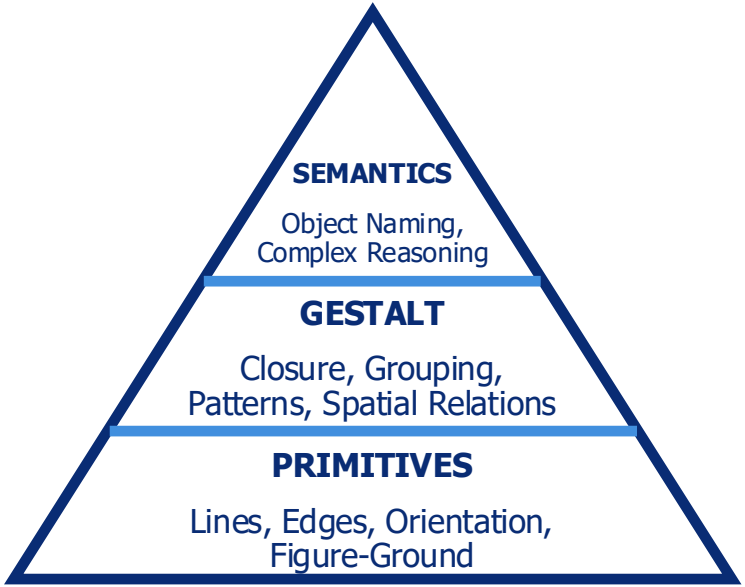
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Castle-in-the-Air: Probing the Foundational Visual Deficits of MLLMs via Bottom-Up Cognitive Factors

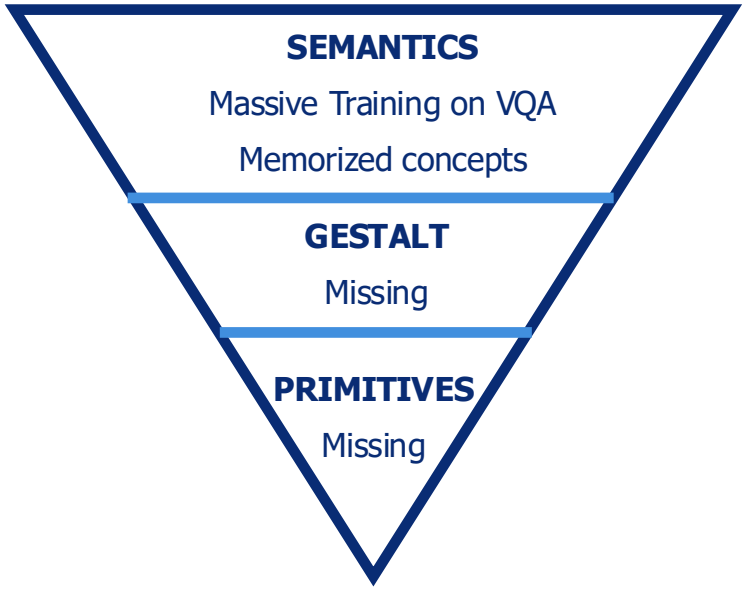
Jen-Tse Huang · Dasen Dai · Jen-Yuan Huang · Youliang Yuan · Xiaoyuan Liu · Wenxuan Wang · Wenxiang Jiao · Pinjia He · Zhaopeng Tu · Haodong Duan

NeurIPS 2026 E&D Track

Human vs. AI: Bottom-Up vs. Top-Down



Human Vision

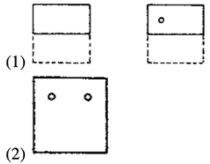


AI Vision

The VisFactor Benchmark

VZ2

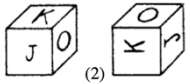
Paper Folding Test



Does the figure in (2) correctly show the paper after folding and punching via steps in (1)?
A: No.

S2

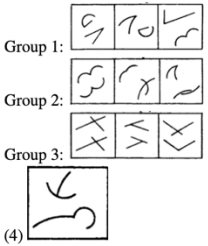
Cube Comparisons Test



Are the two cubes the same?
A: Yes.

