CS 8803-VLM

Vision-Language Models & Multimodal Agents

Zsolt Kira · Fall 2026

Website: https://faculty.cc.gatech.edu/~zk15/teaching/AY2026_cs8803vlm_fall/index.html
Canvas: Fall 2026 course: https://gatech.instructure.com/courses/525870
Ed Discussion: https://edstem.org/us/courses/103116/discussion (linked on Canvas)

WELCOME!

Georgia Tech

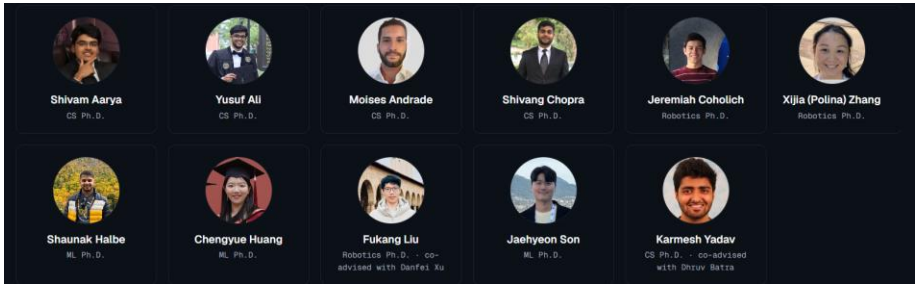
Zsolt Kira

zkira@gatech.edu
Associate Professor
School of Interactive Computing

Georgia Tech Machine Learning

RIPL
Robotics Perception and Learning

- Joined as Assistant Professor in 2018
- **Research Interests:** Intersection of deep learning and robotics, focusing on robustness and decision-making in an open world.
- **Group**
  - **Also a number of M.S. students per year**



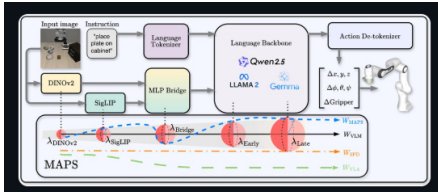
- **Teaching:**
  - CS 8803-VLM Vision-Language Models & Multimodal Agents (Fall),
  - CS 7643 Deep Learning (Spring)

Georgia Tech

Zsolt Kira

RIPL
Robotics Perception and Learning

zkira@gatech.edu
Associate Professor
School of Interactive Computing
Georgia Institute of Technology



Robust Finetuning of Foundation Models

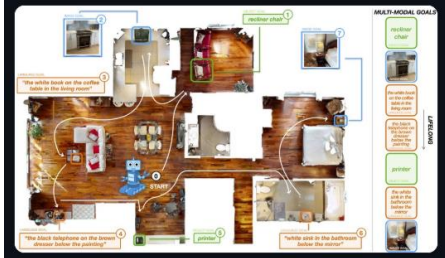
Adapting large pretrained models to a target task (including for decision-makingin VLAs) without destroying the out-of-distribution generalization. We develop projection- and geometry-based constraints on how far finetuning may move the weights, and benchmarks to measure progress.

- **MAPS** C. Huang et al. · CVPR 2026 arXiv Project Code
- **The Geometry of Robustness** S. Chopra et al. · CVPR 2026 arXiv ★ Highlight
- **Grounding Descriptions in Images (GRAIN)** S. Halbe et al. · WACV 2026 arXiv
- **Directional Gradient Projection (DiGraP)** C. Huang et al. · ICLR 2025 arXiv
- **FRAMES-VQA** C. Huang et al. · CVPR 2025 arXiv Code
- **Rethinking Weight Decay** J. Tian et al. · NeurIPS 2024 arXiv
- **Fast Trainable Projection (FTP)** J. Tian et al. · NeurIPS 2023 arXiv Code
- **Trainable Projected Gradient (TPGM)** J. Tian et al. · CVPR 2023 arXiv Code

Memory, Verification, and Reasoning in Agents

We build methods and benchmarks that isolate memory from perception, memory architectures that stay tractable over tens of thousands of steps, and verifiers that resist agreement bias.

- **FindingDory** K. Yadav et al. · ECCV 2026 arXiv Project Code
- **Self-Grounded Verification** M. Andrade et al. · ICLR 2026 arXiv
- **Reasoning Errors in VLM Verifiers** J. Cha et al. · ICLR Workshop 2026 OpenReview
- **EVE** Y. Ali et al. · ICRA Workshop 2026 arXiv PDF Talk PDF
- **Memo** G. Gupta et al. · NeurIPS 2025 arXiv



Simulation, Real-to-Sim-to-Real, and Benchmarks

Simulators and benchmarks for embodied AI, and the pipelines that carry policies between simulation and the real world. This includes personalized real-to-sim capture and viewpoint-robust transfer from fixed-camera data.

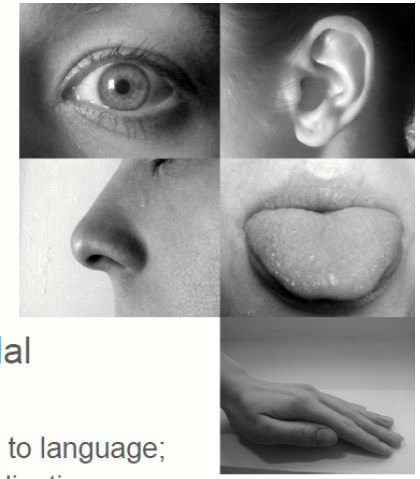
- **Sim2real Image Translation** J. Coholich et al. · ICRA 2026 arXiv Project Code
- **EmbodiedSplat** G. Chhablani et al. · ICCV 2025 arXiv Video
- **Habitat 3.0** X. Puig et al. · ICLR 2024 arXiv Project
- **GOAT-Bench** M. Khanna et al. · CVPR 2024 arXiv Project Code
- **Seeing the Unseen** R. Ramrakhya et al. · CVPR 2024 arXiv Code
- **HomeRobot** S. Yenamandra et al. · CoRL 2023 arXiv Code

- What are Vision-Language & Multi-Modal Models?
- Why now — and what changed since 2023
- The arc: tokens → VLMs → reasoning → agents → action → world models
- Logistics
- Q&A

Agenda for Today

Georgia Tech

Why does multimodality matter?



A range of very good reasons:

- Faithfulness: Human experience is multimodal
- Practical: The internet & many applications are multimodal
- Data efficiency and availability:
  - Efficiency: Multimodal data is rich and “high bandwidth” (compared to language; quoting LeCun, “an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts”), so better for learning?
  - Scaling: More data is better, and we’re running out of high quality text data.

Multimodality is one of the main frontiers of the new foundation model revolution.

