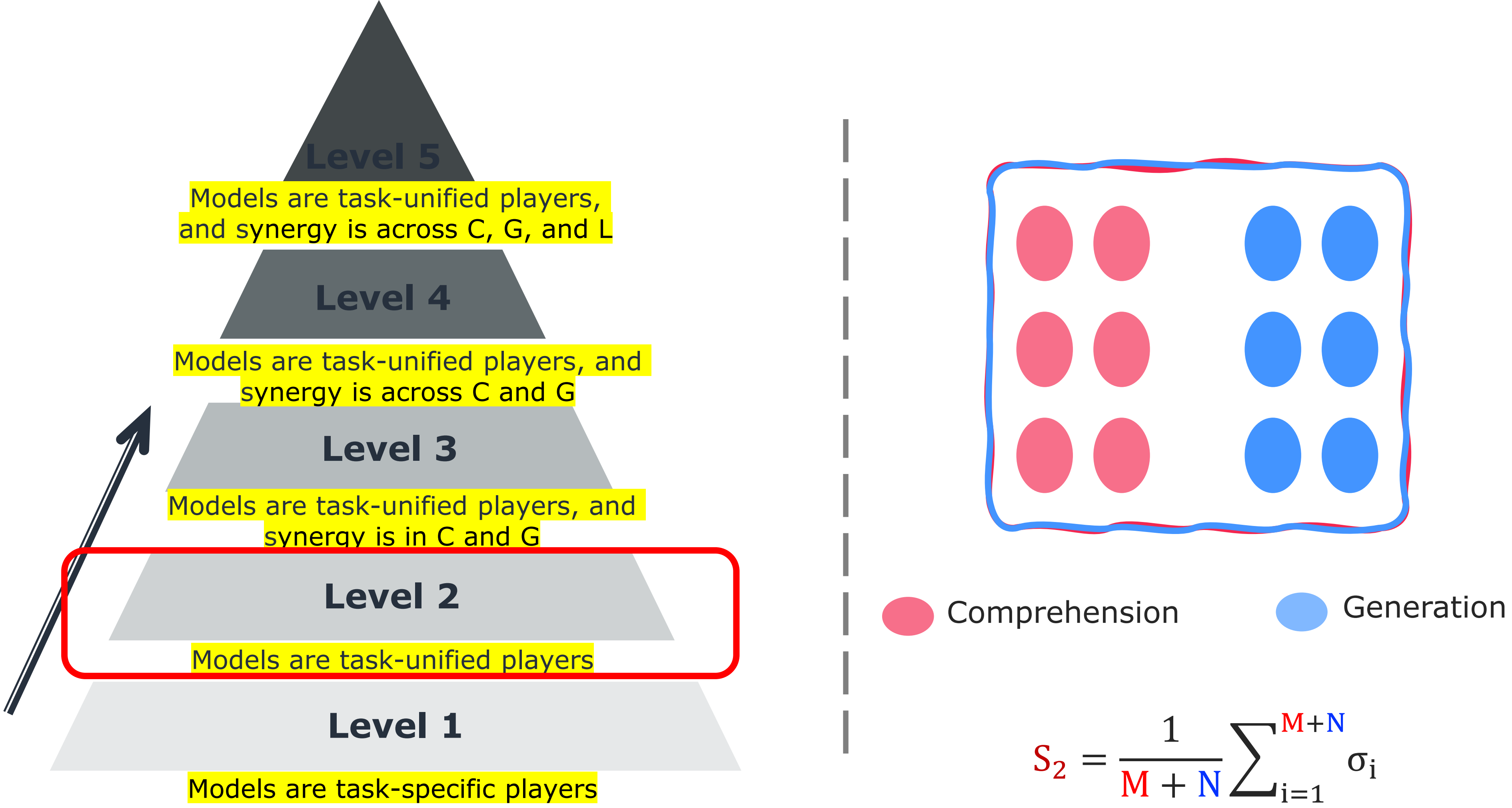
Generation

MM CelebA-HQ	Places2	Imagenet	COCO-Stuff	Manga109
Lafite	Mat	VQ-diff	layoutdiff	Hat

MSCOCO	Visual Genome	Set14	SketchyCOCO	ADE20K
Lafite	layoutdiff	Hat	Piti	Inade

Scribble	BSD100	urban	Cityscapes	HIDE	LOL	FFHQ
Piti	Hat	Hat	Inade	Lakdnet	WaveNet	Mat

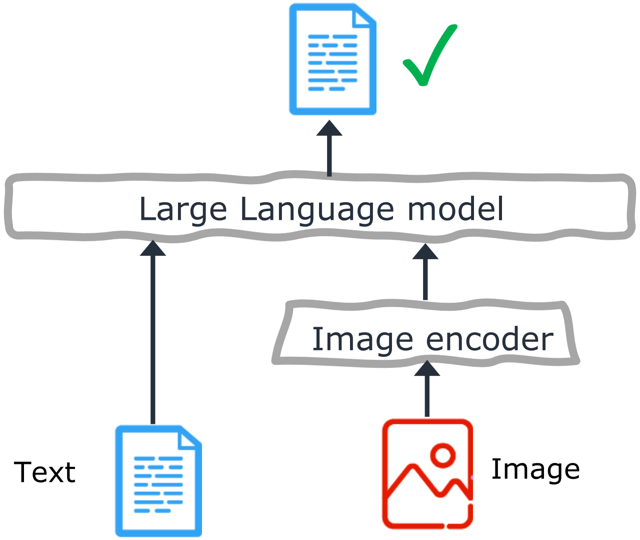
CelebA-HQ	PIE-Bench	VITON-HD	SIDD	GoPro	CUB
Mat	P2P	Mgd	Hinet	Lakdnet	Lafite