I3

Figure Classification



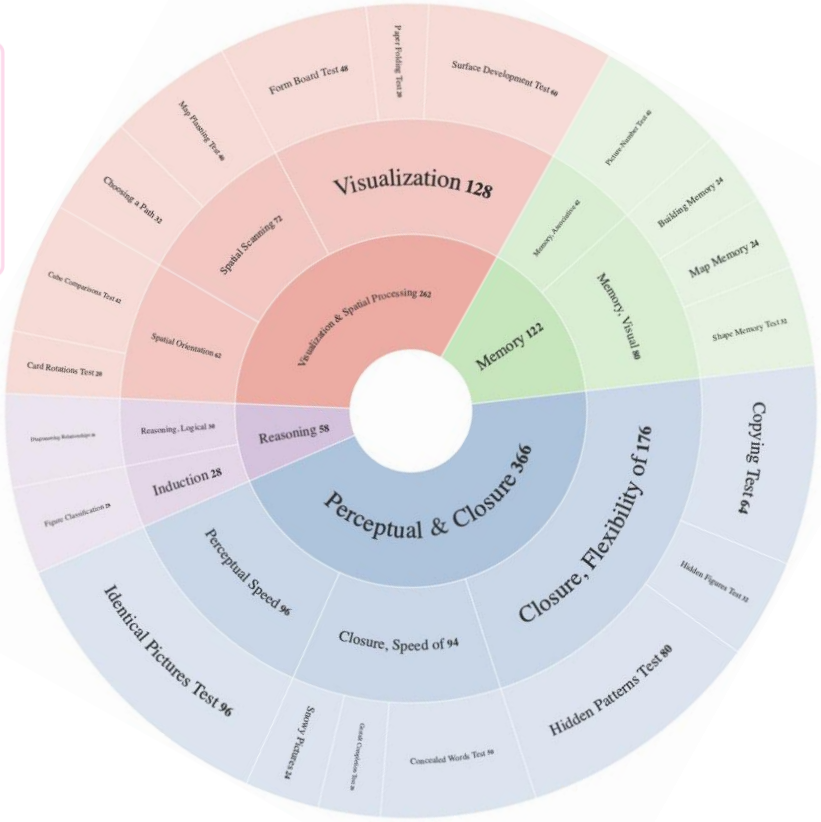
Which group is the figure in (4) from?
A: 2.

RL2

Diagramming Relationships

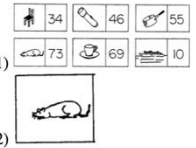


Can the diagram show the relationship between: Desks, furniture, pencils? **A: Yes.**



MA1

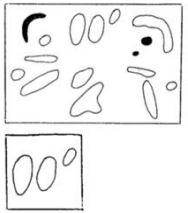
Picture-Number Test



Which number is the object in (2) related to in (1)?
A: 73.

MV1

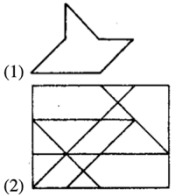
Shape Memory Test



Is the shape in (2) shown in shapes in (1)?
A: Yes.

