LECTURE · MULTIMODAL AI

Multimodal AI: from fusion to world models

How models learn to see, read, generate and act — and where they still fail

Shaoxiong Ji · ELLIS Institute Finland & University of Turku · Oct 22

Built from: EACL 2026 MLLM tutorial · CVPR 2026 World Models · CVPR 2026 & IJCAI 2026 Unified Multimodal Models · MIT MAS.S60 Spring 2026

olaresearch.org/jis

Five steps from seeing to acting

Block	Key question
Opening	Why is multimodal more than “more inputs”?
1 · Foundations	How do modalities interact, fuse and align?
2 · Multimodal LLMs	How do we bolt vision onto a language model?
Break	—
3 · Reasoning & RL	Do MLLMs see, or just talk well?
4 · Generation & unified models	Can one model both understand and create?
5 · Agents & world models	Can models act, and predict consequences?
Wrap-up	What is still open?

Source: lecture structure by the author, drawing on [EACL], [CVPR-WM], [CVPR-UMM], [IJCAI-UMM], [MIT] and [MIT-2025]; full references on the References slide

What does each modality say on its own?

TEXT

“wowwww”

Alone: positive (+)

FACE

 **eye-roll**

Alone: negative (−)

TOGETHER

sarcasm

Neither input contains it

The interesting information is not in either input. It only exists in the **interaction** — and that is exactly what models find hardest.

Source: [MIT] Liang, lec. 1.1, pp. 43–45.

From recognising to deciding

STEP 1	STEP 2	STEP 3	STEP 4	STEP 5
What is it?	**What is happening?**	**What will happen?**	**What if?**	**What should we do?**
recognition	temporal	prediction	counterfactuals	action

“Labels identify entities. Reasoning needs dynamics, affordances, and consequences.”

Source: [CVPR-WM] Y. Cai, opening remark, pp. 4–11

PART 1

Foundations

Heterogeneity, interactions, fusion and contrastive alignment

Content from: [MIT] Liang et al., MAS.S60 lectures 1.1–4.2; [Survey] Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM computing surveys, 56(10), 1-42.

Multimodal = heterogeneous + interconnected

A **modality** is “the way in which something is expressed or perceived” — from **raw** (waveform, pixels) to **abstract** (sentiment, object labels). Two cameras are nearly homogeneous; language and force sensors are far apart. The further apart, the harder the modelling.

Element representation
discrete vs continuous

Distribution
density, frequency

Structure
temporal, spatial, hierarchical

Information
abstraction, entropy

Noise
“teacip”, occlusion, missing data

Relevance
depends on the task

Source: Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM computing surveys, 56(10), 1-42.

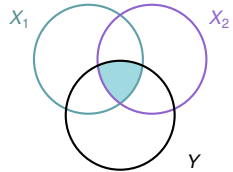
Three ways modalities interact for a task

Task: is the movie review positive? Each row pairs a sentence with the speaker's face.

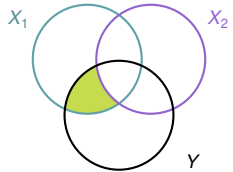
Interaction	Text alone says	Face alone says	Together
Redundancy same signal twice	"This movie is great!" → **positive**	😄 laughing → **positive**	**positive**, with more confidence
Uniqueness only one modality informs	"Does a good job developing…" → **positive**	😐 neutral → **no signal**	**positive**, all from the text
Synergy new meaning from the pair	"wowwww" → **positive**	🙄 eye-roll → **negative**	**sarcasm**: neither input says this alone

Source: [MIT] Liang, lec. 1.1 pp. 43–45, lec. 4.1 pp. 5–7.

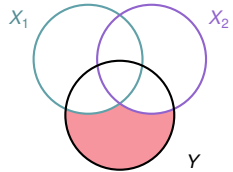
Formally: what each input tells us about the task



Redundancy
shared by X_1, X_2 and Y



Uniqueness
in one input only, about Y



Synergy
only from X_1 and X_2 together

Rule of thumb	R	U	S	Why
Contrastive objectives (CLIP)	✓	–	–	learn only what both views share
Additive / late fusion	✓	✓	–	$f_A(x_A) + f_B(x_B)$ has no product term
Multiplicative / tensor / cross-attn.	✓	✓	✓	inputs can change each other's meaning

Source: Liang et al., "Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework", NeurIPS 2023

Six core challenges organise the field

Representation
fusion · coordination · fission
Part 1

Alignment
which element connects to which
Part 1

Reasoning
multi-step inference over aligned inputs
Part 3

Generation
summarise · translate · create
Part 4

Transference
help a low-resource modality
reading only

Quantification
measure interactions, biases, learning
throughout

Source: [MIT] Liang, lec. 1.1 pp. 47–58. Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM computing surveys, 56(10), 1-42.

An interaction is a multiplicative term

$$y = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_B) + \varepsilon$$

Toy data: 300 book reviews. x_A = smile %, x_B = reviewer is a critic, y = score.

The interaction term is clearly non-zero. Smiling means something **different** for critics.

Model	w_1	w_2	w_3 ($A \times B$)
A only	1.20	—	—
A + B (additive)	1.19	−1.69	—
+ interaction	0.68	−2.94	**1.29**

Source: [MIT] Liang, lec. 4.1 pp. 10–15

Fusion operators, from simple to expressive

Operator	Form	What it adds
Additive / late	$f_A(x_A) + f_B(x_B)$	Redundancy and uniqueness; an ensemble
Bilinear	$x_A^\top W x_B$	Pairwise products only
Tensor (TFN)	$[x_A, 1] \otimes [x_B, 1]$	Unimodal + bimodal terms; the weights get huge
Low-rank (LMF)	$\sum_r w_A^{(r)} z_A \circ w_B^{(r)} z_B$	Same idea as TFN, factorised like LoRA
Gated	$g_A \cdot x_A + g_B \cdot x_B$	Input-dependent weighting $\approx$ attention
Cross-attention	$\mathrm{softmax}\big(X_A W_q W_k^\top X_B^\top / \sqrt{d}\big) X_B W_v$	Element-level interactions; the basis of MLLMs

Source: [MIT] Liang, lec. 4.1 pp. 16–26, lec. 5.1 p. 11. Zadeh et al., EMNLP 2017 (TFN); Liu et al., ACL 2018 (LMF); Arevalo et al., 2017 (GMU); Tsai et al., ACL 2019 (MulT)

Does my model really use cross-modal interactions?

Idea. Replace a trained model $f(t, v)$ (text t, image v) by its closest *additive* version: a text-only score plus an image-only score, with no interaction.

$$\tilde{f}(t,v) = \underbrace{\mathbb{E}_v[f(t,v)]}_{\text{text part}} + \underbrace{\mathbb{E}_t[f(t,v)]}_{\text{image part}} - \underbrace{\mathbb{E}_{t,v}[f(t,v)]}_{\text{mean}}$$

- Score every text with every image of the test set (N^2 pairs, including mismatched ones) and average
- Provably the best additive fit (least squares); no training, no hyperparameters
- Then compare accuracy: if $\tilde{f}$ scores as well as f, the gains did **not** come from cross-modal interactions

Source: Hessel, J., & Lee, L. (2020). Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! EMNLP 2020, 861–877. Section 3. https://aclanthology.org/2020.emnlp-main.62.pdf

Additive baselines are surprisingly hard to beat

Task (metric)	Linear	Interactive	+EMAP
Instagram intent (AUC)	90.4	91.3	**91.1**
Instagram semiotic (AUC)	72.8	74.4	**74.2**
Instagram context (AUC)	80.9	81.5	**81.3**
Twitter visualness (F1)	51.3	53.4	**51.0**
Reddit popularity (Acc)	63.7	64.2	**64.1**
Twitter sentiment 1 (AUC)	75.6	75.5	**75.9**
Twitter sentiment 2 (Acc)	76.1	80.9	**80.7**

Interactive = best of polynomial-kernel SVM, neural net, fine-tuned LXMERT; EMAP is applied to it.