Level 5
Models are task-unified players, and synergy is across C, G, and L
Level 4
Models are task-unified players, and synergy is across C and G
Level 3
Models are task-unified players, and synergy is in C and G
Level 2
Models are task-unified players
Level 1
Models are task-specific players
● Comprehension
● Generation

$$S_2 = \frac{1}{M + N} \sum_{i=1}^{M+N} \sigma_i$$

Level 2: Unified C and G

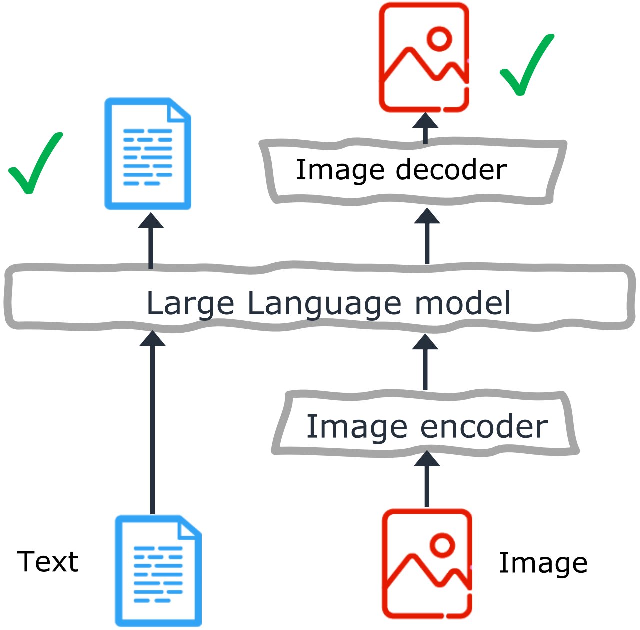
Unified comprehension



Flamingo[58]

(QA: Text+image)