CF1

Hidden Figures Test



Can the figure in (1) be found in the pattern in (2)? **A: No.**

CS1

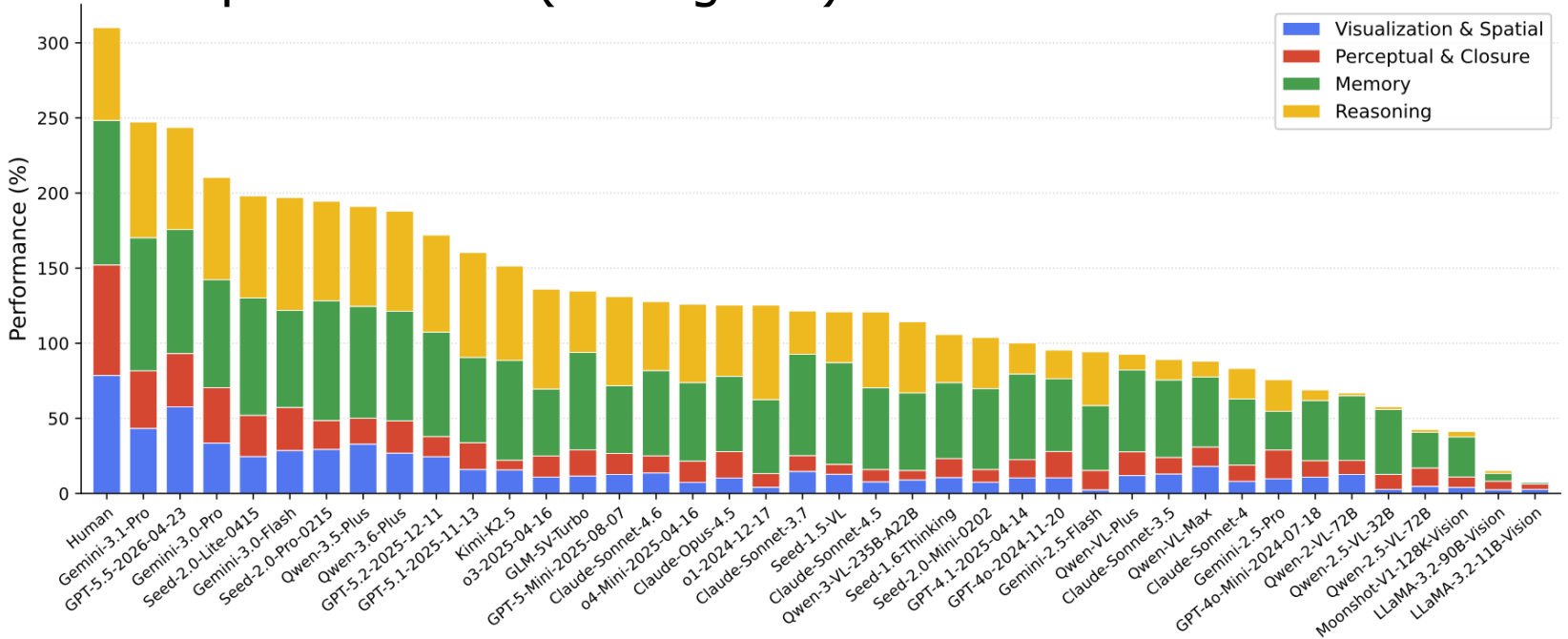
Gestalt Completion Test



What is the object?
A: A hammer head.

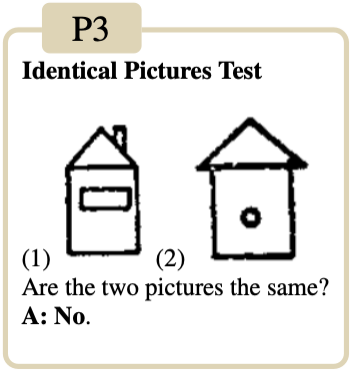
The VisFactor Leaderboard

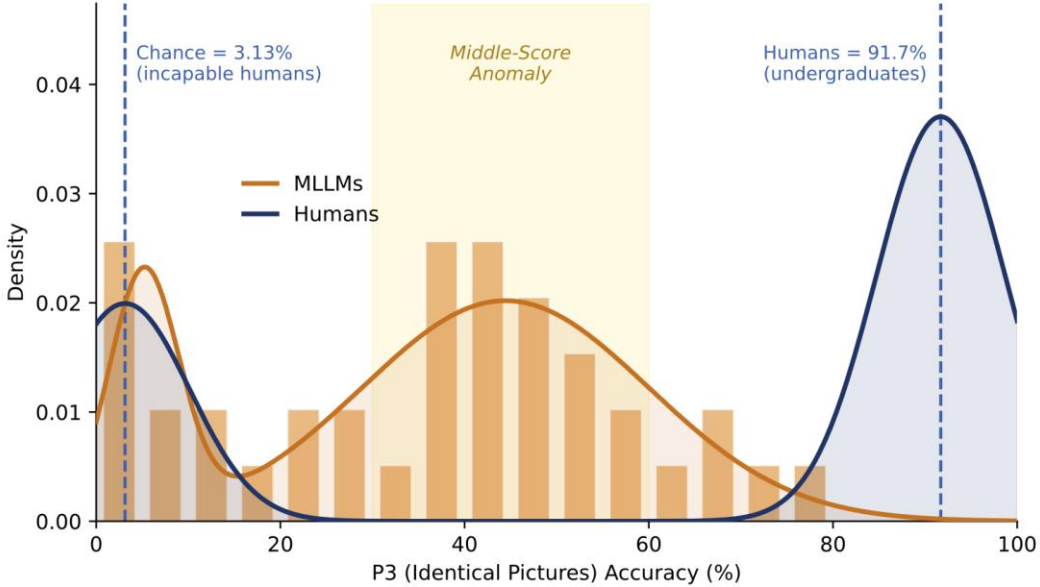
- Human performance (undergrads): 78.8



The "Middle-Score Anomaly"