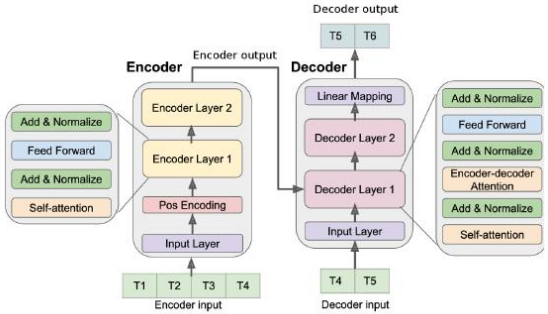
Slide by Douwe Kiela

Homogenization of Deep Learning

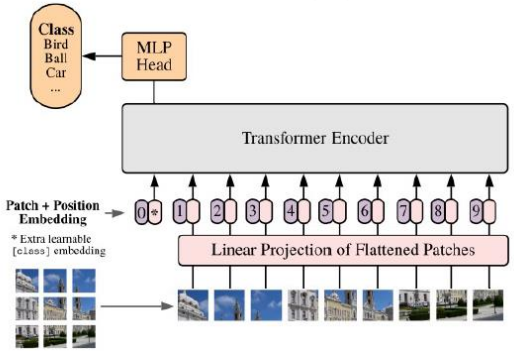
Homogenization is the **consolidation** of methodologies for building machine learning systems across a wide range of applications.

- Enabled by modular, plug-n-play nature of neural networks and training
- Consequence: Multi-modal, unified architectures, unified tasks (next-token prediction)

Example: The Transformer Models (Vaswani et al., 2017)



Transformer Models originally designed for NLP



Almost identical model (Visual Transformers) can be applied to Computer Vision tasks

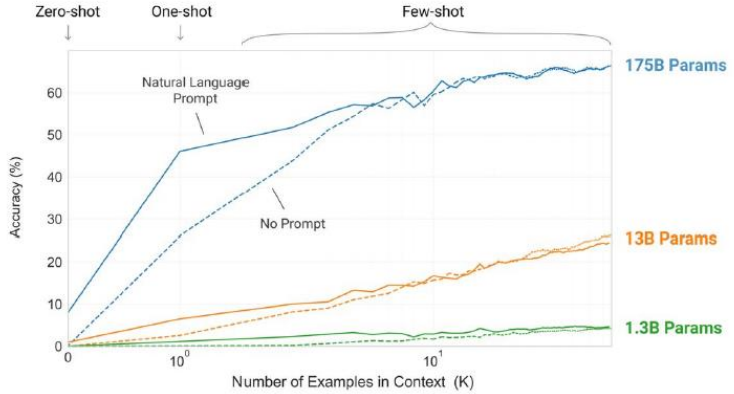
Emergence of new behaviors

Emergence means that the behavior of a system is implicitly induced rather than explicitly constructed. For Deep Learning, emergence is often induced by larger model & more data.

Example: Compared to GPT-2's 1.5B parameter model, GPT-3's 175-billion model permits "prompting" and "in-context learning", i.e., adapting to a new task simply by describing task.

Example input (prompt):
Ask it to translate French to English

$$\underbrace{\text{maison} \rightarrow \text{house}, \text{chat} \rightarrow \text{cat}, \text{chien} \rightarrow}_{\text{prompt}} \underbrace{\text{dog}}_{\text{completion}} .$$



Bommasani et al., On the Opportunities and Risks of Foundation Models

Course Progression

1	2	3	4	5	6
Week 1	Week 2	Weeks 3–5	Weeks 6–9	Weeks 9–11	Weeks 11–13
Pixels & words	**Aligned representations**	**Vision-language models**	**Reasoning & reliability**	**Agents**	**Action & world models**
Patches and sub-words become one token stream.	Vision and language share an embedding space.	A language model that can see, and generate.	Think before answering; know when you are wrong.	Perceive, plan, and take actions in software.	Act in the physical world; imagine what follows.
Transformer · ViT	**CLIP → SigLIP 2 · DINOv3**	**Flamingo · LLaVA → Qwen3-VL**	**MM-CoT · RLVR · POPE**	**ViperGPT · Mind2Web · OSWorld**	**RT-2 · OpenVLA · DreamZero**

Shifts in past few years

01 Native Multimodal

CIRCA 2023

A frozen CLIP encoder with an LLM with a small adapter. 224–336 px, one image, short context.

FALL 2026

Trained multimodal from the start: dynamic native resolution, interleaved image/video/text, long context, dense grounding heads.

02 Post-Training for Reasoning

CIRCA 2023

Chain-of-thought with prompting, ineffective visual reasoning.

FALL 2026

Post-training with verifiable rewards enables test-time compute for accuracy and visual reasoning.

03 From answering to acting

CIRCA 2023

The output was a caption or an answer, benchmarked on a static dataset.

FALL 2026

The output is a click, a keystroke, an API call. The benchmark is a live browser or desktop with success/failure at the end.

04 Action is just another output modality

CIRCA 2023

Robot policies trained separately from language.

FALL 2026

The same next-token machinery emits action chunks at control rate — and world models predict what those actions will cause.

PART 1

Building a VLM

Tokens, alignment, instruction tuning, and generation — Weeks 1–5.

1

Week 1

Pixels & words

Patches and sub-words become one token stream.

Transformer · ViT

2

Week 2

Aligned representations

Vision and language share an embedding space.

CLIP → SigLIP 2 · DINOv3

3

Weeks 3–5

Vision-language models

A language model that can see, and generate.

Flamingo · LLaVA → Qwen3-VL

4

Weeks 6–9

Reasoning & reliability

Think before answering; know when you are wrong.

MM-CoT · RLVR · POPE

5

Weeks 9–11

Agents

Perceive, plan, and take actions in software.

ViperGPT · Mind2Web · OSWorld

6

Weeks 11–13

Action & world models

Act in the physical world; imagine what follows.

RT-2 · OpenVLA · DreamZero

How/what should we train?

