Jigsaw Puzzles Boost AI Visual Reasoning

Originally published at https://www.unite.ai/jigsaw-puzzles-boost-ai-visual-reasoning/

New research indicates that AI models can get smarter at seeing by solving jigsaw puzzles. Rearranging scrambled images, videos, and 3D scenes helps them sharpen their visual skills without the need for extra data, labels, or tools.

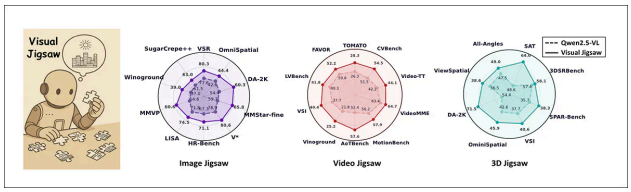
In the current scramble to push Multimodal Large Language Models (MLLMs*) ahead of the pack (or at least stay three releases in front of the nearest rival), there are few easy wins and no free lunches.

Though many of 2025's slew of impressive Chinese FOSS releases are reported to have lower development and running costs, occidental releases tend to throw *more* at the problem: more data volume, more inference grunt, more electricity (though not, as we recently noted, more actual human annotators, since that's too expensive even for the $trillion+ scale gen-AI revolution).

In the research literature, most supposedly 'free' approaches to the evolution of AI architectures tend to offer only minor incremental improvements; or else improvements in areas that are not the most critically pursued. Nonetheless, the search for hitherto undiscovered 'fundamental principles' that could accelerate the pace of development is too tempting to abandon.

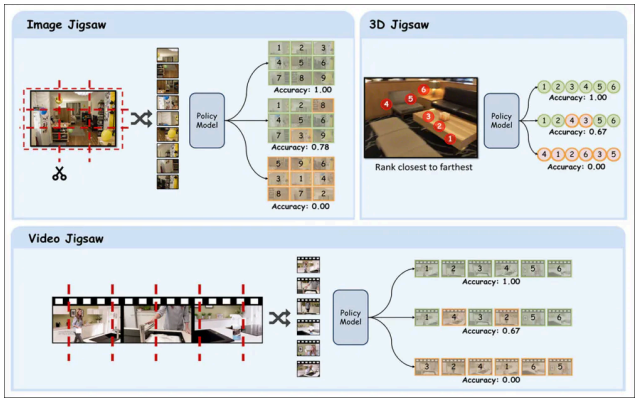
Picking up the Pieces

While not quite in that category, a new academic collaboration between Chinese institutions claims to have determined that making VLMs *solve jigsaw puzzles* improves their performance notably, even though this reinforcement learning approach has previously performed poorly in this area, and even though it requires no extra systems, ancillary models or other 'bolt-on' processes:



Visual Jigsaw is a self-supervised post-training framework that improves vision-centric skills in multimodal large language models. By training on jigsaw tasks across images, videos, and 3D data, the models gain sharper fine-grained, spatial, and compositional perception in images, stronger temporal reasoning in videos, and enhanced geometry-aware understanding in 3D scenes. The radar charts in the image above show consistent gains over base Qwen2.5-VL, with value scales adjusted for each benchmark for clarity. Source: https://arxiv.org/pdf/2509.25190

The system devised by the researchers is titled *Visual Jigsaw*, and involves training existing MLLMs on material that has been fragmented and randomly dispersed, like a jigsaw. The authors developed three modalities for this approach: image, video, and 3D (i.e., CGI-style meshes), and found that a moderate adaptation of the same process benefited all three domains:



A representation of the three Visual Jigsaw tasks. In the Image Jigsaw, an image is split into patches, shuffled, and the model predicts the correct layout; in the Video Jigsaw, clips are shuffled and the model restores their order in time; in the 3D Jigsaw, points with different depths are shuffled and the model ranks them from nearest to farthest. Model outputs are scored against ground truth, with partial credit given for partly correct solutions.

Visual Jigsaw's training method helps AI models improve at understanding visual information by making them reassemble these jumbled-up images, video clips, or 3D data points.