Unified comprehension and Generation

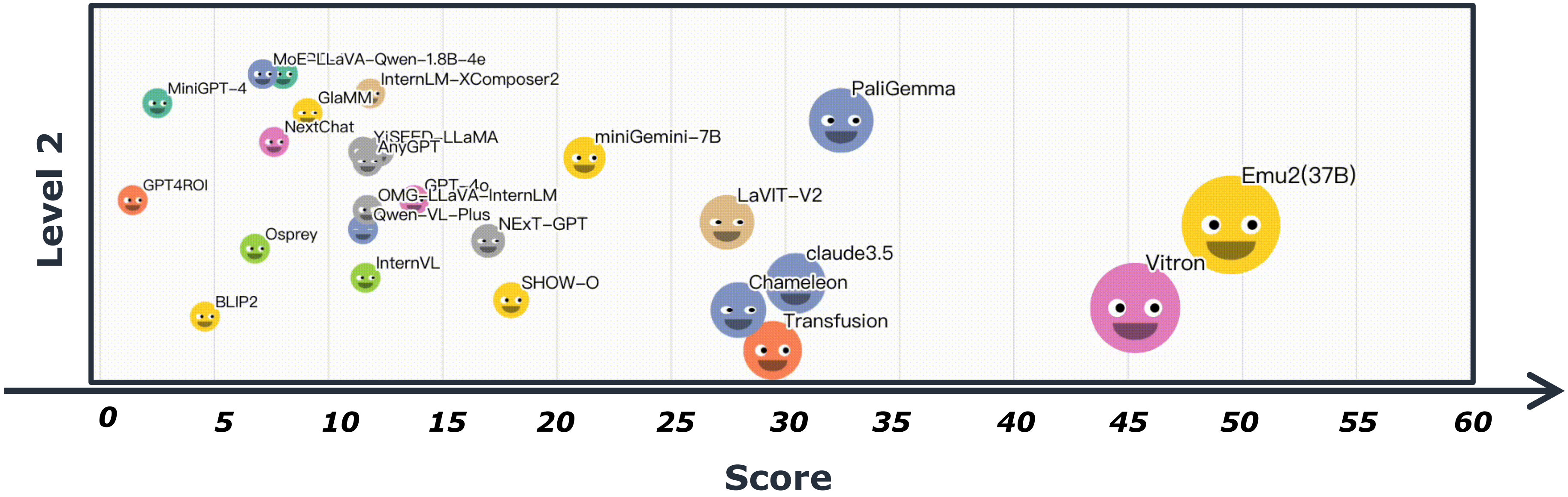


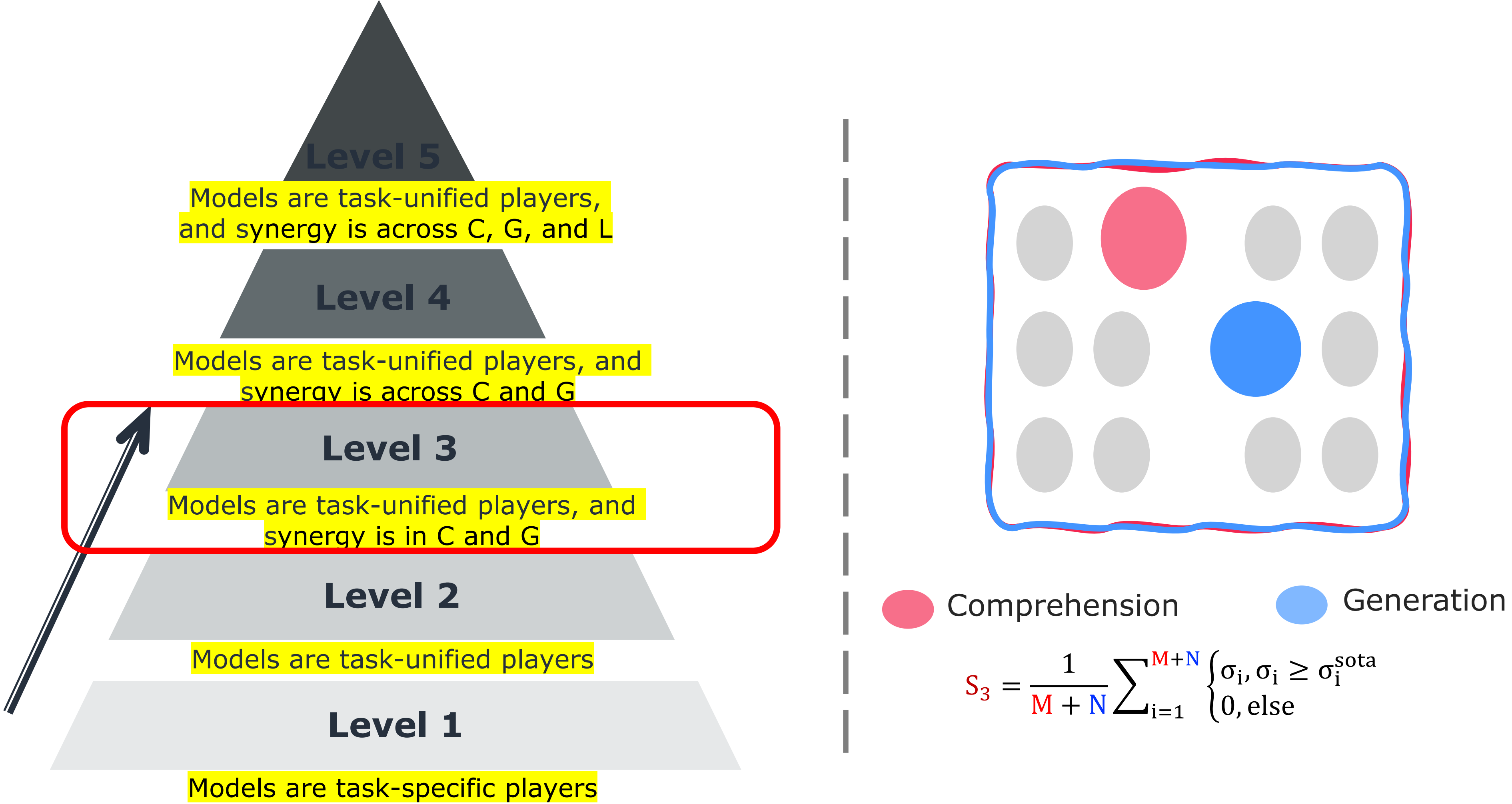
MM-Interleaved[68]

(QA: Text+image; Text-to-image, image editing)

Examples for the framework of unifying C and/or G.