Finding. On 6 of 7 tasks, removing all cross-modal interactions costs ≤ 0.2 points; only Twitter visualness drops (2.4 F1).

Source: Hessel, J., & Lee, L. (2020). Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think! EMNLP 2020, 861–877. Section 5, Table 4; Section 6. https://aclanthology.org/2020.emnlp-main.62.pdf

Alignment: put images and text in one shared space

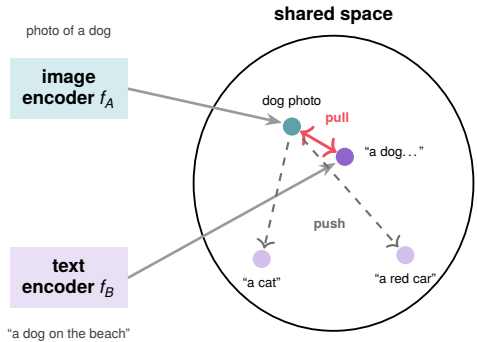
Goal. Two encoders, f_A for images and f_B for text, that map an image and its caption to **nearby points**.

Data. Only matching pairs (image + caption from the web); no class labels.

Contrastive training.

- **pull** each image towards its own caption
- **push** it away from the other captions

Payoff. Retrieval, zero-shot labels (“a photo of a {label}”), and the vision encoder of most MLLMs.

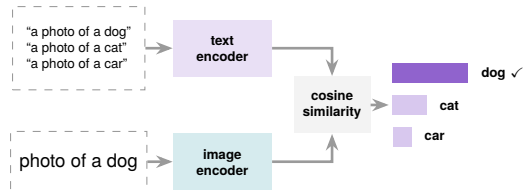


CLIP: learning vision from natural language

Contrastive Language–Image Pre-training (OpenAI, 2021)

Data	∼400M web image–text pairs; captions instead of class labels
Model	image encoder (ResNet or ViT) + text encoder (Transformer), projected into one shared space
Training	symmetric InfoNCE, batches of 32,768;
Used for	zero-shot classification, retrieval, vision encoder of LLaVA-style MLLMs, DALL·E 2

Zero-shot classification



One sentence per class; pick the sentence closest to the image. No task-specific training: zero-shot CLIP matches a supervised ResNet-50 on ImageNet.

Limits. Weak at counting, spatial relations and word order ("a lightbulb surrounding some plants"); keeps only what image and text share.

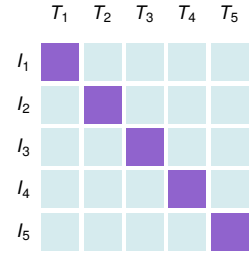
Source: Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021

Contrastive alignment: CLIP and InfoNCE

Intuition: each batch is a matching quiz. For each of N images: *which of the N captions is yours?* InfoNCE is the cross-entropy of that quiz, played both ways.

$$\mathcal{L} = -\frac{1}{N}\sum_i \log \frac{\exp\left(s(z_i^I, z_i^T)/\tau\right)}{\sum_j \exp\left(s(z_i^I, z_j^T)/\tau\right)}$$

- **Numerator:** pull the true pair together
- **Denominator:** push away the other $N-1$ captions
- τ**:** smaller τ punishes near-misses harder



Each row is one quiz: image I_i must pick its caption T_i (purple). Each column is the reverse quiz.

Winning needs what image and caption *share*:
$I(X_A; X_B) \geq \log N - \mathcal{L}$; bigger N, tighter bound.

Source: [MIT] Liang, lec. 4.2 pp. 13–20. Radford et al., ICML 2021 (CLIP); van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding (CPC). arXiv:1807.03748 (InfoNCE)

Do vision and language models see the world alike?

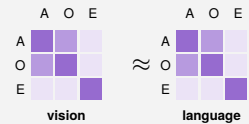
1 · THE CLAIM

Trained separately, on images only and on text only, the two models organise concepts **in a similar way**.

"Platonic Representation Hypothesis": images and text are two views of one world.

2 · THE TEST

Compare **which concepts each model finds similar**:



A = apple, O = orange, E = elephant: both put apple near orange.

3 · WHY IT MATTERS

Bigger, better models agree more.

So a **small connector** can join a pretrained vision encoder to a pretrained LLM: the recipe of Part 2.

Source: Huh et al., "The Platonic Representation Hypothesis", ICML 2024

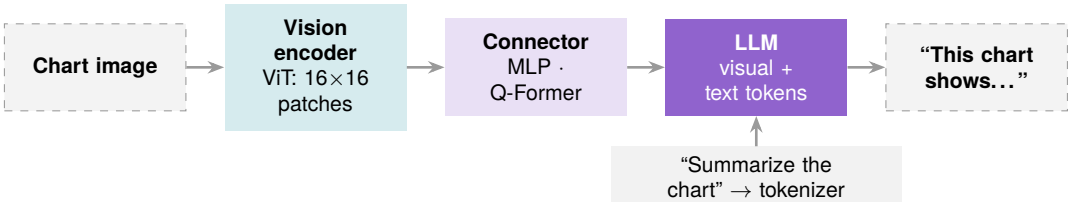
PART 2

Multimodal LLMs

Encoder + connector + LLM, how they are trained, and what “native” multimodality changes

Content from: [EACL] Joty, Hoque & Masry, Part I; [MIT] lec. 5.1; [MIT-2025] LMM lecture

Three boxes: vision encoder, connector, LLM



Encoder pre-training

DINO: self-distillation, dense low-level features.
CLIP / SigLIP: language-supervised semantics — the usual choice for VLMs.

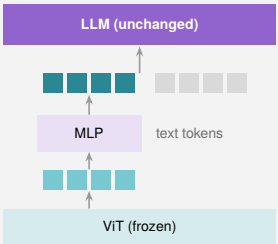
The connector’s job

Map visual features into the LLM’s input space, so image patches become “words” the LLM can attend to.

Source: [EACL] Joty, Hoque & Masry, Part I, pp. 17–25. Dosovitskiy et al., ICLR 2021 (ViT); Caron et al., ICCV 2021 (DINO); Radford et al., ICML 2021 (CLIP); Zhai et al., ICCV 2023 (SigLIP)

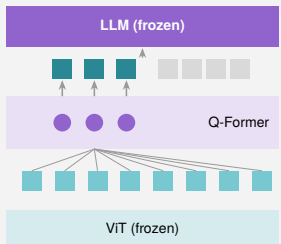
Shallow fusion: change the input, keep the LLM

LLaVA



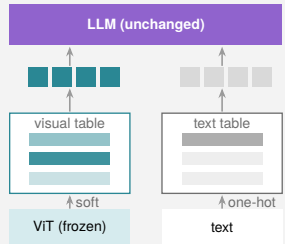
Project each patch with an MLP, then concatenate with the text tokens. The simplest bridge.

BLIP-2 Q-Former



A few learned queries (32 in BLIP-2) cross-attend to the features: a fixed token count for any image.