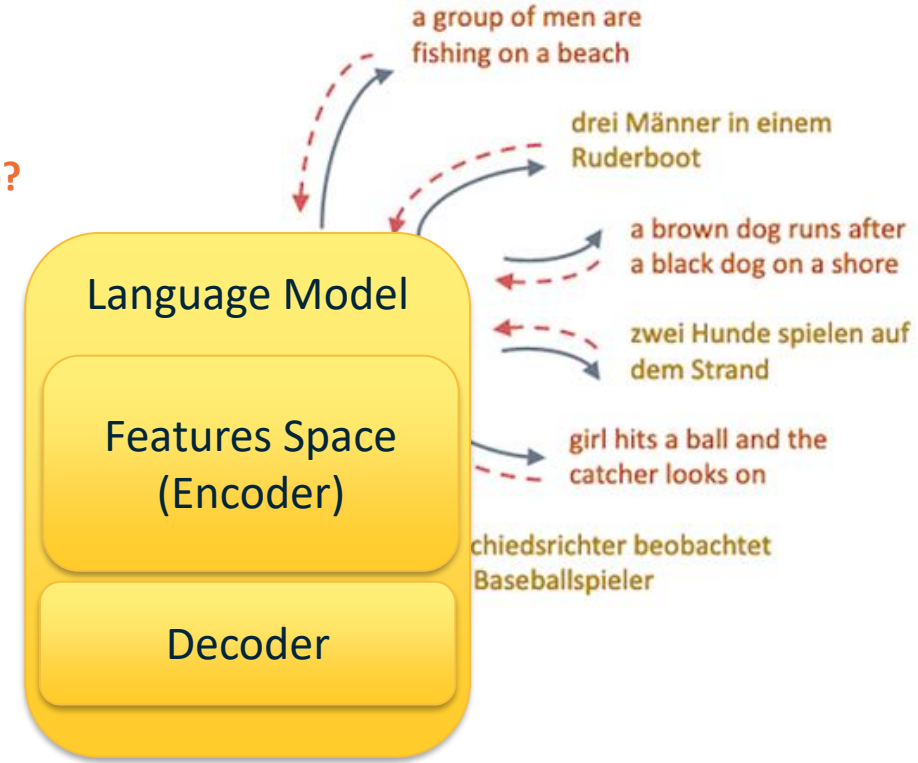
Using what data?

What tasks can we do?

Visual Feature Space

?

What should the interface be?

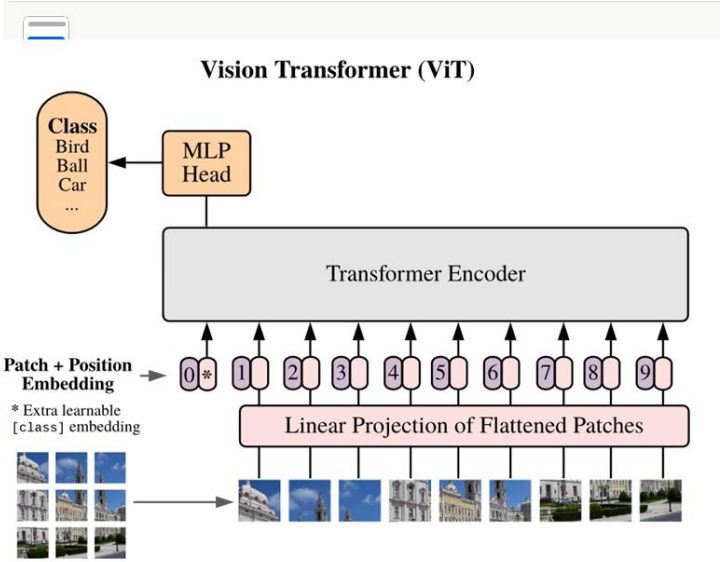


VLMs: Considerations

Georgia Tech

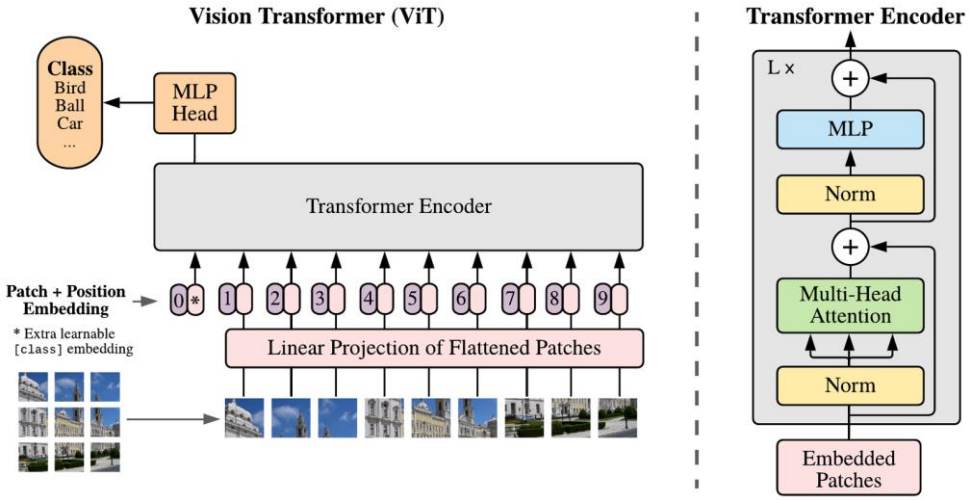
Potential ways of representing an image?

- Image encoder
  - Any architecture: ResNet, Vision transform (ViT)
  - Randomly initialized, SL/SSL pre-trained



Slide by Vicente Ordóñez

Georgia Tech

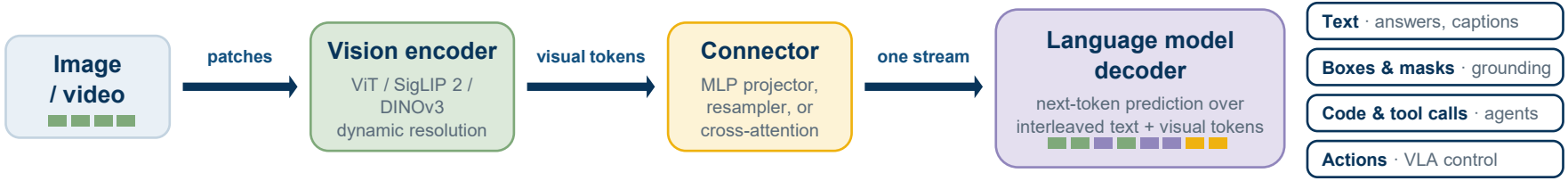


Dosovitskiy et al., An Image is Worth 16×16 Words, ICLR 2021 — figure from the paper.

- A 224×224 image at 16×16 patches is 196 tokens.
- Position embeddings: Information about 2-D layout (learned)
- (Mostly) same block, objective, optimizer as text.
- Frontier: native / dynamic resolution

Images As Tokens

The Three Components



Design Choices

Frozen or trained encoder? How many visual tokens per image? Early, mid, or late fusion?