The process is based on words rather than images, and therefore there's no need for the model to generate images or use any extra visual components. This method fits into a system called *Reinforcement Learning from Verifiable Rewards* (RLVR), where the model gets rewarded for correct answers based on clear, automatic rules; and where, therefore, no human labeling is required.

This crucial fact is actually hard to glean from the new paper: that the system is assembling the puzzle *semantically*, through descriptions, and not in the shape-representative way that humans learn to solve such puzzles:

Prompt for Image Jigsaw

You are given nine shuffled image tiles that were created by slicing one image into a 3*3 grid.

Here are the tiles, each tagged with an index reflecting the current (shuffled) order in which they are shown:

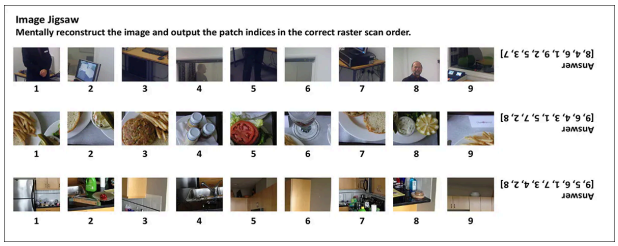
Tile 1: <image>
Tile 2: <image>
Tile 3: <image>
Tile 4: <image>
Tile 5: <image>
Tile 6: <image>
Tile 7: <image>
Tile 8: <image>
Tile 9: <image>

Task:
Mentally reassemble the original image, arranging the tiles into the correct 3*3 layout and provide the tile indices in raster-scan order (left-to-right, top-to-bottom), separated by commas.

Answer format example:
5, 1, 3, 7, 9, 2, 4, 8, 6
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

From the new paper's supplementary material, an example RL task illustrating the text-based nature of this adjunct learning process. The images that would have been shown to the MLLM are not depicted here.

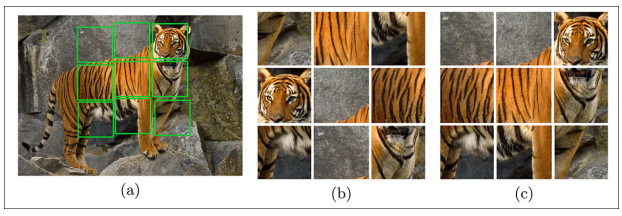
Though MLLMs deal extensively with vision-centric tasks, they are *language-based* architectures, not designed to generate images, videos or shape representations such as 3D meshes.



Examples from the image jigsaw task. Each row shows shuffled patches that the model must reorder into their original layout, with the correct arrangement shown on the right.

In any case, this type of training is done after the main learning phase, when the model already has some ability to understand images.

Prior approaches such as the 2017 Swiss paper *Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles* used this kind of reinforcement approach with rather less success, on convolutional neural networks (CNNs), a notably different kind of architecture in comparison to a modern MLLM.



From the 2017 release 'Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles ', an early example of the use of fragmentation as a reward-based challenge for a neural system. Source: https://arxiv.org/pdf/1603.09246

In tests, Visual Jigsaw led to what the authors claim to be consistent and measurable improvements across a wide range of benchmarks: the image jigsaw task improved fine-grained perception, spatial layout understanding, and compositional reasoning; the video jigsaw task enhanced the model's ability to track temporal sequences and reason about event order; and the 3D jigsaw task strengthened depth-based understanding and spatial reasoning using only RGB-D inputs.

Across all three modalities, the paper reiterates, the new method outperformed several competitive baselines, despite requiring no architectural changes, extra visual modules, or additional supervised data:

> *'Extensive experiments demonstrate substantial improvements in fine-grained perception, temporal reasoning, and 3D spatial understanding. Our findings highlight the potential of self-supervised vision-centric tasks in post-training MLLMs and aim to inspire further research on vision-centric pretext designs.'*

The new paper is titled *Visual Jigsaw Post-Training Improves MLLMs*, and comes from six researchers across Nanyang Technological University, Linköping University, and SenseTime Research. The paper is accompanied by a project site with live demos (and you can even load your own image into the image-based jigsaw demo). The code and weights for the project have been made generally available,

Method

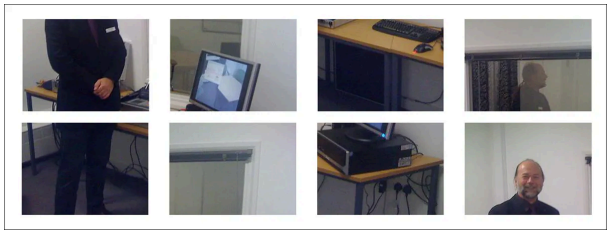
Though we'll take a look at the way information is split up to accommodate the three tested modalities, we should first consider the reward design for the new system.