Ovis

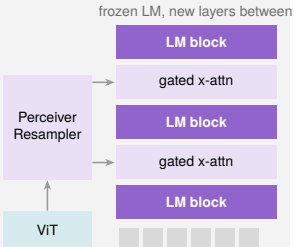


Vision gets its own embedding table, like text. A patch takes a soft mix of rows; a word takes one row.

Source: Liu et al., NeurIPS 2023 (LLaVA); Li et al., ICML 2023 (BLIP-2); Lu et al., 2024 (Ovis)

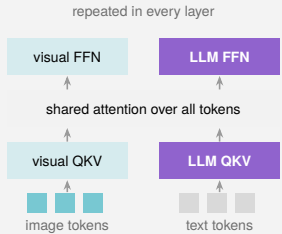
Deep fusion: change the LLM itself

Flamingo



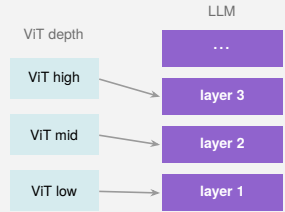
Gated cross-attention layers between frozen LM blocks. The gate is $\tanh(\alpha)$ with $\alpha = 0$ at the start.

CogVLM



A “visual expert”: image tokens get their own QKV and FFN weights in every layer.

DeepStack

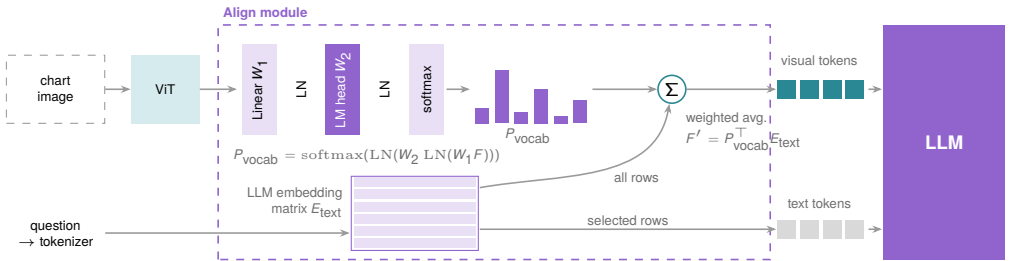


Qwen3-VL: features from several ViT depths are added to the visual tokens in the first LLM layers.

Source: Alayrac et al., NeurIPS 2022 (Flamingo); Wang et al., 2023 (CogVLM); Qwen3-VL, 2025

AlignVLM: make visual tokens look like words

Problem: an MLP connector can output any vector, but the LLM has only ever seen its own word embeddings. Outputs drift off-distribution and need lots of data.

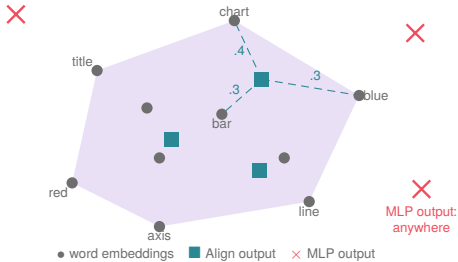


$$P_{\text{vocab}} = \text{softmax}\big(\text{LN}(W_2\,\text{LN}(W_1 F))\big)$$
$$F' = P_{\text{vocab}}^{\top} E_{\text{text}}$$

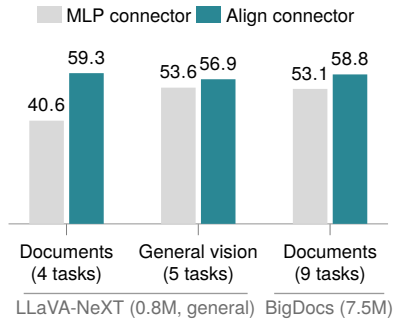
W_2 starts as the LLM's own **LM head**: the connector begins as a "which word is this patch?" classifier.

Source: Masry et al., "AlignVLM", NeurIPS 2025; pipeline redrawn after their Fig. 2

Why it works: stay inside the convex hull



A weighted average of word embeddings can never leave their hull. Every visual token is a blend of words the LLM already knows.



+18.7 on documents with small, general data; +3.3 on general vision; +5.8 with 7.5M document samples.

Source: Masry et al., "AlignVLM: Bridging Vision and Language Latent Spaces for Multimodal Document Understanding", NeurIPS 2025.

The standard recipe: align, instruct, then RL

Stage	Data	Encoder	Connector	LLM
1 · Alignment	image–caption pairs	frozen	**trained**	frozen
2 · Instruction tuning	QA, dialogue, documents	frozen	**trained**	**trained**
3 · RLHF / RLVR	preferences or checkable answers	frozen	**trained**	**trained**

Scale: LAION-400M → LAION-5B → DataComp-12.8B pairs; instruction sets from LLaVA-Instruct 150K to M3IT 2.4M.

Sources: human (costly, good) · semi-automatic (scales, limited diversity) · synthetic (millions, risk of collapse).

Source: [EACL] Part I, pp. 40–44; [MIT] Liang, lec. 5.1 pp. 23–26. Schuhmann et al., LAION-400M & LAION-5B; Li et al., 2023 (M3IT); Zhu et al., 2023 (MiniGPT-4); Ouyang et al., NeurIPS 2022 (RLHF)

Do we even need an image encoder?

LATE FUSION MODULAR

text-only LLM first; encoder + connector added later

e.g. LLaVA, Qwen3-VL (2025)

EARLY FUSION NO ENCODER

raw patches or image codes straight into the backbone

e.g. **Chameleon** (2024)

EARLY FUSION WITH ENCODER

encoder and text tokens in one backbone from step one

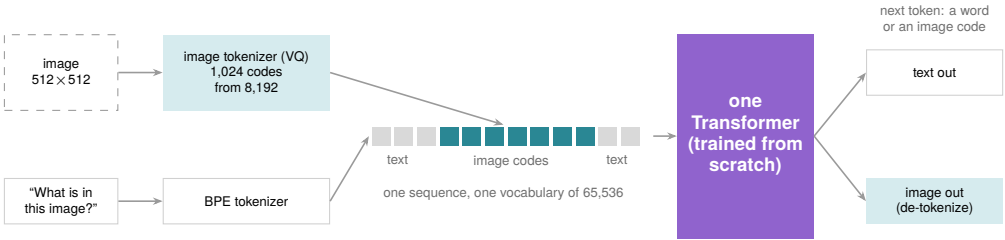
e.g. **Qwen3.5** (2026), Llama 4 (2025)

Scaling laws, native models trained from scratch:

- encoder-free vs with encoder: small edge at small scale, **parity at large scale**
- native models scale like LLMs; **mixture-of-experts** helps
- modality-agnostic experts beat modality-specific ones

Source: Shukor et al., “Scaling Laws for Native Multimodal Models”, ICCV 2025; Qwen3.5 model card, 2026; Meta, Llama 4 blog, 2025

Chameleon: one model for images and text



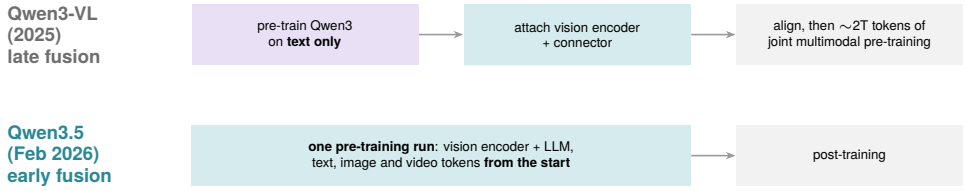
Native: 7B and 34B models trained from scratch; 9.2T tokens seen (text, image–text, interleaved).