The score of Level 2





Level 5

Models are task-unified players, and synergy is across C, G, and L

Level 4

Models are task-unified players, and synergy is across C and G

Level 3

Models are task-unified players, and synergy is in C and G

Level 2

Models are task-unified players

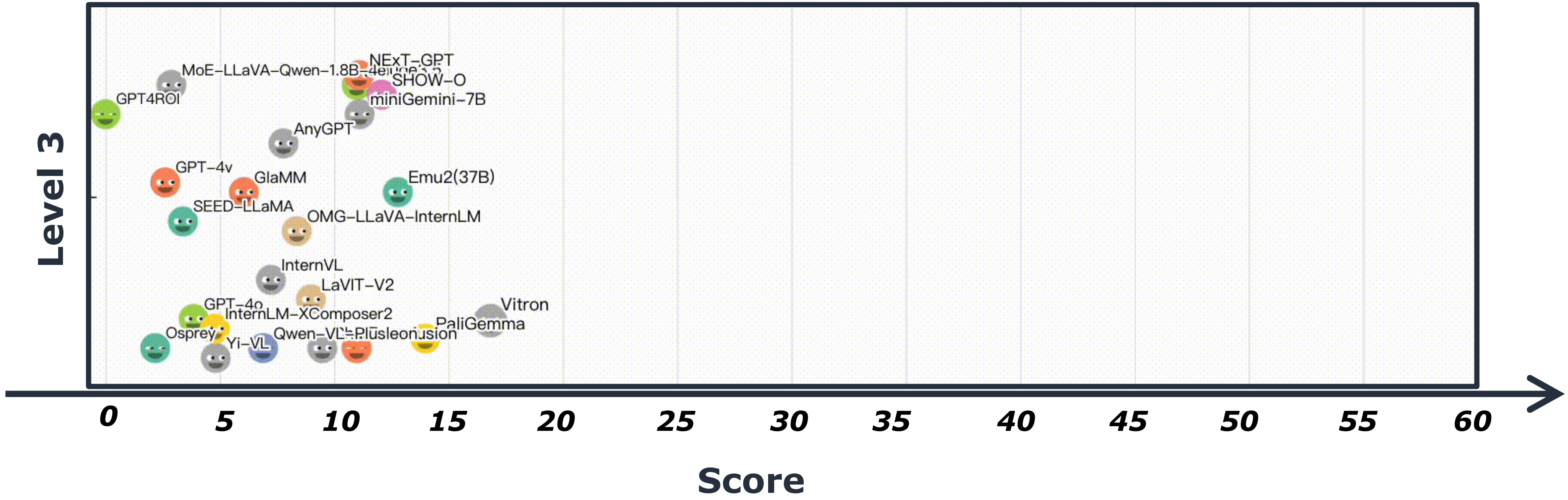
Level 1

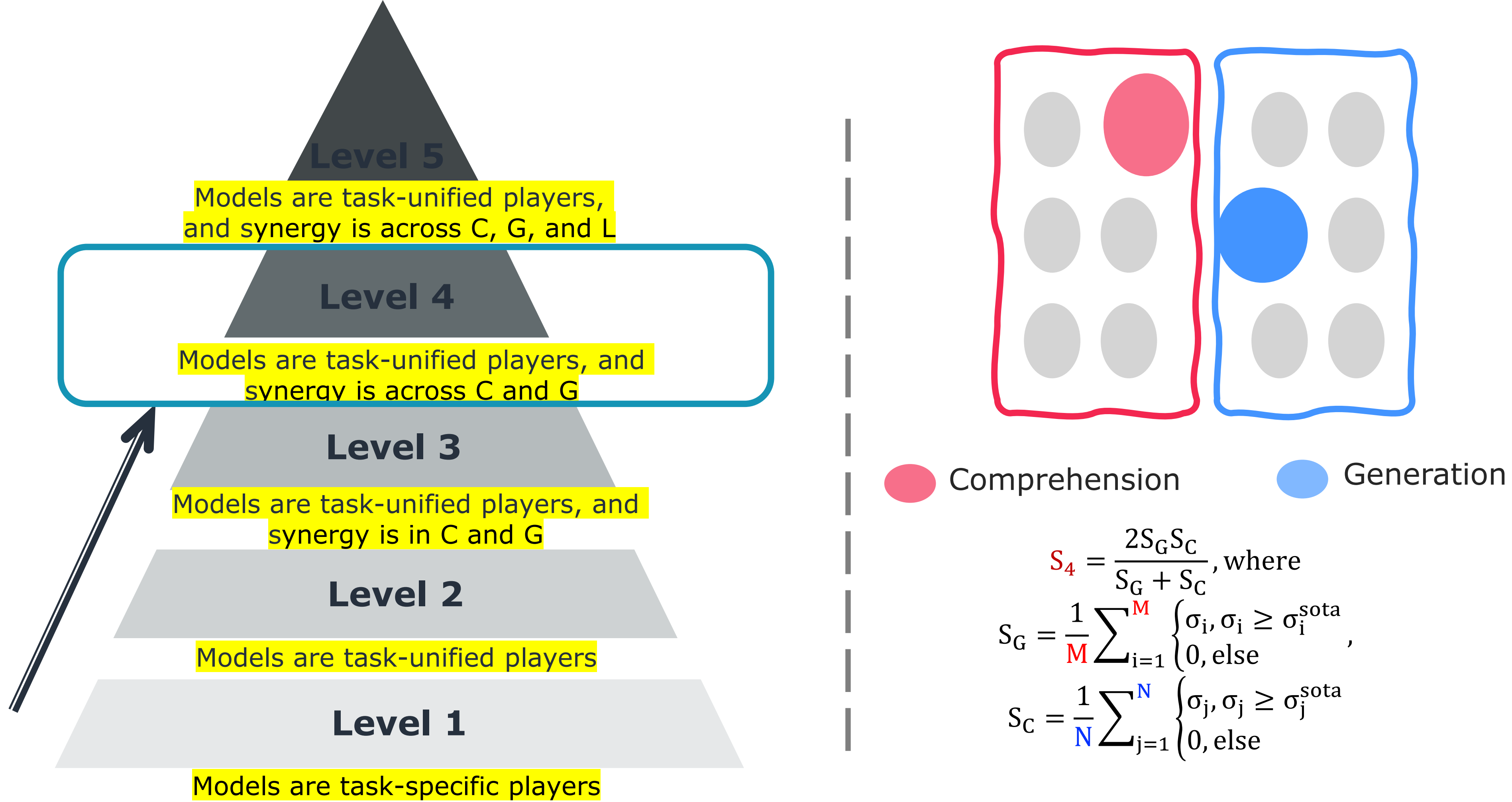
Models are task-specific players

● Comprehension ● Generation

$$S_3 = \frac{1}{M+N}\sum_{i=1}^{M+N}\begin{cases}\sigma_i, \sigma_i \geq \sigma_i^{sota}\\ 0, \text{else}\end{cases}$$