The Visual Jigsaw approach scores model responses using a *graded reward*, not just a simple pass or fail. If the model predicts the exact correct order of the jigsaw pieces, it receives a full reward; if the answer is mostly right, but not perfect, the model gets partial credit, scaled by a *discount factor* to avoid overvaluing near-misses (this prevents the model from gaming the system by repeating guesses that are only partially correct).

Invalid answers, such as the 'cheat' of using the same number repeatedly, receive a score of zero. To encourage consistent formatting, the model must place its reasoning inside *<think>* tags and its final answer inside *<answer>* tags. It receives a small bonus if this format is used correctly.

Images

To create a puzzle for the *images* modality, a picture is first divided into a grid of patches by slicing it into equal-sized blocks:



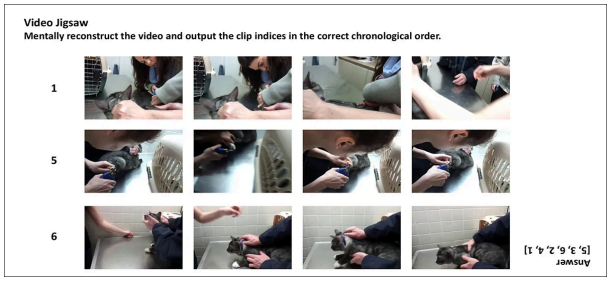
Examples of image-based patches for the system to solve.

The patches are laid out in a fixed order from top-left to bottom-right, as with the reading order on a page, before being scrambled with a random shuffle. The model is then exposed to this shuffled set of patches, and must figure out the original order by predicting the correct permutation that restores the original layout.

In training, 118,000 images from the COCO dataset were used, with each image yielding nine patches (i.e., 'puzzle pieces'). The prompt given to the system is shown earlier in this article (the image above with the caption beginning *'From the new paper's supplementary material'*).

Video

For the video jigsaw task, a video is chopped into a sequence of clips by slicing it evenly across time, and then shuffling the clip segments. The model is then shown this scrambled sequence, and must figure out their correct, original chronological order.



Truncated examples of the video puzzle challenge, from the paper's supplementary material.

The training for this modality used 100,000 videos from the LLaVA-Video dataset, with each video split into six clips. To stop the model from exploiting obvious frame-matching cues at clip boundaries, 5% of the frames at the start and end of each clip were trimmed.

Clips contained no more than 12 frames, each at a maximum resolution of 128x28x28 pixels. Videos shorter than 24 seconds were excluded.

The prompt for this task is shown below:

Prompt for Video Jigsaw

You are given four **shuffled** video clips that were created by slicing one original video into 6 equal-length temporal segments.

Here are the clips, each tagged with an index reflecting the current (shuffled) order in which they are shown:

Clip 1: <video>
Clip 2: <video>
Clip 3: <video>
Clip 4: <video>
Clip 5: <video>
Clip 6: <video>

Task:
1. Mentally reassemble the original video by arranging the clips in their correct chronological order (earliest segment first, latest segment last).
2. Output the clip indices in that order, separated by commas.

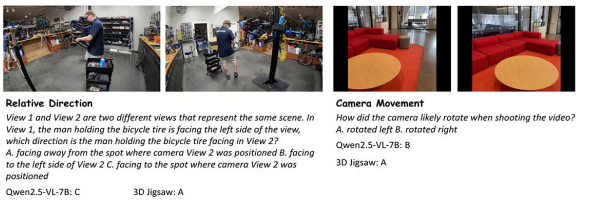
Answer format example:
2, 3, 1, 4, 6, 5
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

The reinforcement learning prompt presented to the MLLMs for the video task, without the clips that would have been presented to the MLLM..