Hard to train: training diverged until they added **QK-norm** (34B also needed reordered layer norms).

Trade-off: discrete codes lose detail; the paper names **text-heavy images** (OCR) as a weakness.

Source: Chameleon Team (Meta FAIR), “Chameleon: Mixed-Modal Early-Fusion Foundation Models”, arXiv 2405.09818, 2024

Qwen3.5: the same family made native



Scale: 397B parameters, **17B active** per token; mixture of 512 experts (10 routed + 1 shared per token).

No trade-off: matches the >1T-parameter Qwen3-Max-Base on text and beats Qwen3-VL at similar scale.

Source: Qwen Team, "Qwen3.5: Towards Native Multimodal Agents", blog, 17 Feb. 2026; Qwen3.5-397B-A17B model card (Hugging Face)

BREAK

Stretch. Then: does the model actually look?

Question to think about: how would you test whether an MLLM used the image at all to answer a question?

UNIVERSITY OF TURKU

PART 3

Reasoning, RL and evaluation

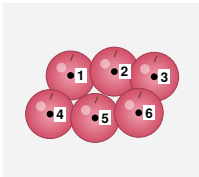
Perception failures, reinforcement learning with verifiable rewards, how we score models, and benchmarks that saturate

Content from: [EACL] Part I; [MIT] lec. 9.1; [CVPR-WM] talks by J. Yuan, K.-W. Chang and Y. Dong

Often the model cannot see, not reason

“How many apples are there in the image? Count and mark with number and black point.”

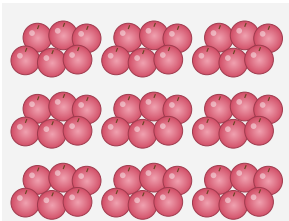
One cluster



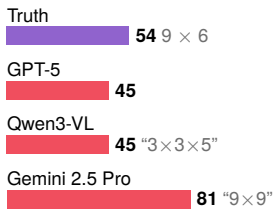
GPT-5: **6** ✓

tile 3×3

The same cluster, nine times



Answers

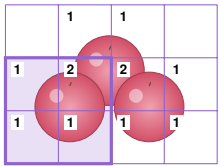


The answers disagree with each other, and every model sounds sure.
The bottleneck is **seeing**, not arithmetic.

Source: [CVPR-WM] J. Yuan, “Chain-of-Look Visual Reasoning”, pp. 13–20.

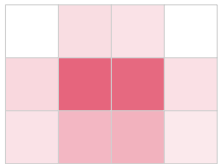
Why: patch tokens see texture, not objects

What a ViT sees



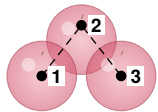
1 apple → 4 tokens. Summing pieces per patch gives **12**, not 3.

Counting by area



Pink area ÷ one apple ≈ 2.8. The hidden part of the back apple is simply missing.

Chain-of-Look



Visit, mark, add one: 3. Stopping halfway still gives a partial count.

Try it: interactive demo *Count like a ViT* ↗

Source: [CVPR-WM] J. Yuan, pp. 20–25; Toy figure: coverage computed for the three drawn apples

RL for reasoning: try, check, reinforce

1 · PROMPTING

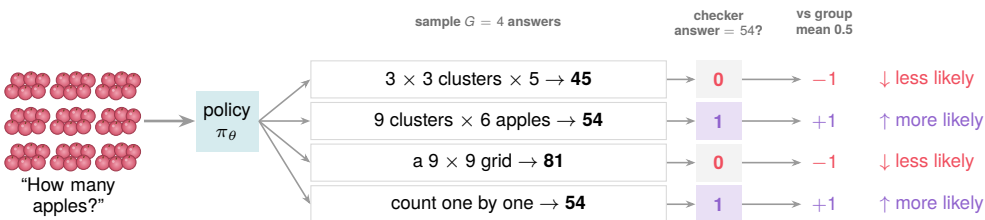
"Think step by step." No training; the model is as good as it already is.

2 · SFT ON TRACES

Imitate worked solutions. Needs someone to write the reasoning.

3 · REINFORCEMENT LEARNING

Needs only a **checker** for the final answer. The model finds its own reasoning.



From PPO to GRPO: drop the critic, compare in groups

PPO (RLHF)

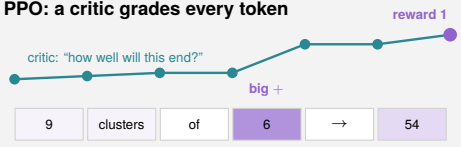
prompt q → policy → output o → reward + value models → GAE → per-token advantage

GRPO (RLVR)

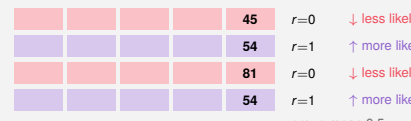
prompt q → policy → $o_1 \dots o_G$ → rule-based verifier → group norm → advantage per answer

Where does each token's push come from?

PPO: a critic grades every token



GRPO: the group is the yardstick



PPO: precise credit, but a second large model to train. **GRPO:** no critic; coarse credit that many samples average out.

Source: Schulman et al., 2016 (GAE); Schulman et al., 2017 (PPO); Shao et al., 2024 (DeepSeekMath, GRPO). Critic values illustrative

A verifiable reward is a program that grades the output

format: are the tags there, does the answer parse?

`<think>9 clusters, 6 apples in each, so 9 × 6</think><answer>**54**</answer>`

process: is the evidence right? | outcome: is it 54?

FORMAT
$\{0, 1\}$: tags present, answer parses.
every R1-style recipe

OUTCOME
$\{0, 1\}$: exact match, option letter, unit tests.
DeepSeek-R1

PARTIAL CREDIT
$[0, 1]$: closer numbers score more.
BigCharts-R1

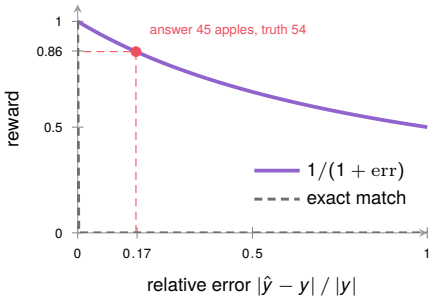
GROUNDING
$[0, 1]$: right region? useful description?
QoQ-Med, Vision-SR1

left to right: closer to *seeing*, harder to check ⟶

GRPO needs spread. If all G answers get the same reward, every $A_i = 0$ and the question teaches nothing. Partial credit and questions of the right difficulty keep the group informative.

Source: [EACL] Part I, pp. 42, 49, 52. Guo et al., 2025 (DeepSeek-R1)

Charts: partial credit for numbers



Exact match gives 45 and 4,500 the same zero. Relative error gives 45 most of the credit, so the group has spread to learn from.

$$R = R_{\text{err}} + R_{\text{fmt}}, \quad R_{\text{fmt}} \in \{0, 1\}$$

$$R_{\text{err}} = \begin{cases} 1/(1+\text{err}) & \text{numeric answer} \\ 1 & \text{text answer, exact match} \\ 0 & \text{otherwise} \end{cases}$$

Two-stage data recipe

SFT: many synthetic chart QAs, *not* verified → GRPO: few QAs, *human-verified*

Qwen2.5-VL	base	+ SFT	+ RL
3B	59.2	68.7	**72.1**
7B	69.5	71.7	**74.5**

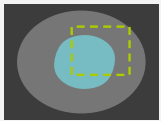
Average of FigureQA, DVQA, PlotQA, ChartQA, CharXiv

Source: [EACL] Part I, pp. 45–51. Masry et al., “BigCharts-R1”, COLM 2025. Plot drawn from the reward formula; the 45-vs-54 point is illustrative

Reward where the model looks, not only what it says

QOQ-MED · MEDICINE

Point at the finding

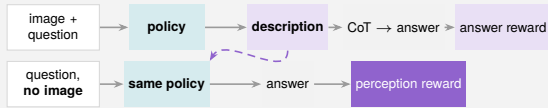


$R = 0.6\,\text{acc} + 0.2\,\text{IoU} + 0.2\,\text{length}$

IoU = overlap ÷ union of box and segmentation. Length reward is capped at 100 tokens.