The score of Level 3





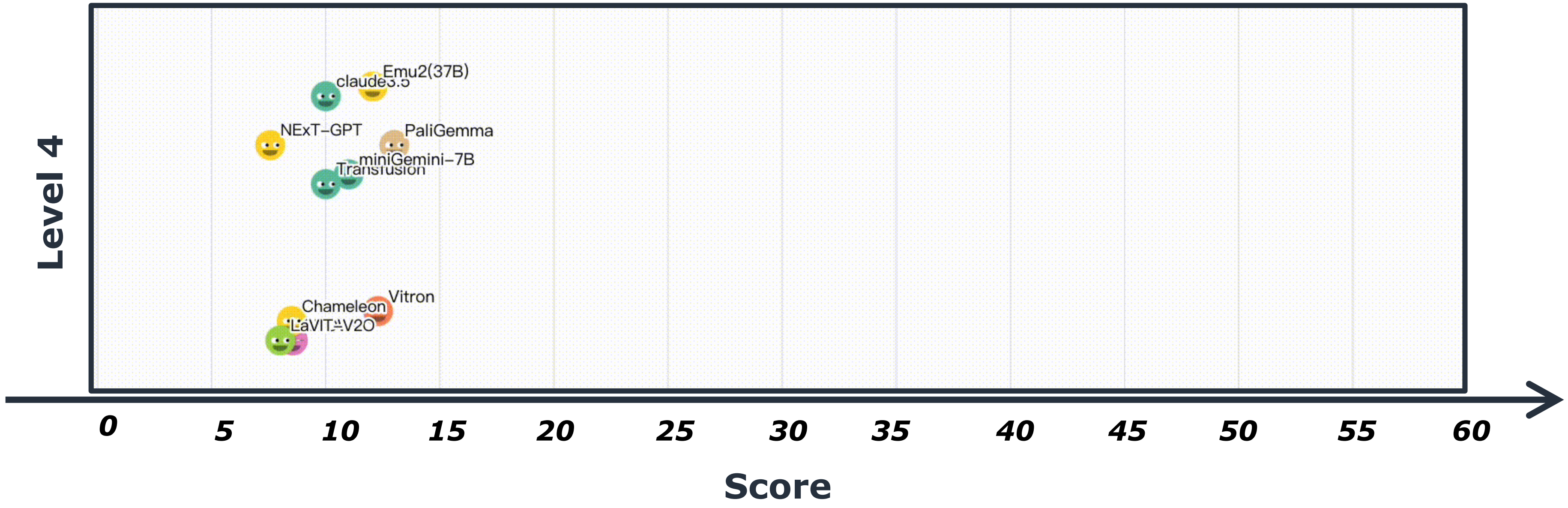
Level 5
Models are task-unified players, and synergy is across C, G, and L
Level 4
Models are task-unified players, and synergy is across C and G
Level 3
Models are task-unified players, and synergy is in C and G
Level 2
Models are task-unified players
Level 1
Models are task-specific players
Comprehension
Generation

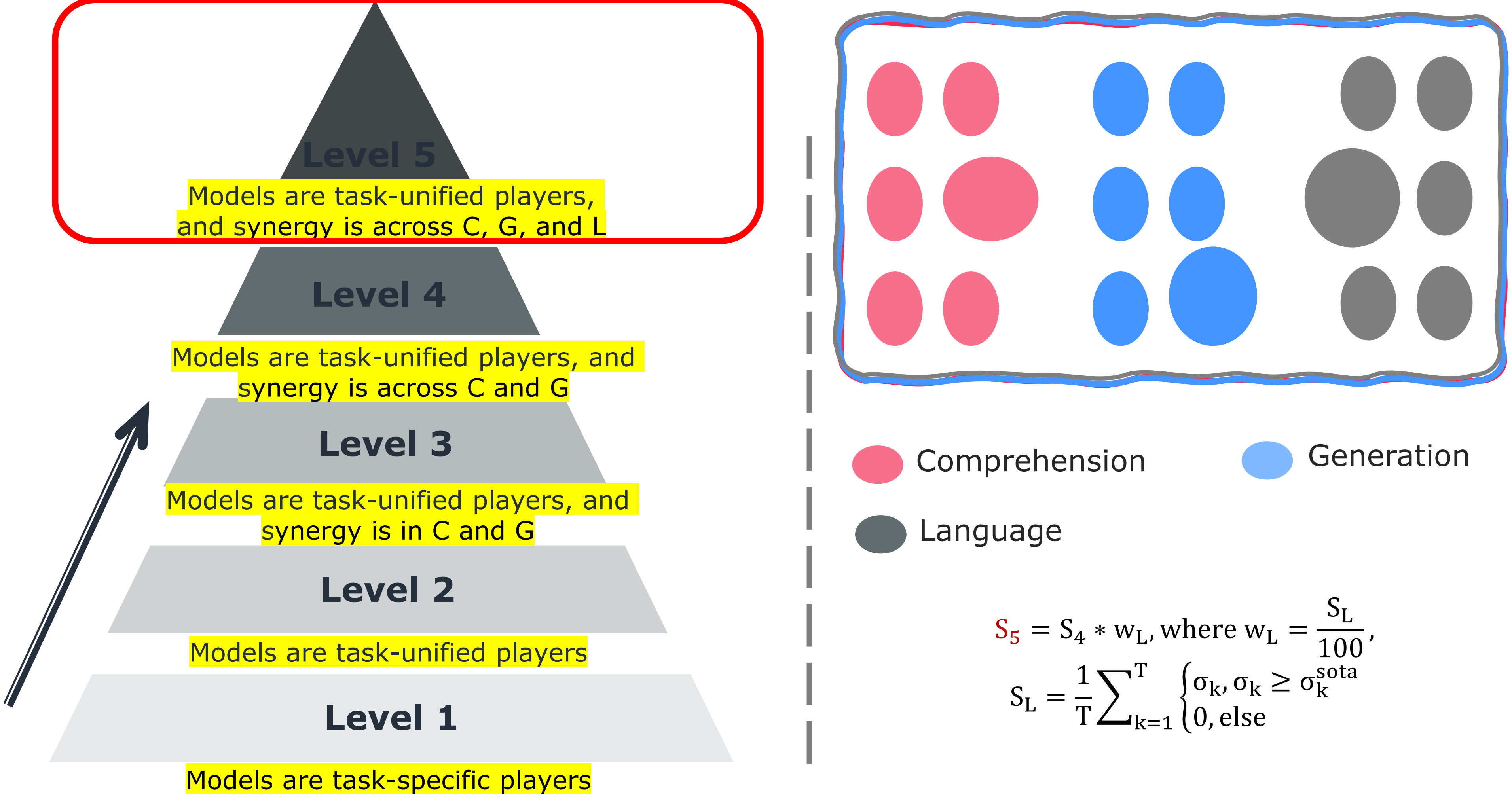
$$S_4 = \frac{2S_G S_C}{S_G + S_C}, \text{where}$$

$$S_G = \frac{1}{M}\sum_{i=1}^{M} \begin{cases} \sigma_i, \sigma_i \geq \sigma_i^{sota} \\ 0, \text{else} \end{cases},$$

$$S_C = \frac{1}{N}\sum_{j=1}^{N} \begin{cases} \sigma_j, \sigma_j \geq \sigma_j^{sota} \\ 0, \text{else} \end{cases}$$