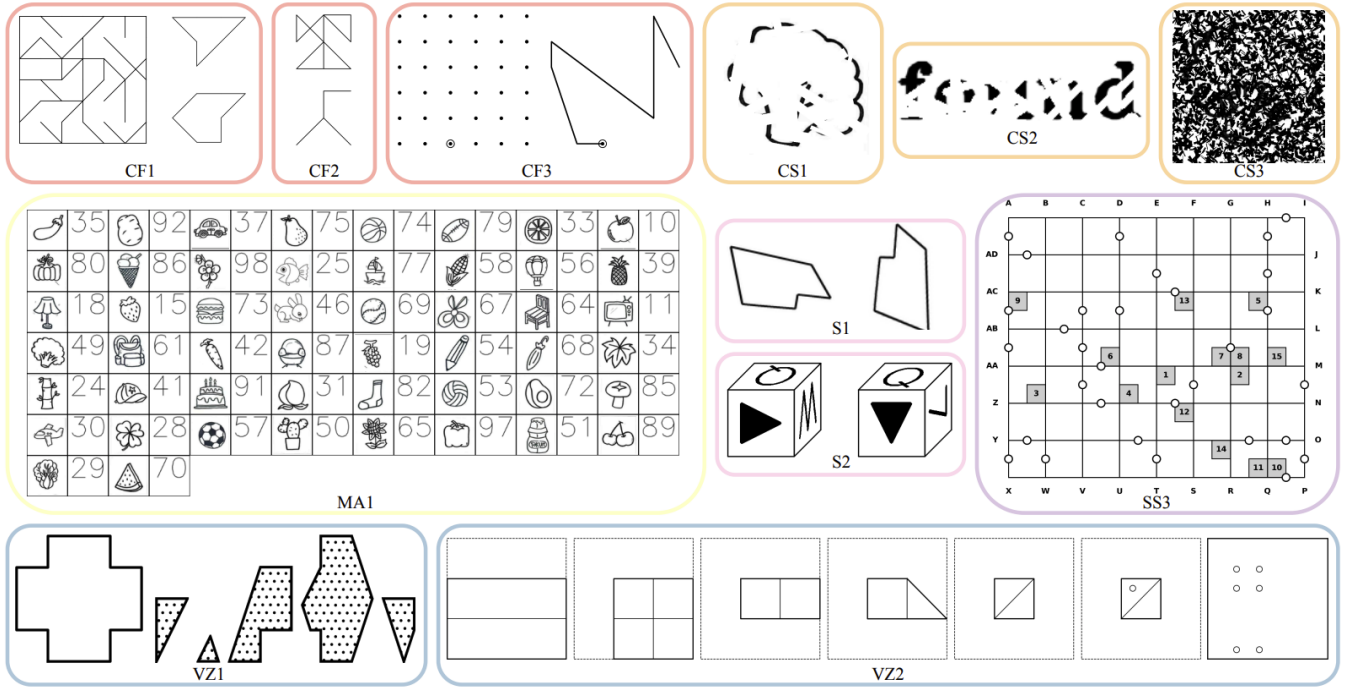
- Consider a very simple task
- What will humans perform?





Future-Proofing: Our Generator

- ✓ Unlimited tests
- ✓ Auto-verification
- ✓ Difficulty-control
- ✓ Step-by-step answers



Model Performance on Generated Tests

GPT-4.1

	CF1	CF2	CF3	CS1	CS2	CS3	MA1	S1	S2	SS3	VZ1	VZ2	AVG
Easy	3.1	13.8	15.6	40	78	45.8	88.1	0	2.4	45	14.6	0	28.9
Normal	0	12.5	4.7	35	76	16.7	90.5	0	0	30	12.5	0	23.2
Hard	0	17.5	4.7	35	52	25	78.6	0	0	32.5	18.8	0	22.0

GPT-5.5

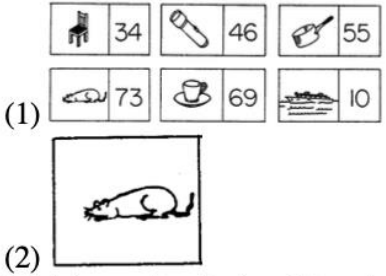
	CF1	CF2	CF3	CS1	CS2	CS3	MA1	S1	S2	SS3	VZ1	VZ2	AVG
Easy	93.8	100	100	30	60	20.8	100	60	40.5	100	87.5	90	73.6
Normal	25	98.8	100	15	44	4.2	100	40	35.7	95	87.5	80	60.4
Hard	6.3	98.8	100	5	22	4.2	100	25	35.7	92.5	75	40	50.4