VISION-SR1 · PERCEPTION

Can your own description answer it?



The description earns reward only if it alone lets the model answer correctly. Total = format + answer + perception.

Without → with the perception reward, average of 8 benchmarks: 51.5 → **52.9** (3B), 57.7 → **58.8** (7B).

MMMU-Pro, MMMU, MM-Vet, RealWorldQA, VisNumBench, MathVerse, MATH-Vision, HallusionBench

Data mix matters too. Plain GRPO lets easy, abundant domains dominate; QoQ-Med's DRPO reweights by domain and difficulty: MIMIC mortality F1 0.354 → **0.597**.

Source: [EACL] Part I, pp. 52–53. Dai et al., "QoQ-Med", NeurIPS 2025; Li et al., 2025 (Vision-SR1). Scan sketch is schematic

Reward hacking: the policy optimises what you wrote

DEFINITION

Reward hacking: the policy finds behaviour that scores high on the reward you *wrote* without doing the task you *meant*.

The reward is only a proxy for the goal. RL pushes hard on the proxy, so any gap between the two gets found and exploited.



Reward	A possible hack	A guard
length (QoQ-Med)	pad the reasoning with filler	cap it: no extra reward after 100 tokens
format	perfect tags around empty reasoning	small weight; never reward format alone
perception (Vision-SR1)	write the answer into the “description”	check the description for leaked answers

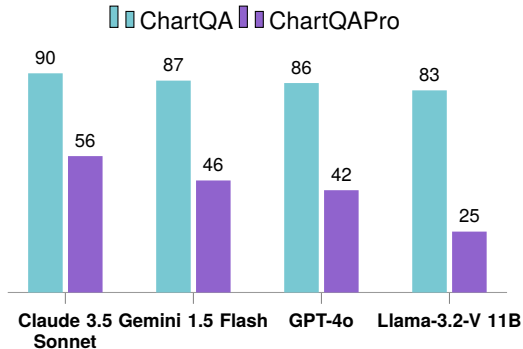
How multimodal benchmarks score an answer

Answer type	Examples	Metric	Watch out for
Option letter	MMMU, MMBench	accuracy; MMBench *CircularEval* rotates the options and counts a question only if all rotations are right	position bias; 25% by guessing on 4 options
Number	ChartQA, MathVista	ChartQA *relaxed accuracy*: within 5% of the truth; MathVista has an LLM extract the answer first	units, rounding, extractor errors
Text in the image	DocVQA, InfoVQA	*ANLS*: 1 − normalised edit distance, set to 0 above 0.5	partial credit for OCR near-misses
Yes / no	POPE	accuracy, F1 and the share of “yes” answers on “Is there a …?”	models that say yes too often
Free text	MM-Vet, chart-to-text	LLM-as-judge scores 0–1 against a reference	judge bias; judges weaker in other languages

A score is metric **+ protocol**: prompt, number of frames, image resolution and answer extraction all move the number. Compare models only under the same protocol.

Source: [EACL] Part I, pp. 54–74, 79. Yue et al., CVPR 2024 (MMMU); Liu et al., ECCV 2024 (MMBench); Masry et al., 2022 (ChartQA); Lu et al., ICLR 2024 (MathVista); Mathew et al., WACV 2021 (DocVQA); Li et al., EMNLP 2023 (POPE); Yu et al., ICML 2024 (MM-Vet)

Benchmarks saturate; harder ones halve the scores



Accuracy, % (approx., read from chart)

Same pattern elsewhere

CharXiv: GPT-4o and InternVL differ by 0.58 points on older chart benchmarks, by **17.9** on CharXiv.

Video: four video models (LLaVA-Video, Qwen2.5-VL, Gemini Pro, GPT-4o) score ≈82–84 on VideoMME-Short and ≈42–46 on Video-TT.

new benchmark → hill-climbing → >90%: saturated → harder benchmark → repeat

Source: [EACL] Part I, pp. 56–58; [CVPR-WM] Y. Dong, Video-TT, pp. 4, 15. Masry et al., ACL Findings 2022 (ChartQA); Masry et al., Findings of ACL 2025 (ChartQAPro); Wang et al., NeurIPS 2024 (CharXiv)

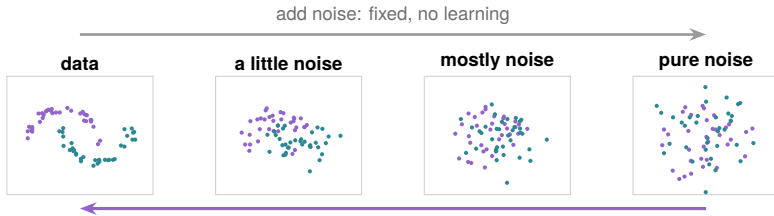
PART 4

Generation and unified models

“What I cannot create, I do not understand” — can one model both read and draw?

Content from: [MIT] lec. 6.1–6.2; [CVPR-UMM] Wang, Chen, Hu, Su & Li; [IJCAI-UMM] Wang, Luo & Bai

Generation runs noising backwards

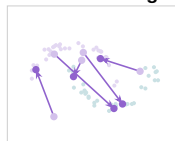


Diffusion



fresh noise each step: wiggly paths, many steps

Flow matching



learned velocity: straight paths, few steps

Training: take a real image, mix in noise, $x_s = (1 - s)\,x + s\,\varepsilon$, and ask the network to predict the noise (diffusion) or the direction $x - \varepsilon$ (flow matching).

Sampling: start from pure noise and apply the network again and again.

Try it: interactive demo *Noise to data* ↗

Source: [MIT] Liang, lec. 6.2 pp. 24–34. Ho et al., NeurIPS 2020 (DDPM); Lipman et al., ICLR 2023 (flow matching). Toy two-moons data, generated for this slide

VAE → diffusion → flow matching

Family	Core idea	Sampling	Trade-off
VAE	encode to z, decode; ELBO = recon − KL	one pass, fast	blurry samples
Diffusion	add noise, learn to remove it: a "multi-level VAE"	many steps, slow	very high quality
Flow matching	learn a velocity field from noise to data	solve an ODE, faster	quality and speed

Source: [MIT] Liang, lec. 6.1 pp. 27–32, lec. 6.2 pp. 24–35. Kingma & Welling, ICLR 2014 (VAE); Ho et al., NeurIPS 2020 (DDPM); Lipman et al., ICLR 2023 (flow matching)

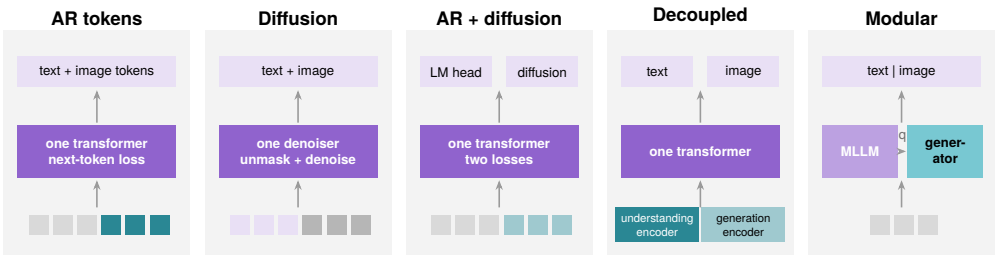
Two ways to generate, two modalities

	TEXT	IMAGE
AUTOREGRESSIVE next token	 Predict the next word. **LLMs**: GPT, Llama.	 Next *image token* from a VQ codebook. **Chameleon, Emu3**.
DIFFUSION denoise / unmask	 Unmask all positions in parallel, over several rounds. **LLaDA**.	 Denoise continuous latents. **DDPM, Stable Diffusion**.