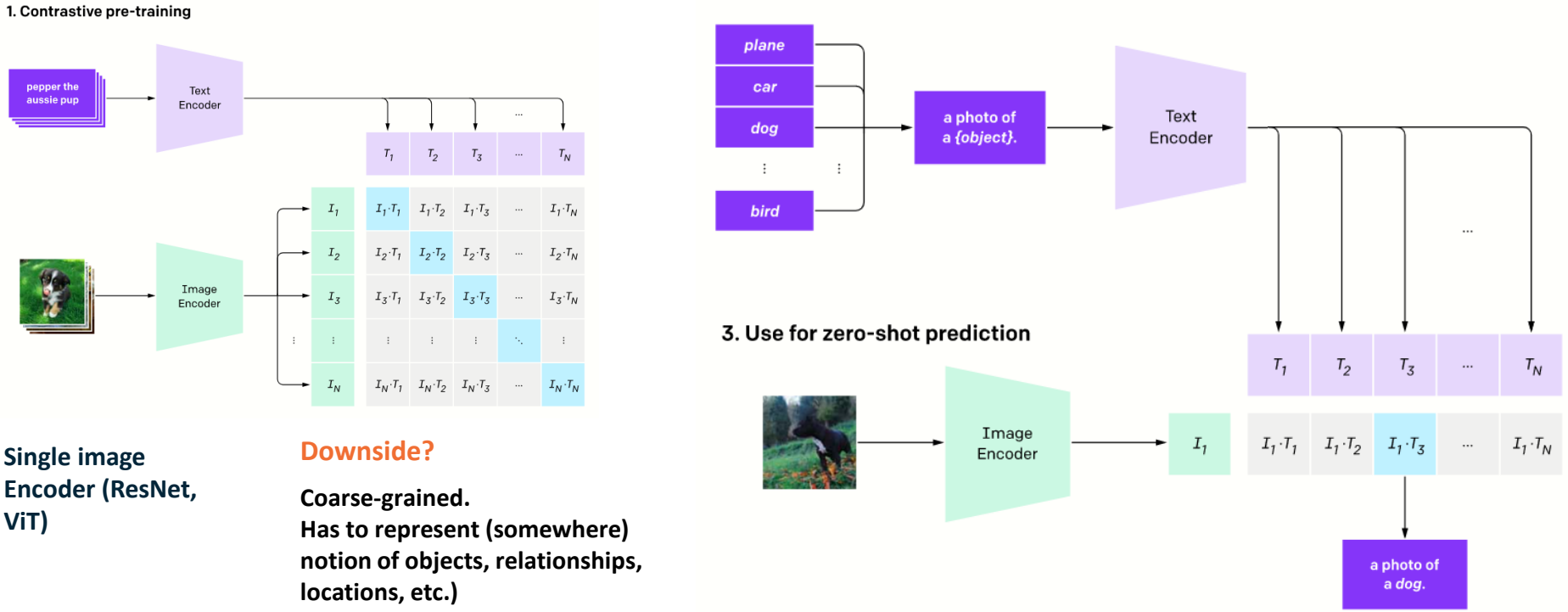
Constraints

Token budget vs. resolution; long video; the connector is a bottleneck on fine detail.

Why it generalises

Any output that can be written as tokens — a mask, a click, a joint velocity — trains the same way.

Method of alignment: Contrastive Learning



Single image Encoder (ResNet, ViT)

Downside?

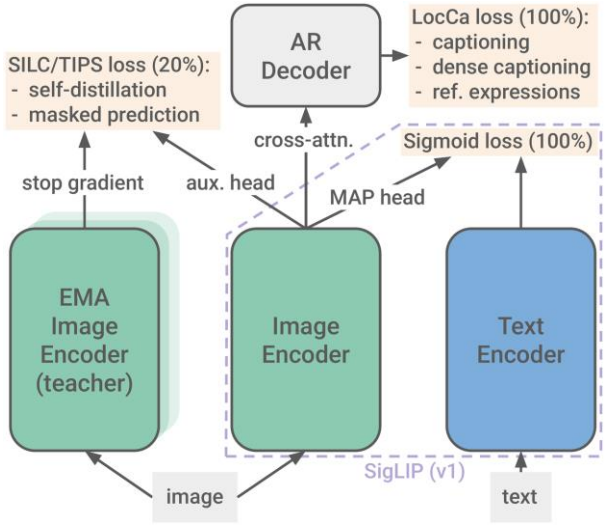
Coarse-grained.
Has to represent (somewhere) notion of objects, relationships, locations, etc.)

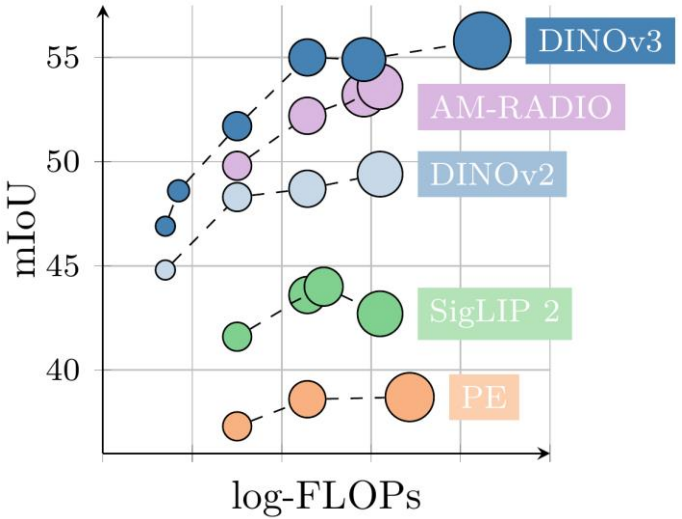
Radford et al., Learning Transferable Visual Models From Natural Language Supervision

CLIP: Learning More Aligned Representations

Georgia Tech

Two directions since 2021: better language supervision, and language-free dense features.