The score of Level 4



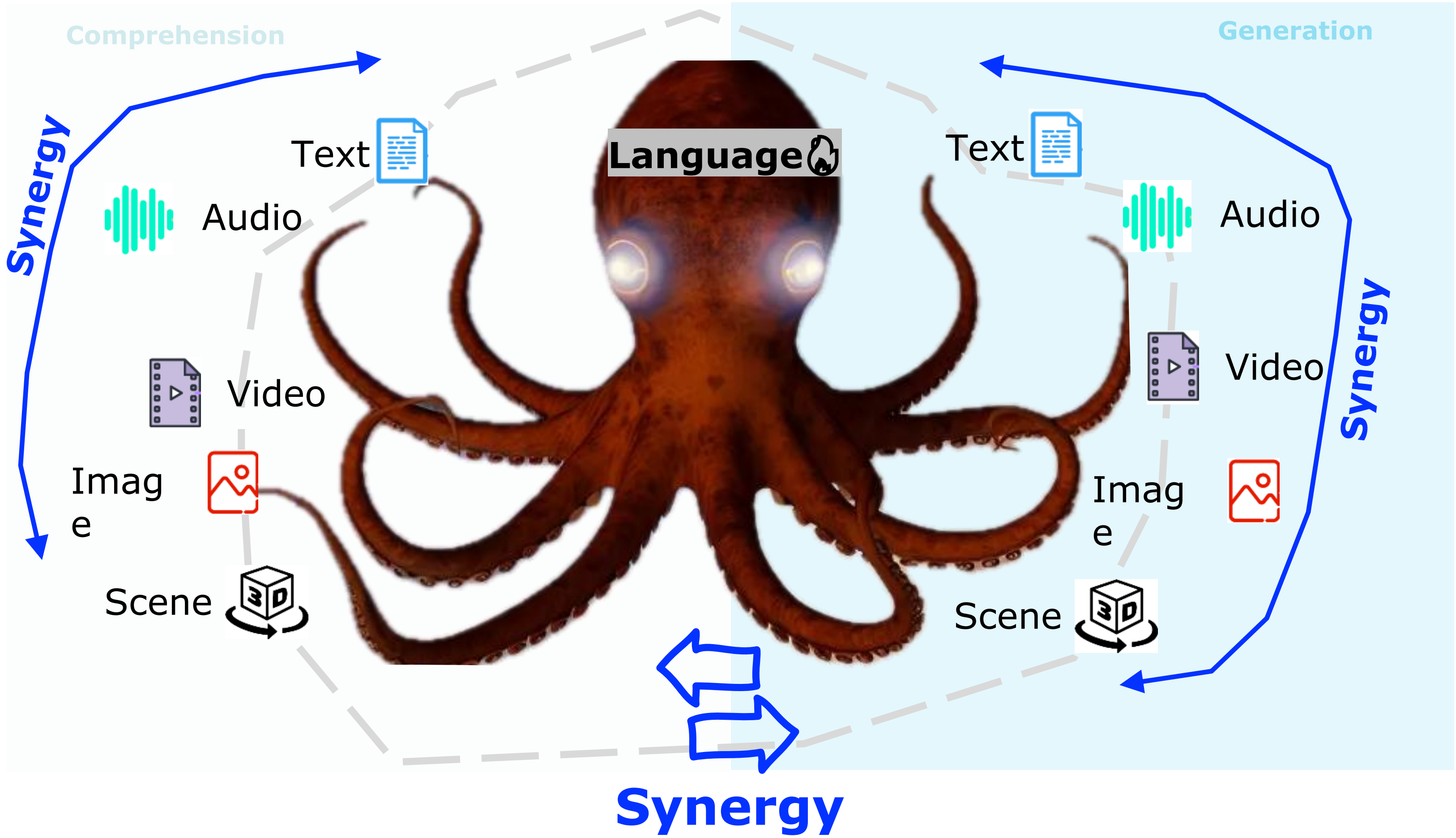

Level 5
Models are task-unified players, and synergy is across C, G, and L
Level 4
Models are task-unified players, and synergy is across C and G
Level 3
Models are task-unified players, and synergy is in C and G
Level 2
Models are task-unified players
Level 1
Models are task-specific players
Comprehension
Generation
Language

$$S_5 = S_4 * w_L, \text{where } w_L = \frac{S_L}{100},$$

$$S_L = \frac{1}{T}\sum_{k=1}^{T}\begin{cases}\sigma_k, \sigma_k \geq \sigma_k^{sota}\\ 0, \text{else}\end{cases}$$

This is our goal!