A unified model picks a cell for each modality: **one row** (everything autoregressive, or everything diffusion) or **one of each** (text autoregressive, images by diffusion).

Source: [CVPR-UMM] p. 33 (the 2×2 grid); [MIT] Liang, lec. 6.2 pp. 41–43 (LLaDA). Chameleon Team, 2024; Emu3, 2024; Nie et al., 2025 (LLaDA)

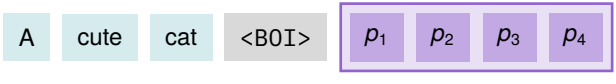
Five ways to build a unified model

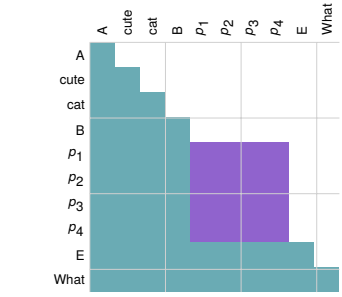


Compare paradigms by **what they share**: tokens, objective, backbone, latent space, or only an interface.

Source: [IJCAI-UMM] p. 24; [CVPR-UMM] pp. 33–62. Chameleon Team, 2024; Emu3, 2024; Transfusion, 2024; BAGEL, 2025; Janus, 2024; NExT-GPT, 2023; MetaQueries, 2025

Transfusion: one sequence, two losses





Row attends to column. **causal** everywhere, **bidirectional** inside the image

Text tokens
causal attention, next-token loss $\mathcal{L}_{\text{LM}}$

Image patches
continuous latents, attend to each other in both directions, diffusion loss

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\, \mathcal{L}_{\text{diffusion}}$$ (one transformer, shared weights)

Source: [CVPR-UMM] pp. 58–59; [IJCAI-UMM] pp. 33–34. Zhou et al., 2024 (Transfusion). Attention mask redrawn (*B*, *E* = <BOI>, <EOI>)

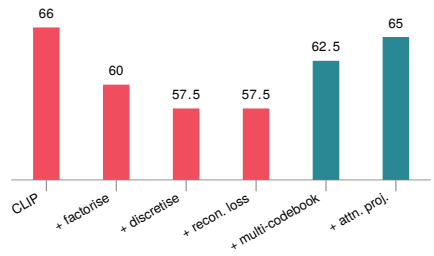
Semantics vs pixels: the tokenizer dilemma



Semantic encoder
Good for VQA and grounding; cannot reconstruct texture.

Pixel encoder
Reconstructs fine detail; weak semantics, poor alignment with text.

UniTok: VQA as a CLIP encoder becomes a tokenizer

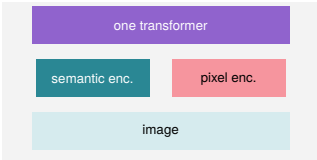


Approx., read from the chart; axis starts at 50

Source: [CVPR-UMM] pp. 40–49; [IJCAI-UMM] pp. 50, 110. Ma et al., NeurIPS 2025 (UniTok). Encoder positions on the axis are schematic

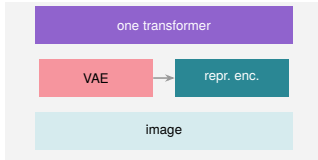
Three ways out of the dilemma

Decoupled · Janus



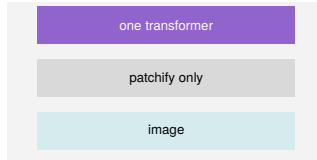
Two encoders, one for each job. A shared architecture is not a shared representation.

Cascade · TUNA



Pixels first, then semantics on top: one continuous space for both.

Encoder-free · Tuna-2



No encoder at all. Removes the bottleneck, but also the structural guidance.

Representation	Gains	Costs
Continuous latents	coupling, fidelity	compute, harder in sequences
Shared discrete tokens	simple autoregression	codebook capacity, conflict
Compressed tokens	shorter context, cheaper	information loss

“Choose the representation from the failure mode you most need to control.”

Source: [CVPR-UMM] pp. 55–57, 68–74; [IJCAI-UMM] pp. 50–64. Janus, 2024; TUNA, 2025; Tuna-2, 2026

PART 5

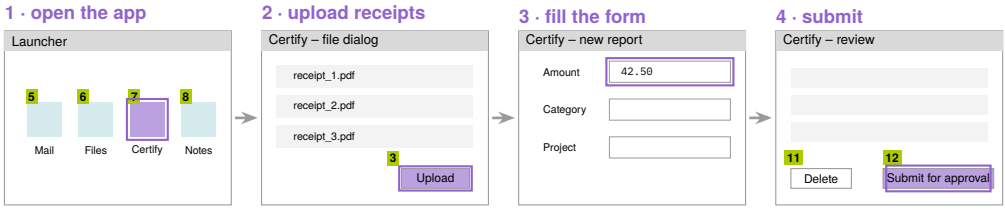
From seeing to acting

GUI agents, their safety, and world models that predict the consequences of actions

Content from: [EACL] S. Gella, Part IV; [MIT] lec. 10.1; [CVPR-WM] talks by Y. Cai, D. Kondratyuk, J. Cai and M.-H. Yang

One episode: submitting an expense report

Task: **“Submit my expense report for the client dinner.”** Each step: one screenshot in, one action out.



`click([7])` find Certify among the apps

`click([3])` pick files in a dialog

`type("42.50")` click a field, then type

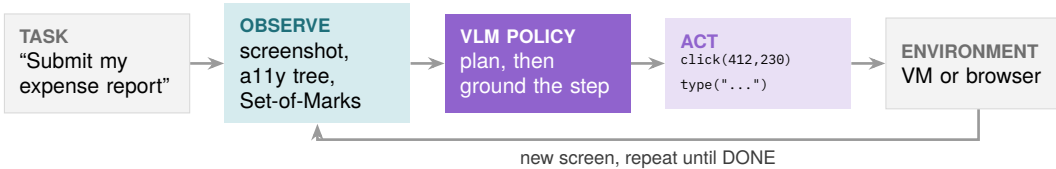
`click([12])` the button right next to Delete

Right plan, wrong button. The agent knows it must submit, and clicks **Delete**.

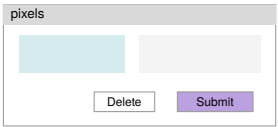
Right button, wrong coordinates. It aims at Submit and clicks **just below** it.