3D Data

A full 3D jigsaw task would usually involve breaking up a 3D space (such as voxel blocks or point cloud fragments) into smaller chunks and training a model to reconstruct their original spatial layout.

However, the average MLLM is not equipped to handle raw 3D data directly, instead relying on semantically-interpreted image or video inputs. Therefore, to create a task that still taps into 3D reasoning while remaining compatible with current MLLMs, the authors introduced a more tractable variant using the aforementioned RGB-D images (i.e., 2D images that include depth information for each pixel).



Example questions from the 3D Spatial Understanding Benchmark used to evaluate performance on relative viewpoint and camera motion reasoning. The 3D Jigsaw model correctly infers both the spatial relation between two views of a scene and the likely direction of camera rotation, outperforming the Qwen2.5-VL-7B baseline.

From each RGB-D image, the model is given a shuffled list of points that originally had distinct depths, ranging from near to far, the goal being to recover their correct depth-based ordering using only the 2D image as reference:

Prompt for 3D Jigsaw

<image>
You are given an indoor RGB image. Six points are marked on the image with red circular labels (1, 2, 3, . . .).

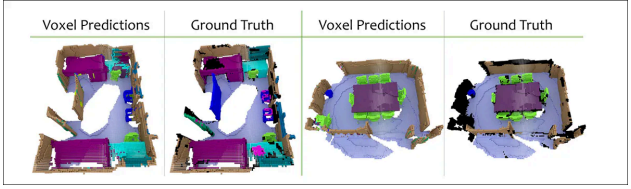
Your task is to order the points from closest to farthest relative to the camera, judging the distance based on the center of the red circular marker.

Answer with the ordered sequence of point numbers.
You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think></think> tags. The final answer MUST BE put within <answer></answer> tags.

The RL prompt for the 3D jigsaw.

Each point is marked on the image (which is shown to the model, but not visualized in the prompt example in the image directly above), and the model must predict which was closest, second closest, and so on, effectively reconstructing the original depth sequence without access to the raw depth values.

The 3D jigsaw task is trained on RGB-D images from the ScanNet dataset, using 300,000 samples created by selecting random sets of six depth points per image.



Examples of point clouds from the ScanNet dataset used in the 3D jigsaw. Source: https://arxiv.org/pdf/1702.04405

Each point was required to lie within a depth range of 0.1 to 10 meters, and to promote diversity, no two points in a set allowed to be closer than 40 pixels in the image plane, or less than 0.2 meters apart in depth.

Tests

For initial tests, the system utilized Qwen2.5-VL-7B-Instruct as the base multimodal model. Training used the Group Relative Policy Optimization (GRPO) algorithm, with both KL regularization and entropy loss removed.

For partial predictions, a discount factor of 0.2 was applied. Image jigsaw training used a global batch size of 256, while video and 3D jigsaws used 128. The learning rate was set to 1×10^{-6}.

For each prompt, the model generated 16 responses at a decoding temperature of 1.0. Both image and video jigsaw tasks were trained for 1,000 steps, while the 3D jigsaw task was trained for 800 steps.

Image Jigsaw

The image jigsaw model was tested across three categories of vision-centric benchmarks: for fine-grained perception and understanding, MMVP, the fine-grained perception subset of MMStar; MMBench; HR-Bench; V*; MME-RealWorld (lite); LISA-Grounding; and OVD-Eval.

For monocular spatial understanding, the benchmarks were VSR; OmniSpatial; and Depth Anything V2 (DA-2K). For compositional visual understanding, the tests employed Winoground and SugerCrepe++.

Three baselines were tested, all derived from Qwen2.5-VL-7B: ThinkLite-VL for multimodal reasoning; VL-Cogito for general vision and scientific tasks; and LLaVA-Critic-R1 for image perception.

All were evaluated using short answers only, since chain-of-thought (CoT) reasoning sometimes reduced performance.