- GPT-4.1: A floor effect
- GPT-5.5: Monotonic decreasing

Case Study: MA1

MA1

Picture-Number Test



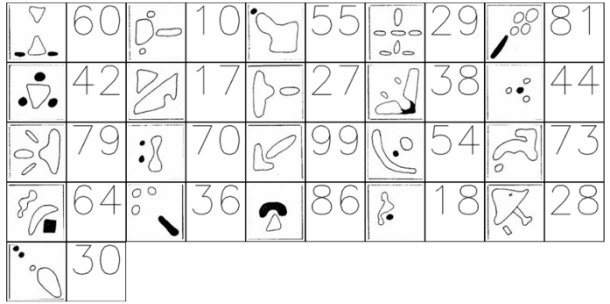
Which number is the object in (2) related to in (1)?

A: 73.

- Models perform well on MA1 with semantically meaningful images
- What if we change to some semantically meaningless images?



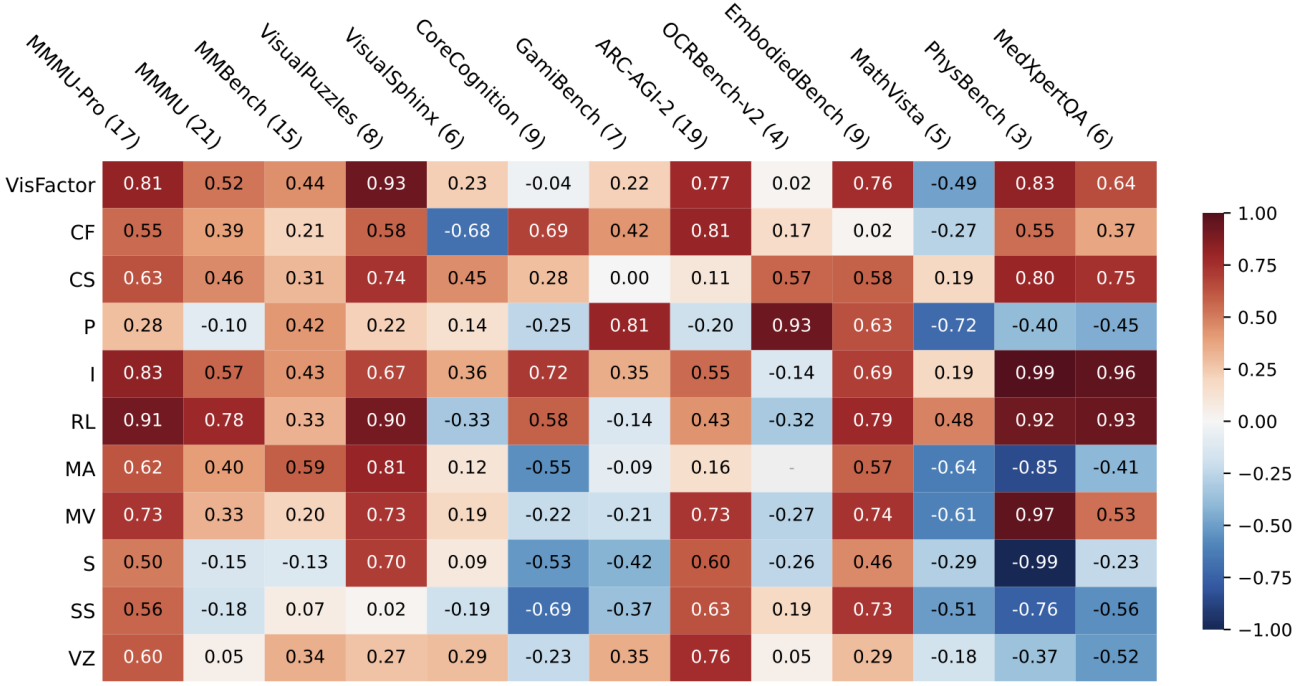
(a) MA1 generated with CF2 figures



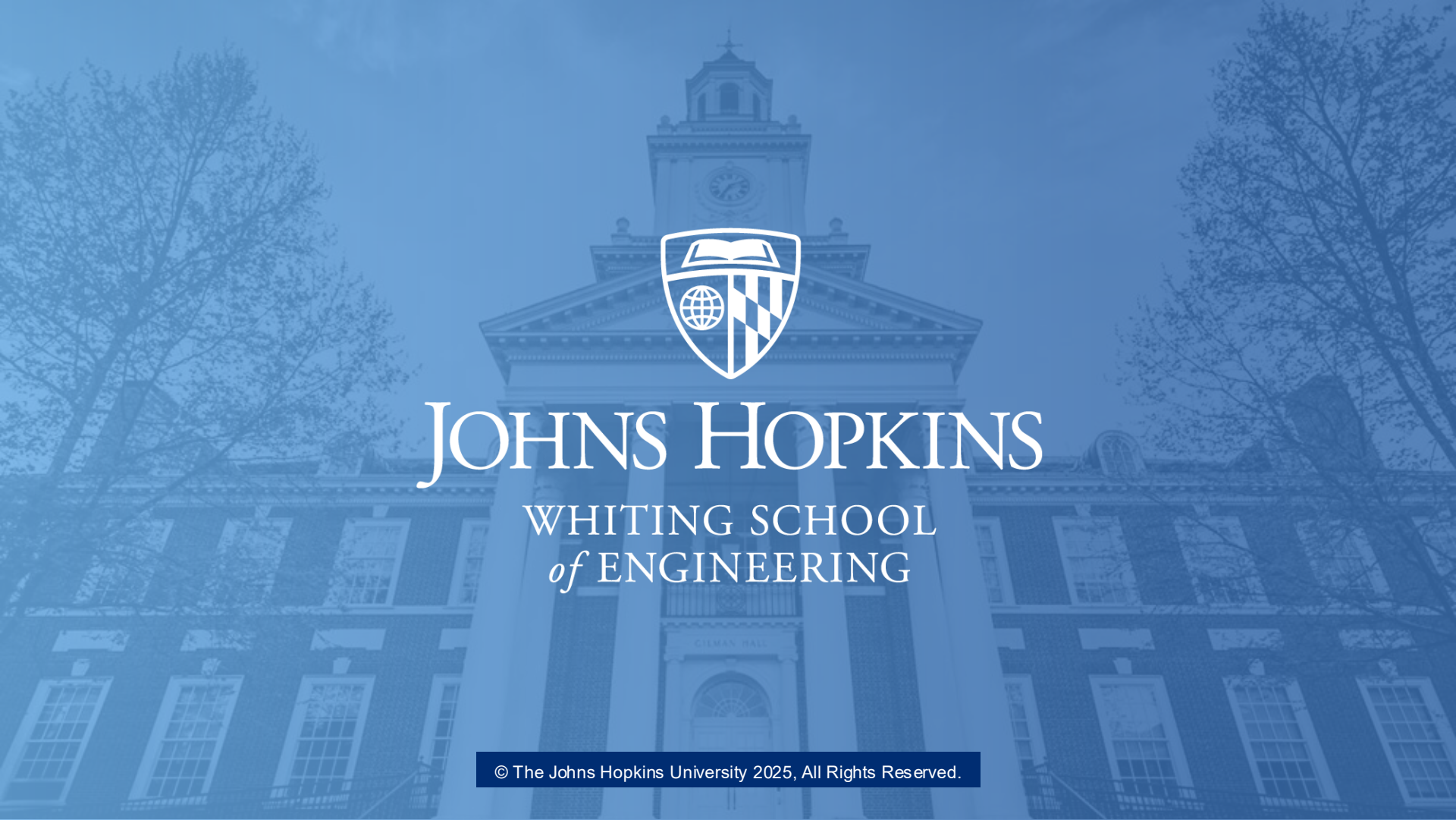
(b) MA1 generated with MV1 figures.

Number of Pairs	10		20		40		80	
	MA1	CF2	MA1	CF2	MA1	CF2	MA1	CF2
GPT-4.1-2025-04-14	90.48	78.57	83.33	57.14	88.10	52.38	92.86	33.33
Claude-3.7-Sonnet	97.62	73.81	73.81	45.24	85.71	38.10	85.71	9.52
Qwen-VL-Max-250408	97.62	83.33	88.10	47.62	90.48	2.38	73.81	7.14

The VisFactor Leaderboard



- Correlations between VisFactor and existing visual benchmarks



© The Johns Hopkins University 2025, All Rights Reserved.