Level 5: Total Synergy



Upgrade

Level	Definition	Score	Example
1: Specialist	Models are task-specific players	$\mathrm{N.A.}$	**Dino, polygon-former, SegCLIP**
2:Unified C ***and*** G	Models are task-unified players	$\mathrm{S_2} = \frac{1}{\mathrm{M+N}}\sum_{\mathrm{i=1}}^{\mathrm{M+N}} \sigma_{\mathrm{i}}$	**miniGPT4, NextChat**
3: Synergy in C ***and*** G	Models are task-unified players, and synergy is in C and/or G	$\mathrm{S_3} = \frac{1}{\mathrm{M+N}}\sum_{\mathrm{i=1}}^{\mathrm{M+N}} \begin{cases}\sigma_{\mathrm{i}}, \sigma_{\mathrm{i}} \geq \sigma_{\mathrm{i}}^{\mathrm{sota}} \\ 0, \mathrm{else}\end{cases}$	**GPT4v, GPT4o, LLaVA1.5-7b, Qwen-VL-Pius, InternVL, MoE-LLaVA-1.8B-4e, Yi-vl, SEED-LLaMA-14B-SFT, Osprey, GlaMM**
4: Synergy ***across*** C and G	Models are task-unified players, and synergy is across C and G	$\mathrm{S_4} = \frac{2\mathrm{S_G}\mathrm{S_C}}{\mathrm{S_G}+\mathrm{S_C}}, \mathrm{where}$ $\mathrm{S_G} = \frac{1}{\mathrm{M}}\sum_{\mathrm{i=1}}^{\mathrm{M}} \begin{cases}\sigma_{\mathrm{i}}, \sigma_{\mathrm{i}} \geq \sigma_{\mathrm{i}}^{\mathrm{sota}} \\ 0, \mathrm{else}\end{cases},$ $\mathrm{S_C} = \frac{1}{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{N}} \begin{cases}\sigma_{\mathrm{j}}, \sigma_{\mathrm{j}} \geq \sigma_{\mathrm{j}}^{\mathrm{sota}} \\ 0, \mathrm{else}\end{cases}$	**miniGemini-7B, Emu2-37B, Vitron,Next-GPT,LaVIT-V2-7B, SHOW-O, Claude3.5 Chameleon, PaliGemma, Transfusion**
5: Total Synergy Synergy ***across*** C, G, and L	Models are task-unified players, and synergy is across C, G, and L	$\mathrm{S_5} = \mathrm{S_4} * \mathrm{w_L}, \mathrm{where}\ \mathrm{w_L} = \frac{\mathrm{S_L}}{100},$ $\mathrm{S_L} = \frac{1}{\mathrm{T}}\sum_{\mathrm{k=1}}^{\mathrm{T}} \begin{cases}\sigma_{\mathrm{k}}, \sigma_{\mathrm{k}} \geq \sigma_{\mathrm{k}}^{\mathrm{sota}} \\ 0, \mathrm{else}\end{cases}$	**None, this is our goal!**

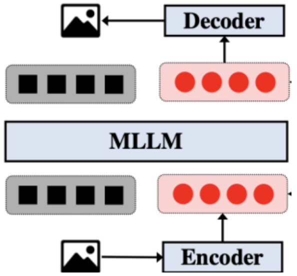
Table1. The roadmap to L5 MM Generalist

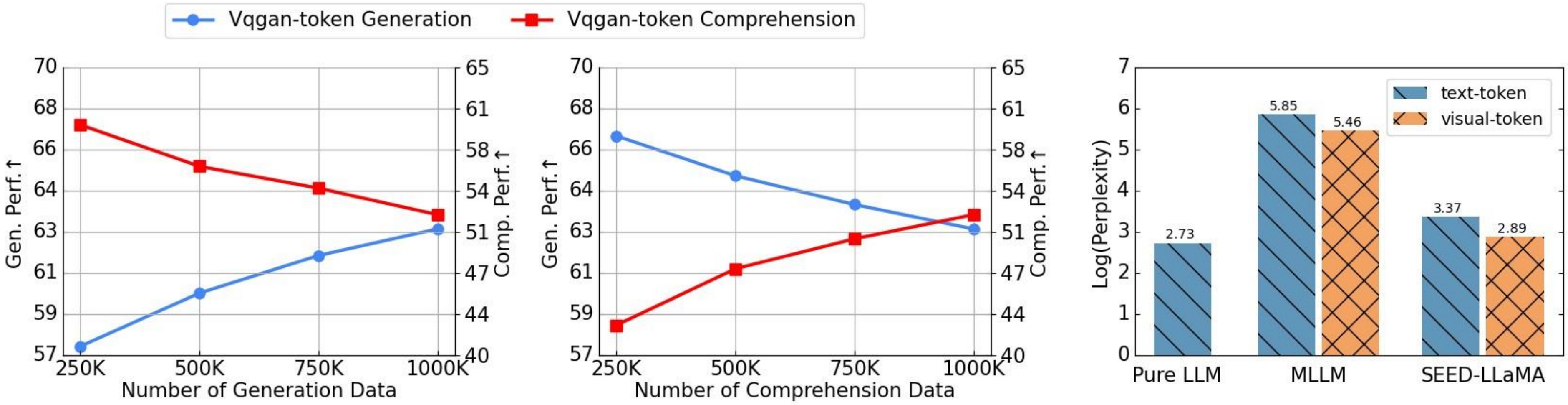
Climbing: L3->L5

- Most are <=L3
- Why?
  - Comprehension: V-token to **lose** info to match T-token
  - Generation: V-token must **preserve** info
  - V-Language **!=** T-Language
  - V-Generation **!=** T-Generation

Comprehension and Generation Inconsistency

- Conflicting objectives





(a)

(b)

(c)

Auto-Encoding Morph-Tokens for Multimodal LLM. ICML’24 spotlight

Challenge 1: Language (V) ≠ Language (L)