	Fine-grained Perception & Understanding								Spatial Und (Mono)			Compositional Und	
Model	**MMVP**	**MMStar (fine-grained)**	**MMBench**	**HR-Bench-8K**	**V***	**MME-RealWorld**	**LISA-Grounding**	**OVD-Eval**	**VSR**	**OmniSpatial**	**DA-2K**	**Winoground**	**SugarCrepe++**
	test	fine	en_dev	test	test	lite	test	test	test	test	val	g-acc	test
ThinkLite-VL	55.33	59.95	84.19	68.12	76.96	46.17	73.70	35.78	78.09	42.60	58.46	35.25	61.49
VL-Cogito	55.33	56.64	82.98	69.62	79.58	47.63	72.26	35.78	79.82	44.29	56.43	38.25	63.59
LLaVA-Critic-R1	53.33	57.80	83.16	67.50	78.01	45.18	68.52	35.28	78.50	42.73	53.82	34.75	61.93
Qwen2.5-VL-7B	54.66	59.75	83.33	67.38	76.96	43.41	71.89	35.07	77.68	42.66	54.45	37.00	61.59
Image Jigsaw (SFT)	56.00	60.94	83.67	69.75	80.10	43.88	66.59	34.35	80.68	43.55	61.46	38.75	62.03
Image Jigsaw	60.66	65.81	84.45	71.13	80.63	45.96	74.54	36.49	80.36	44.49	60.35	39.00	63.02
(Gain)	+6.00	+6.06	+1.12	+3.75	+3.66	+2.55	+2.65	+1.42	+2.68	+1.83	+5.90	+2.00	+1.43

Evaluation results on image benchmarks. Image Jigsaw improved on the Qwen2.5-VL-7B base model across all task categories (i.e., fine-grained perception and understanding; monocular spatial understanding; and compositional visual reasoning), outperforming prior post-trained baselines.

Of the results for the image jigsaw, shown above, the authors state:

> *'[The image above] shows that our method consistently improves the vision-centric capabilities on the three types of benchmarks. These results confirm that incorporating image jigsaw post-training significantly enhances MLLMs' perceptual grounding and fine-grained vision understanding beyond reasoning-centric post-training strategies.*
>
> *'We attribute these improvements to the fact that solving image jigsaw requires the model to attend to local patch details, infer global spatial layouts, and reason about inter-patch relations, which directly benefits fine-grained, spatial, and compositional understanding.'*

Video Jigsaw

For video jigsaw, evaluation was conducted on AoTBench; Vinoground; TOMATO; FAVOR-Bench; TUNA-Bench; Video-MME; TempCompass; TVBench; MotionBench; LVBench; VSI-Bench; Video-TT; and CVBench.

Video-R1 was used as a baseline, having been trained with cold-start supervised fine-tuning followed by reinforcement learning for video understanding and reasoning. In this case, the evaluations included the full reasoning process, since this consistently produced better results than direct answers.

All models were restricted to 256x28x28 pixels, and tested across three frame settings – 16, 32, and 64:

Model	Frames	**AoTBench**	**Vinoground**	**TOMATO**	**FAVOR-Bench**	**TUNA-Bench**	**VideoMME**	**TempCompass**	**TVBench**	**MotionBench**	**LVBench**	**VSI-Bench**	**Video-TT**	**CVBench**
		vqa	group	test	test	test	wo subs	mc	test	val	test	test	mcq	test
Video-R1	16	45.06	9.40	27.29	49.47	53.00	56.62	70.19	51.80	55.82	34.53	34.34	42.95	47.50
Video-R1	32	47.53	10.20	27.29	49.90	54.26	59.88	71.77	53.54	56.12	38.61	35.11	42.63	48.10
Video-R1	64	48.68	10.60	27.36	50.51	54.33	60.85	72.59	53.43	56.09	38.80	36.61	42.74	48.69
Qwen2.5-VL-7B	16	45.52	12.60	25.87	48.54	53.14	57.44	71.77	49.94	55.56	33.51	32.79	38.39	47.70
Qwen2.5-VL-7B	32	49.48	18.20	26.34	49.34	54.88	60.70	72.59	51.96	56.47	39.19	35.34	41.57	49.60
Qwen2.5-VL-7B	64	52.41	21.80	26.35	50.86	55.79	63.44	72.84	53.74	56.29	40.35	37.74	42.25	51.50
Video Jigsaw	16	51.67	15.20	27.56	49.69	55.10	58.07	73.10	51.33	56.87	36.41	35.39	40.19	49.80
(Gain)		+6.15	+2.60	+1.69	+1.15	+1.96	+0.63	+1.33	+1.39	+1.31	+2.90	+2.60	+1.80	+2.10
Video Jigsaw	32	55.00	21.40	28.03	50.56	56.49	62.37	73.60	53.31	57.99	39.70	38.47	43.27	51.60
(Gain)		+5.52	+3.20	+1.69	+1.22	+1.61	+1.67	+1.01	+1.35	+1.52	+0.51	+3.13	+1.70	+2.00
Video Jigsaw	64	57.64	25.20	28.30	52.27	56.63	64.74	73.60	54.18	57.91	41.83	40.40	44.11	54.50
(Gain)		+5.23	+3.40	+1.95	+1.41	+0.84	+1.30	+0.76	+0.44	+1.62	+1.48	+2.66	+1.86	+3.00

Evaluation results on video benchmarks, with Video Jigsaw outperforming the baseline consistently across all tasks and frame settings.

Video Jigsaw produced consistent improvements across all video understanding benchmarks and frame settings, with gains especially strong on tasks requiring temporal reasoning and directionality, such as those in AoTBench, and also in cross-video reasoning benchmarks, such as CVBench:

> *'These results confirm that solving video jigsaw tasks encourages the model to better capture temporal continuity, understand relationships across videos, reason about directional consistency, and generalize to holistic and generalizable video understanding scenarios.'*

3D Data

For the 3D modality, the model was evaluated on SAT-Real; 3DSRBench; ViewSpatial; All-Angles; the aforementioned OmniSpatial; VSI-Bench; SPARBench (tiny); and the aforementioned DA-2K.

Model	SAT-Real	3DSRBench	ViewSpatial	All-Angles	OmniSpatial	VSI-Bench	SPARBench	DA-2K
	test	test	test	test	test	test	tiny	test
Qwen2.5-VL-7B	48.66	57.42	36.52	47.56	42.66	37.74	35.75	54.45
3D Jigsaw	64.00	58.13	38.62	49.06	45.99	40.64	38.31	71.56
(Gain)	+15.34	+0.71	+2.10	+1.50	+3.33	+2.90	+2.56	+17.11

Evaluation results on 3D benchmarks: 3D Jigsaw improved performance across depth comparison tasks such as DA-2K, as well as wider 3D perception benchmarks covering single-view, multi-view, and egocentric video inputs.

Here the authors state:

> *'[3D] Jigsaw achieves significant improvements across all benchmarks. Unsurprisingly, the largest gain is on DA-2K, a depth estimation benchmark that is directly related to our depth-ordering pre-training task. More importantly, we observe consistent improvements on a wide range of other tasks, including those with single-view (e.g. 3DSRBench, [OmniSpatial]), multi-view (e.g. ViewSpatial, All-Angles), and egocentric video inputs (e.g. VSI-Bench).*
>
> *'These results demonstrate that our approach not only teaches the specific skill of depth ordering but also effectively strengthens the model's general ability to perceive and reason about 3D spatial structure.'*

Conclusion

What is not entirely clear from this paper is the exact relationship between image and descriptions that is powering this improvement in MLLM performance.

At first glance, the process of learning through puzzle-solving seems very analogous to our own early forays and development. However, our own spatial interpretation of the world instinctively feels less semantic and related to language, and a deeper look at the paper reveals the extent to which language serves MLLM as an intriguing bridge between visual and semantic realities.

** Please note that the paper's authors prefer the less-common term 'Multimodal Large Language Models', shortened to 'MLLMs'. This is an emerging or infrequently-used term, and applies to models which can extensively reason and analyze images spatially, but which do not generate images. As new paradigms and models emerge, this lexicon is under constant revision.*

First published Thursday, October 2, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More