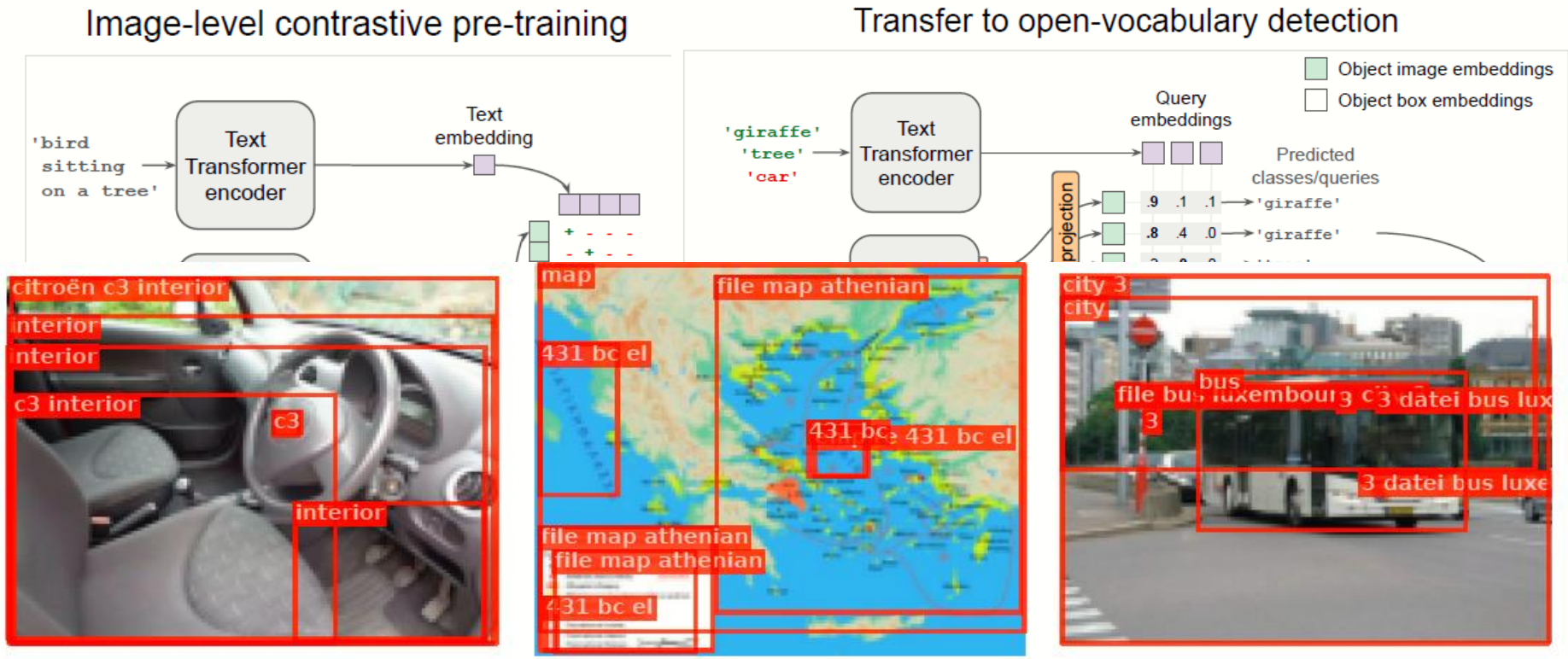




SigLIP 2 keeps the sigmoid loss and adds captioning, self-distillation and masked prediction — plus multilingual data and variable aspect ratios.

DINOv3 trains with no text at all and still wins on dense tasks. Ask why language supervision is not the whole story.

Tschannen et al., SigLIP 2, 2025 · Siméoni et al., DINOv3, 2025 — figures from the papers.

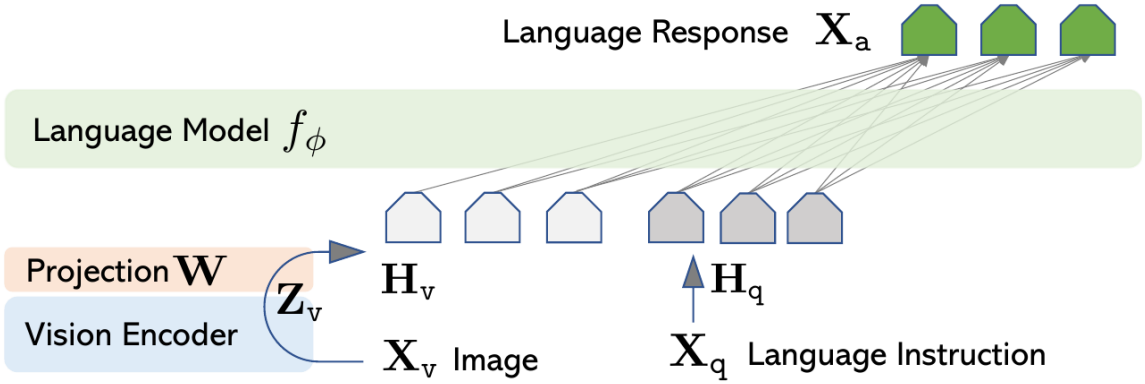


Minderer et al., Simple Open-Vocabulary Object Detection with Vision Transformers
Minderer et al., Scaling Open-Vocabulary Object Detection

Open-Vocabulary Detection → Segmentation

Georgia Tech

Freeze a vision encoder, train one projection, fine-tune on generated conversations.



- The connector can be a single linear layer. Most of the capability already lives in the two pretrained halves.
- Data Contribution: instruction-following conversations synthesized from captions and boxes by a text-only model.
- This is why “which data mixture?” became the central question of 2023–2025.
- Frontier: stop stitching two frozen models together and pretrain multimodally from the start.

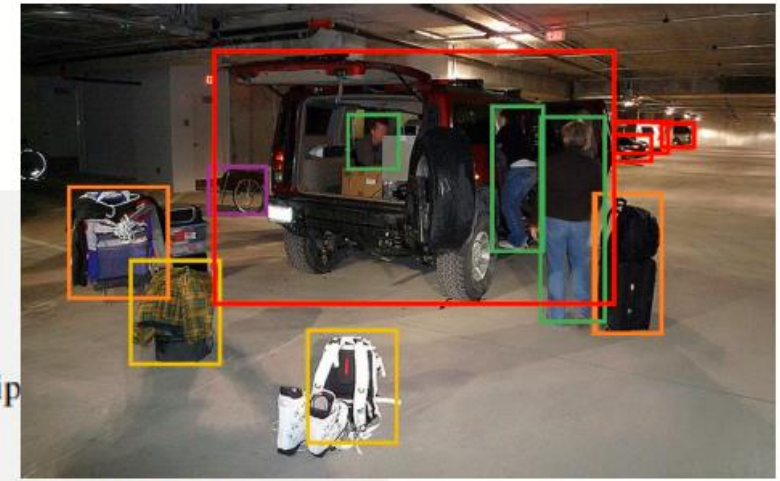
Liu et al., Visual Instruction Tuning (LLaVA), NeurIPS 2023 — figure from the paper.

GPT-assisted Visual Instruction Data Generation

Downside?

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

→ Text-only GPT-4



Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.

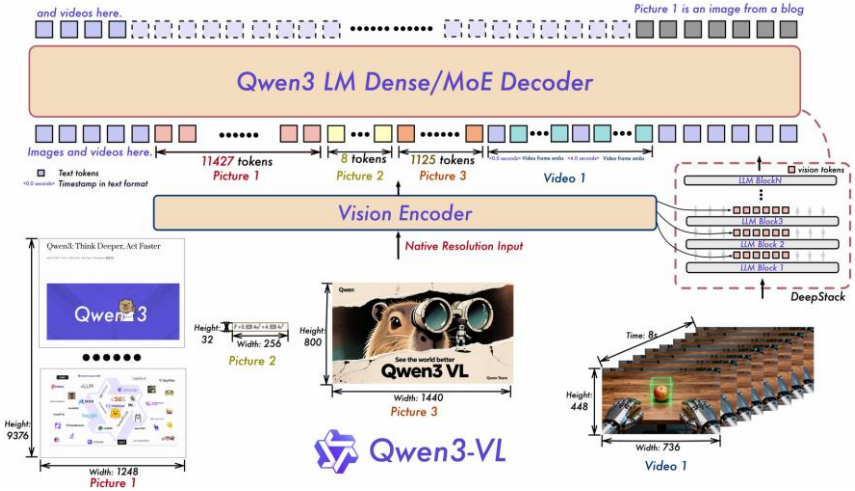
Context type 2: Boxes

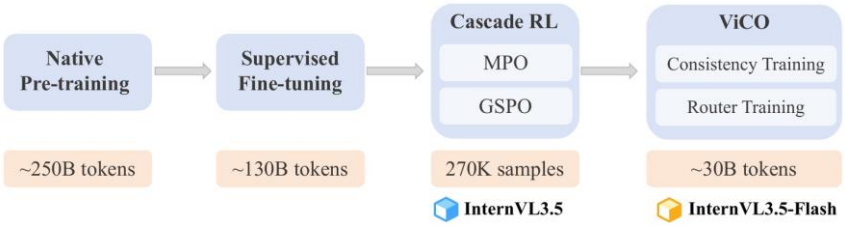
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Data – Using Captions