- Spatial visual tokens are just word spelling, not language

Noam Chomsky
1928-present

A man

A [*old white*] man

A [old white] man [*with white hair*]

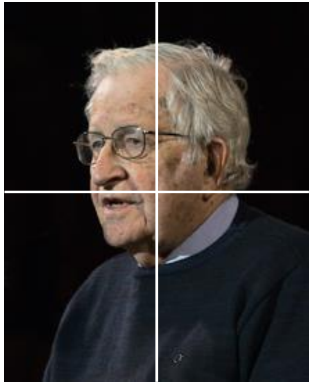
A [old white] man [with white hair] [*in black clothes*]

⋮

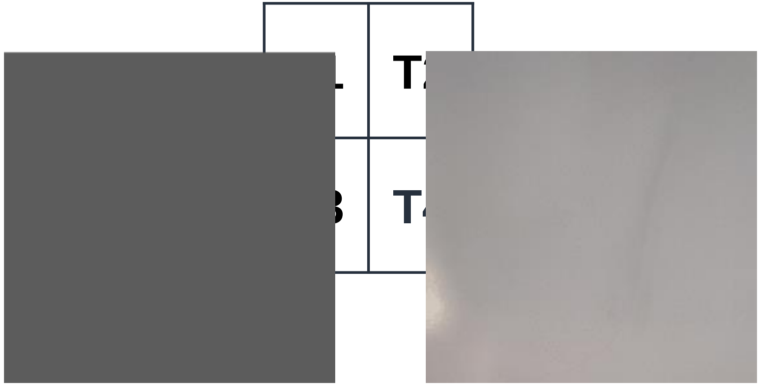
⋮

Recursive Syntax

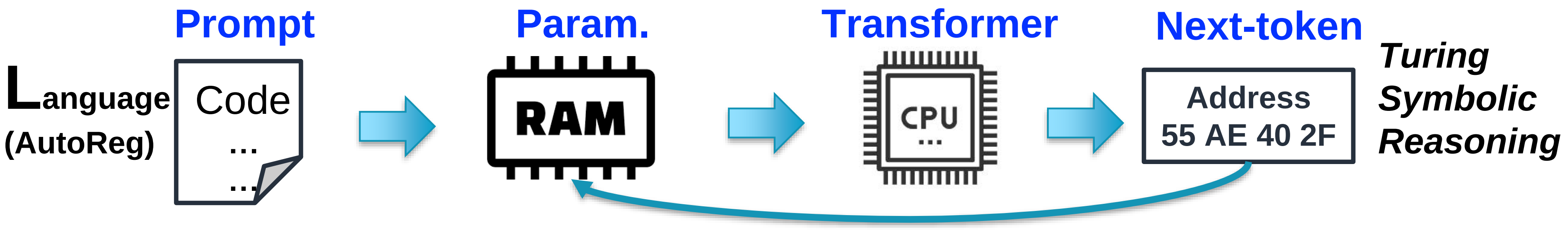
A furry dog sitting in a striped sofa.



Whole image is “spelled” as T1 T2 T3 T4



Challenge 2: Generation (V) ≠ Generation (L)



Prompt

Param.

Transformer

Next-token

Language (AutoReg)

Code … …

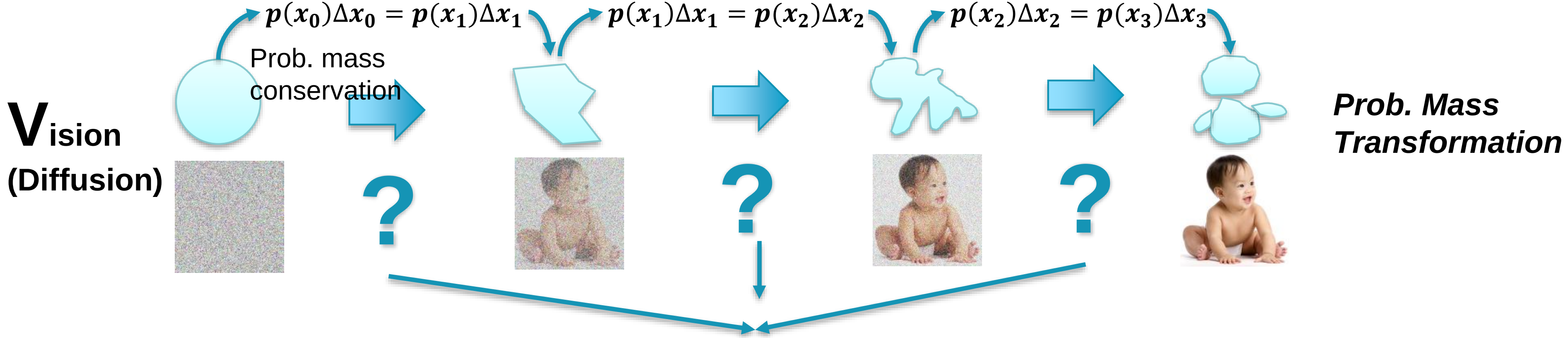
RAM

CPU …

Address 55 AE 40 2F

Turing Symbolic Reasoning

≠ ***Training objectives never align ☹***



$p(x_0)\Delta x_0 = p(x_1)\Delta x_1$

$p(x_1)\Delta x_1 = p(x_2)\Delta x_2$

$p(x_2)\Delta x_2 = p(x_3)\Delta x_3$

Prob. mass conservation

Vision (Diffusion)

? ? ?

Prob. Mass Transformation

the cause of transformation?

Thank you

Road to L5 MM Generalist

https://path2generalist.github.io