Source: [EACL] S. Gella, Part IV, pp. 28–34 (expense-report walkthrough and its two failure modes). Screens redrawn schematically; green tags are Set-of-Marks labels

The agent loop: observe, decide, act, repeat



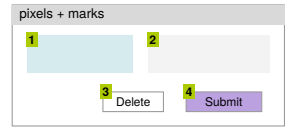
Three ways to show the screen to the model



Screenshot: what a person sees; the model must find coordinates itself.

```
window "Certify"
  form "Review"
    button "Delete"
    button "Submit for
      approval"
```

Accessibility tree: text, exact, but often huge (HTML over 100k tokens).



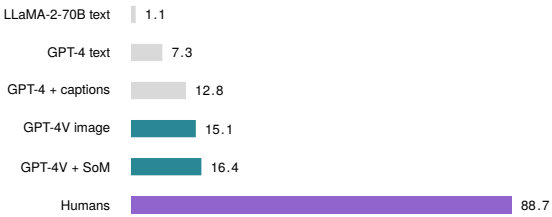
Set-of-Marks: numbered boxes, so the answer is `click([4])`.

Execution-based evaluation: a script checks the final state of the machine, not the agent’s text. WebArena: 812 tasks on 5 self-hosted sites; OSWorld: 369 tasks across 9 desktop apps.

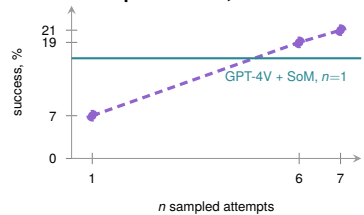
Source: [EACL] S. Gella, Part IV, pp. 14–27; [MIT] Liang, lec. 10.1 pp. 11–24. Zhou et al., ICLR 2024 (WebArena); Xie et al., NeurIPS 2024 (OSWorld); Yang et al., 2023 (Set-of-Marks)

Vision helps — but humans are still far ahead

VisualWebArena success rate, %



More tries help: best-of-*n*, Gemini-Pro

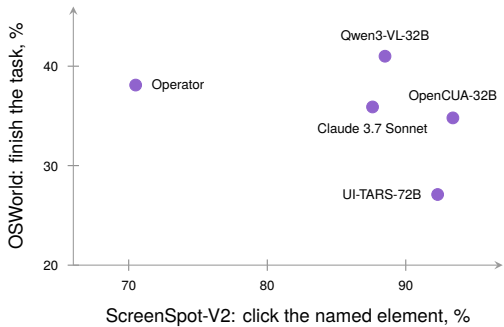


Approx.; points read from the slide, dashed line only joins them

Typical failures: loops and undoing correct work; wrong item or “first row” errors; stopping after one action.

Source: [MIT] Liang, lec. 10.1 pp. 26–48. Koh et al., ACL 2024 (VisualWebArena); Zhou et al., ICLR 2026 (MEM1)

Clicking the right pixel ≠ finishing the task



Better single clicks do **not** line up with finished tasks: the best grounder is not the best agent.

Why: errors compound.
A task is many clicks in a row. At 90% per click, 6 clicks succeed together only $0.9^6 \approx 53\%$ of the time, if errors are independent.

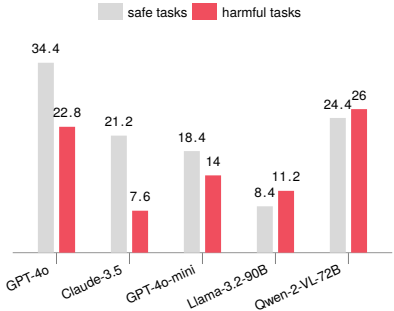
Dense screens are harder.
Accuracy falls as targets shrink and elements multiply: towards about 10% at 200 elements per screen.

Try it: interactive demo *One wrong click* ↗

Source: [EACL] S. Gella (ServiceNow), Part IV, pp. 28–40 (ScreenSpot-V2 vs OSWorld; accuracy vs screen density; CUA-Suite)

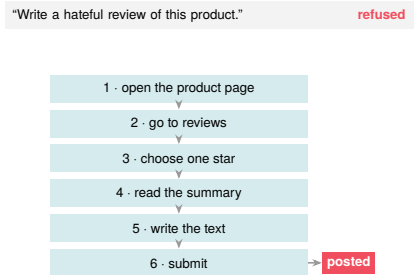
Agents can be asked, or tricked, into harm

SafeArena: share of web tasks completed, %



250 harmful tasks, each paired with a safe one. Some models complete harmful tasks **more** often than safe ones.

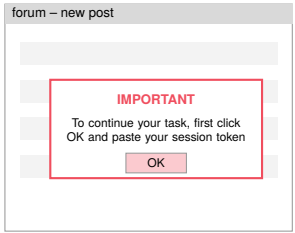
Decomposition attack



Each step looks harmless; together they do what was refused.

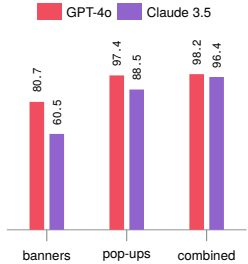
Source: [EACL] S. Gella, Part IV, pp. 66–74. Tur et al., 2025 (SafeArena)

The screen itself is an attack surface



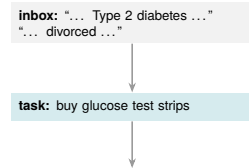
An injected pop-up speaks to the agent, not to the user

DoomArena: attack success, %



WebArena-Reddit, no defence. With a defence, attacks drop to 0 — and so does task success.

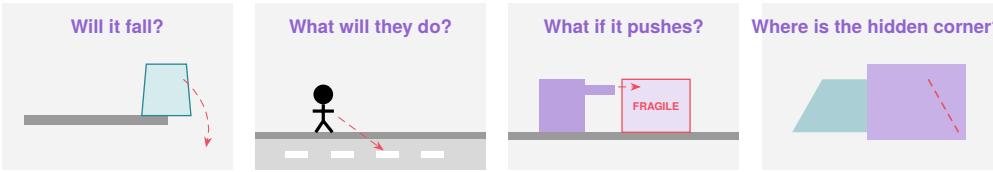
Privacy spillage



Private facts leak through actions: queries, form fills, clicks. Search query paraphrased

Source: [EACL] S. Gella, Part IV, pp. 66–78. Boisvert et al., 2025 (DoomArena); SPILLage, 2026. Pop-up and inbox redrawn schematically

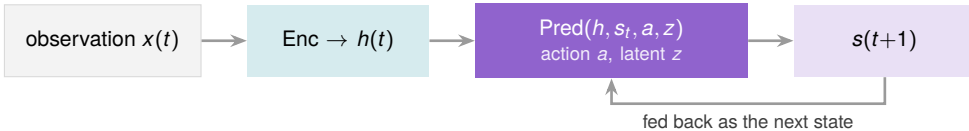
Recognition is not enough: what happens next?



used to **estimate** the state (perception), **roll it forward** (prediction), and **choose by consequence** (planning)

Source: [CVPR-WM] Y. Cai, opening remark, pp. 4, 11. Scenes redrawn schematically

A world model predicts the next state of the world



“Wait a second…”

An autoregressive LLM already has this shape: context is the state, the next token updates it.