Georgia Tech

The 2025–26 recipe: multimodal from scratch, long context, and a real post-training stack.

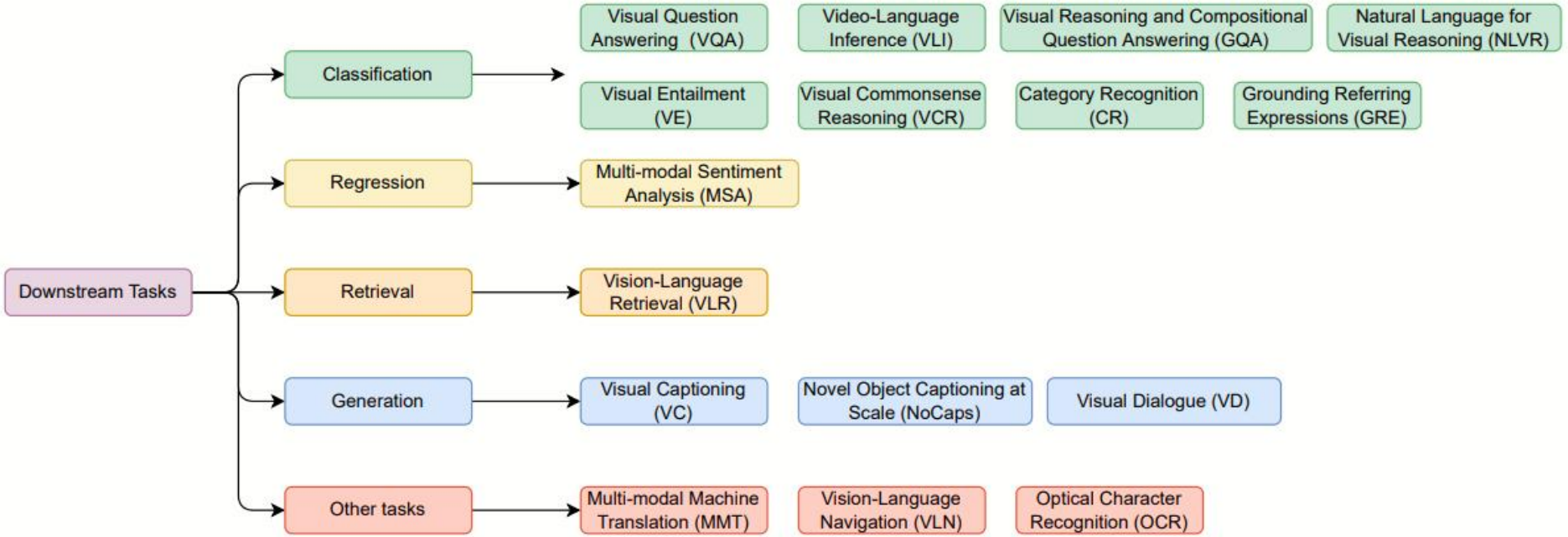




Qwen3-VL: interleaved image / video / text, native-resolution input, timestamped video, and a dense-or-MoE decoder.

InternVL3.5: native pretraining → supervised fine-tuning → cascade RL → efficiency distillation. Post-training is now most of the pipeline.

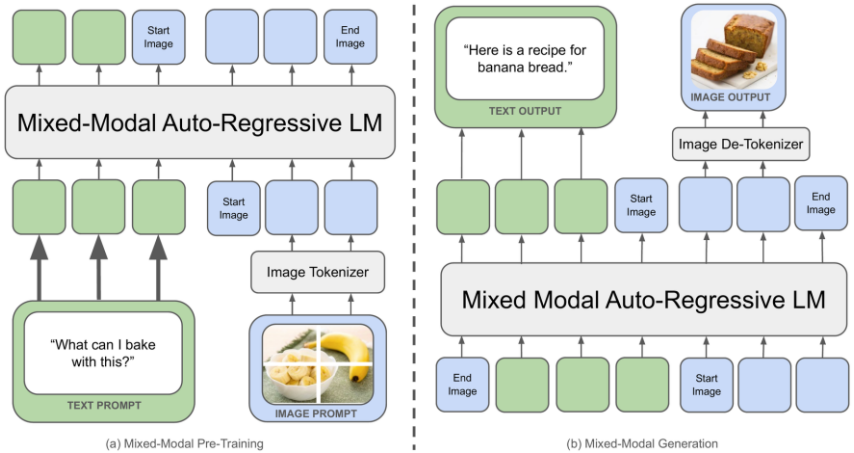
Bai et al., Qwen3-VL, 2025 · Wang et al., InternVL3.5, 2025 — figures from the papers.



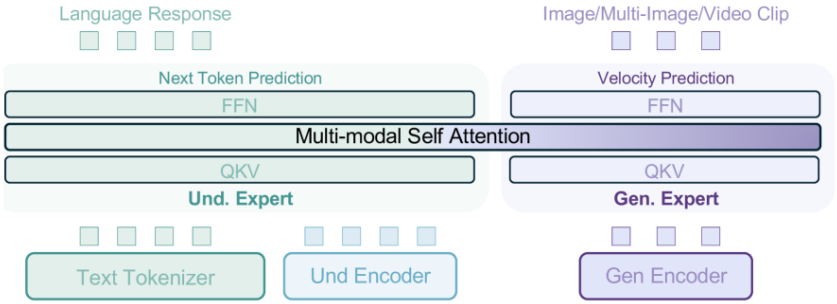
VLP: A Survey on Vision-Language Pre-training, https://arxiv.org/pdf/2202.09061

Tasks

Universal token-in and token-out multi-modal models

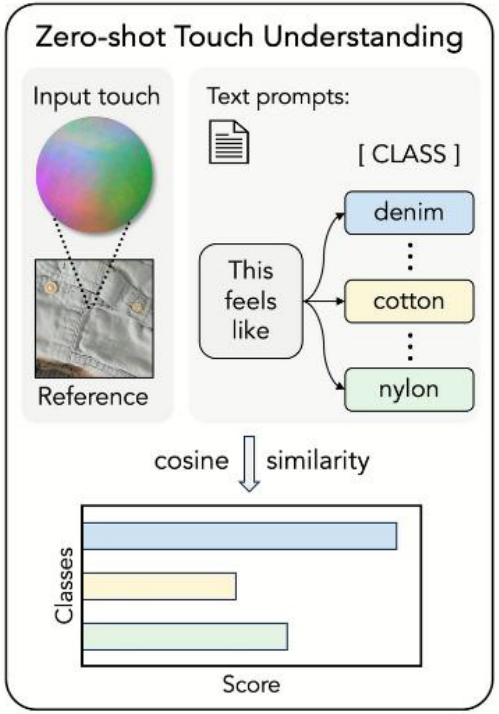


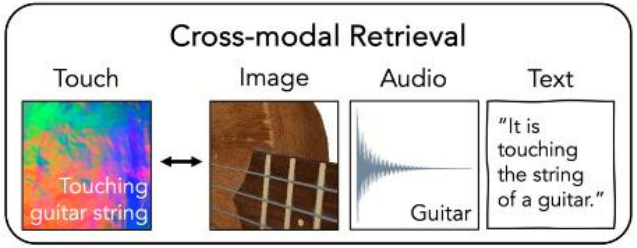
Chameleon: one mixed-modal autoregressive model, early fusion, image and text tokens in a single vocabulary.

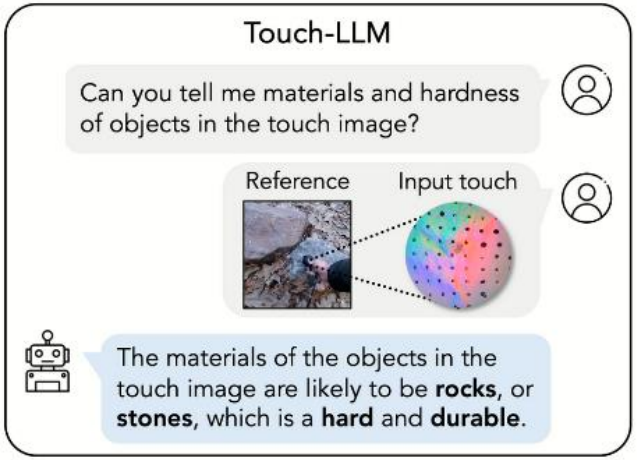


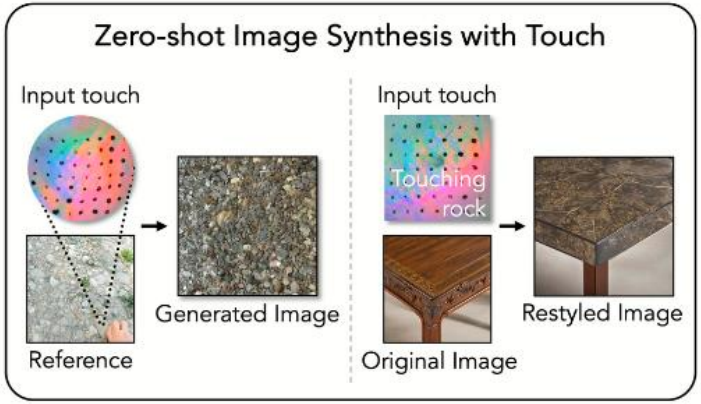
BAGEL: separate understanding and generation experts sharing one attention stack — a different bet on the same goal.

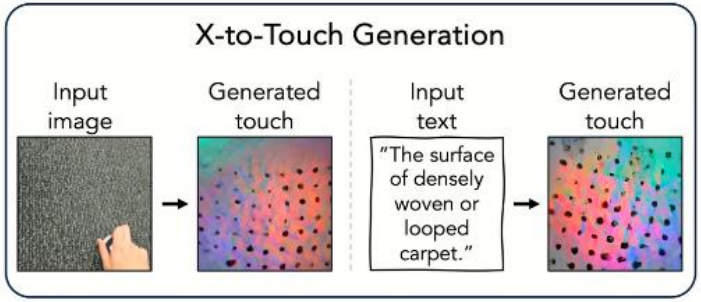
Chameleon Team, 2024 · Deng et al., BAGEL, 2025 — figures from the papers.











Yang et al., Binding Touch to Everything: Learning Unified Multimodal Tactile Representations, CVPR 2024

Other Modalities

Georgia Tech

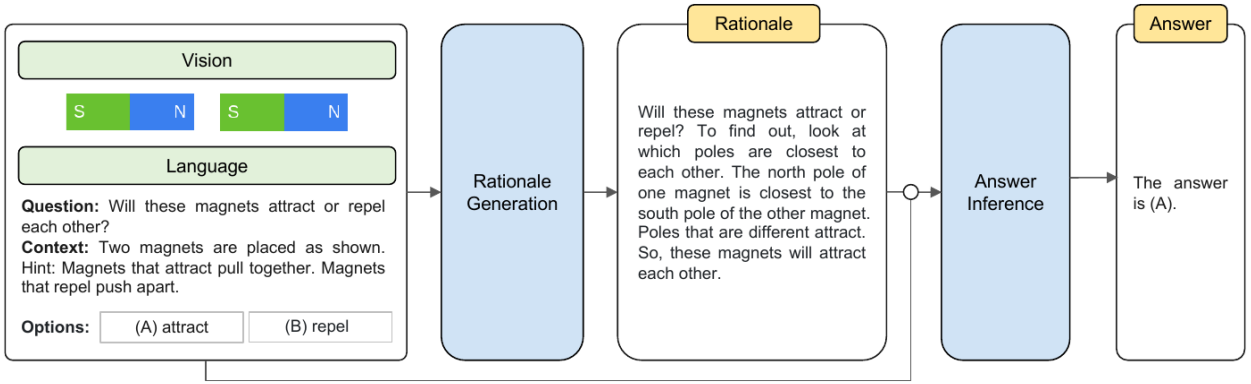
PART 2

Making It Think

Chain-of-thought, RL post-training, and uncertainty modeling — Weeks 6–8.