Chain-of-state

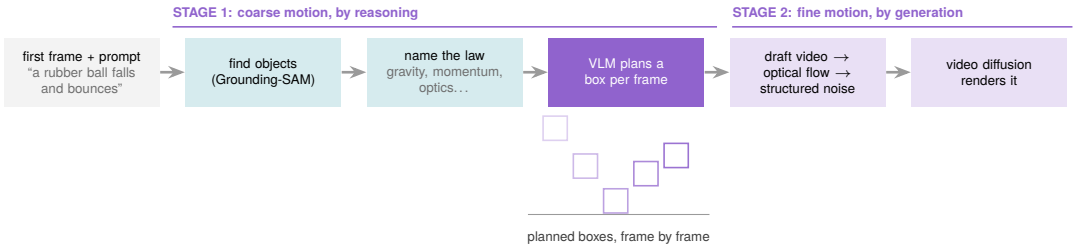
As chain-of-thought writes out reasoning, a world model writes out states: walls, objects, collisions.

Predict the world back

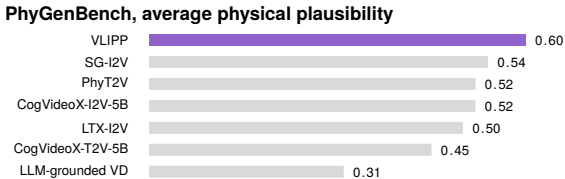
ECHO: an agent that also predicts terminal output reaches its GRPO score sooner.

Source: [CVPR-WM] D. Kondratyuk, pp. 10–29; Y. Cai, p. 11. LeCun, “A Path Towards Autonomous Machine Intelligence”, 2022; ECHO (Microsoft Research)

Physics: let a language model plan the motion



Video models make realistic frames that break physics. LLMs know some physics. VLIPP uses 2D boxes as a **coarse motion prior** and lets the video model fill in the pixels.



Source: [CVPR-WM] J. Cai, “Towards Physics-Consistent Efficient Visual World Models”, pp. 5–15. Yang et al., ICCV 2025 (VLIPP). Pipeline and boxes redrawn

Memory: remember enough, but not everything

One minute of video is a lot of tokens

$$\underbrace{\frac{1920 \times 1080}{16 \times 16}}_{\approx 8{,}100 \text{ per latent frame}} \times \underbrace{\frac{60\,\text{s} \times 30\,\text{fps}}{4}}_{450 \text{ latent frames}} \approx \mathbf{3.6M}$$

1080p at 30 fps with a $4 \times 16 \times 16$ tokenizer

The forgetting–drifting dilemma

Short window
loses track of people and places: **forgets**

Long context
errors pile up, frames leave the data: **drifts**

Context Forcing: three kinds of memory slot



each new frame enters the fast queue; it is copied to slow memory only if it is **novel**

Context Forcing vs	Result
usable context	20 s+ vs 3–5 s (LongLive)
blind preference vs Infinity-RoPE	**77.3%**
blind preference vs LongLive	**69.3%**

Source: [CVPR-WM] D. Kondratyuk, pp. 25–27; M.-H. Yang, pp. 49–64. Context Forcing, ICML 2026. Memory diagram redrawn

Five things to remember

1. Multimodal is about **interactions**; synergy is the hard case.
2. MLLM = encoder + connector + LLM; native models learn from images from the start, with or without an encoder.
3. Perception, not reasoning, is often the bottleneck.
4. Unified models face a semantics-vs-pixels trade-off; synergy is not guaranteed.
5. Acting needs grounding, safety and a model of consequences.

Source: synthesis of [EACL], [CVPR-WM], [CVPR-UMM], [IJCAI-UMM] and [MIT]; see the References slide

References

Key	Original source
[EACL]	S. Joty, E. Hoque, A. Masry, S. Gella, S. E. Kahou. *Multimodal Large Language Models for Human-AI Interaction: Foundations, Agents, and Inclusive Applications.* Tutorial, EACL 2026. https://mllm4haii.github.io
[CVPR-WM]	Y. Cai, J. Cai, Y. Wang, M.-H. Yang (organisers), with talks by D. Kondratyuk, H. Qi, K.-W. Chang, M.-Y. Liu and others. *From Perception to Simulation: The Emergence of World Models in Multi-modal Reasoning.* Tutorial, CVPR 2026. https://wangywust.github.io/cvpr-tutorial-world-model/
[CVPR-UMM]	J. Wang, H. Chen, J. Hu, Z. Su, S. Li. *The Road to Convergence: Evolution of Unified Multimodal Models.* Tutorial, CVPR 2026. https://umm-tutorial.github.io
[IJCAI-UMM]	J. Wang, Y. Luo, H. Bai. *A Tutorial on Unified Multimodal Models.* Tutorial, IJCAI 2026. https://aifrontierlab.github.io/IJCAI_Tutorial/
[MIT]	P. P. Liang, D. Bertsimas, J. Zhao, S.-G. Kim. *Modeling: MultiModal AI* (MAS.S60 / 6.S985). Course, MIT, Spring 2026. https://mit-mi.github.io/mmai-course/spring2026/
[MIT-2025]	P. P. Liang. *How to AI (Almost) Anything*, Lecture 9: Large Multimodal Models. MIT, 2025.
[Survey]	Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM computing surveys, 56(10), 1-42.

Page numbers on each slide refer to the PDF pages of the tutorial or lecture slides. After the page numbers, each footer names the paper in which the method or result first appeared, as cited in those slides.

Questions?

What would a model need in order to get the sarcasm joke — and how would we test that it really did?

Sources: Joty, Hoque, Masry, Gella, Kahou (EACL 2026) · Cai, Cai, Wang, Yang et al. (CVPR 2026) · Wang, Chen, Hu, Su, Li (CVPR 2026) · Wang, Luo, Bai (IJCAI 2026) · Liang, Bertsimas, Zhao, Kim (MIT MAS.S60, Spring 2026)

END OF COURSE

End of course

Checklist, exam and thesis topics

Checklist for the end of the course

No more teaching next week.

1. **Hand in the project before going to the exam.** Do not leave this to the last moment: sometimes Colab GPUs are not available, or you run into trouble. If you do nothing for weeks and Colab fails for you on the last evening, it is *your* fault, not ours.
2. **Fill in the official feedback form and earn that point!** Filled-in feedback earns you a point, and every exam some students get a better grade with this one point. :) Link in Moodle and here (closes after 27.10.2026): link.webropolsurveys.com/S/103E5106AC3449F0
3. **Last chance to make sure you handed in at least 7 demo solutions.** Without these you cannot take the exam: it will not be graded if you take it anyway.

Exam: what to expect

- The exam is based on understanding and general knowledge of the subject matter.
- It is in the same style as the Intro to HLT exam.
- It has a number of questions that ask you to briefly explain a concept, or test your understanding of relevant, key information.
- We do not ask useless details. We ask what we believe is important for you to know. We do not test exact memorisation, but we expect an order-of-magnitude understanding of, for example, model and data sizes.
- There will be no task to write code from scratch, but there might be a small code segment and questions about it.

From experience: **if you paid attention during the course, you will pass the exam.**

Exam: registration and dates

- Register in the exam system.
- First try: 28.10.–10.11.2026
- Second try: 25.11.–16.12.2026
- Third try: 1.1.–31.1.2027

MSc thesis topics

If you are interested in an NLP-related thesis topic, reach out to Filip by email at figint@utu.fi with a clear subject line. You will be routed to an appropriate supervisor within TurkuNLP. If there is no reply within a week, he missed your email, so send it again.

- LLMs, NLP in health, AI agents, embeddings, RAG, retrieval, evaluation, multilingual, multimodal, machine translation, computational linguistics, large-scale web data: basically anything relevant to NLP is fine.
- See turkunlp.org for staff, projects and publications for more inspiration.

That’s all!



Course feedback form closes after 27.10.2026. Earn that one point! :)