1	2	3	4	5	6
Week 1	Week 2	Weeks 3–5	Weeks 6–9	Weeks 9–11	Weeks 11–13
Pixels & words	**Aligned representations**	**Vision-language models**	**Reasoning & reliability**	**Agents**	**Action & world models**
Patches and sub-words become one token stream.	Vision and language share an embedding space.	A language model that can see, and generate.	Think before answering; know when you are wrong.	Perceive, plan, and take actions in software.	Act in the physical world; imagine what follows.
Transformer · ViT	**CLIP → SigLIP 2 · DINOv3**	**Flamingo · LLaVA → Qwen3-VL**	**MM-CoT · RLVR · POPE**	**ViperGPT · Mind2Web · OSWorld**	**RT-2 · OpenVLA · DreamZero**

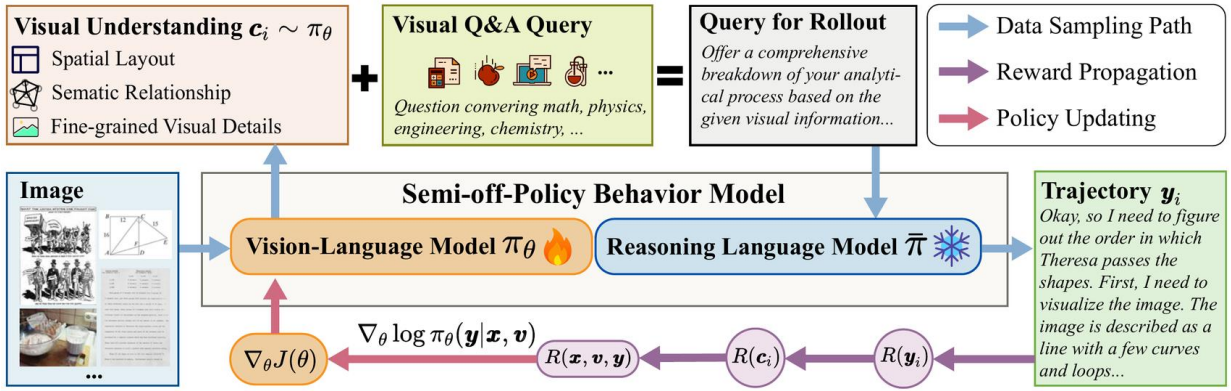
Generate the rationale first, then the answer — and let the rationale see the image.



- Text-only chain-of-thought can make multimodal answers worse: the rationale confidently describes things it cannot see.
- Two stages — rationale generation conditioned on vision, then answer inference conditioned on both.
- What is the mechanism? Is the rationale doing the reasoning, or just producing more tokens to attend over?

Zhang et al., Multimodal Chain-of-Thought Reasoning, TMLR 2024 — figure from the paper.

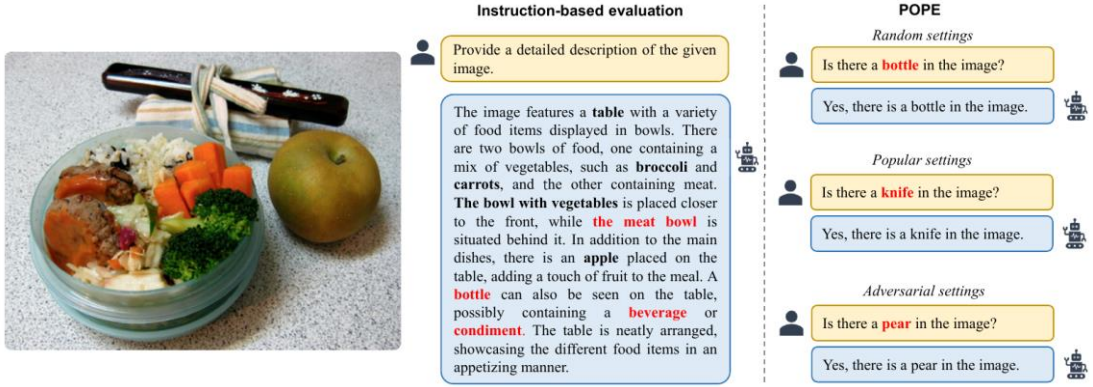
Post-training



- RLHF taught models to follow instructions. Verifiable rewards push further (check correctness)
- Multimodal makes reward design hard — what is a verifiable answer for a perception question?
- Semi-off-policy and behaviour-cloning hybrids exist because on-policy rollouts on a large VLM are expensive.
- Test-time compute as a knob

Ouyang et al., InstructGPT, 2022 → Shen et al., SOPHIA, 2025 — figure from SOPHIA.

Hallucinations



- Free-form caption scoring hides hallucination. POPE uses a balanced yes/no probe
- The failure is systematic: models hallucinate frequent objects, and objects that co-occur with what is there.
- Frontier work attacks the decoding path — attention sinks, grounded decoding — not just the training data.
- Reliability even more important for decision-making

Li et al., POPE, EMNLP 2023 — figure from the paper.

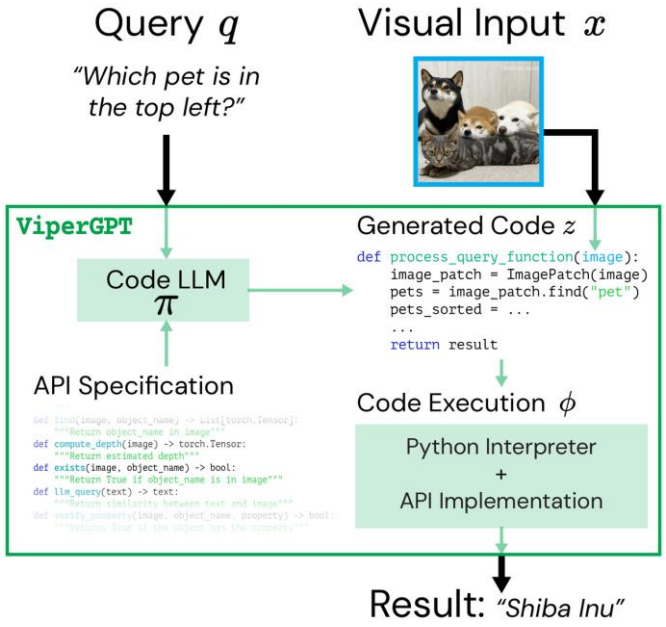
PART 3

Making It Act

Tools, browsers, and the whole desktop — Weeks 9–11.

1	2	3	4	5	6
Week 1	Week 2	Weeks 3–5	Weeks 6–9	Weeks 9–11	Weeks 11–13
Pixels & words	**Aligned representations**	**Vision-language models**	**Reasoning & reliability**	**Agents**	**Action & world models**
Patches and sub-words become one token stream.	Vision and language share an embedding space.	A language model that can see, and generate.	Think before answering; know when you are wrong.	Perceive, plan, and take actions in software.	Act in the physical world; imagine what follows.
Transformer · ViT	CLIP → SigLIP 2 · DINOv3	Flamingo · LLaVA → Qwen3-VL	MM-CoT · RLVR · POPE	ViperGPT · Mind2Web · OSWorld	RT-2 · OpenVLA · DreamZero

Program Synthesis for Visual Reasoning

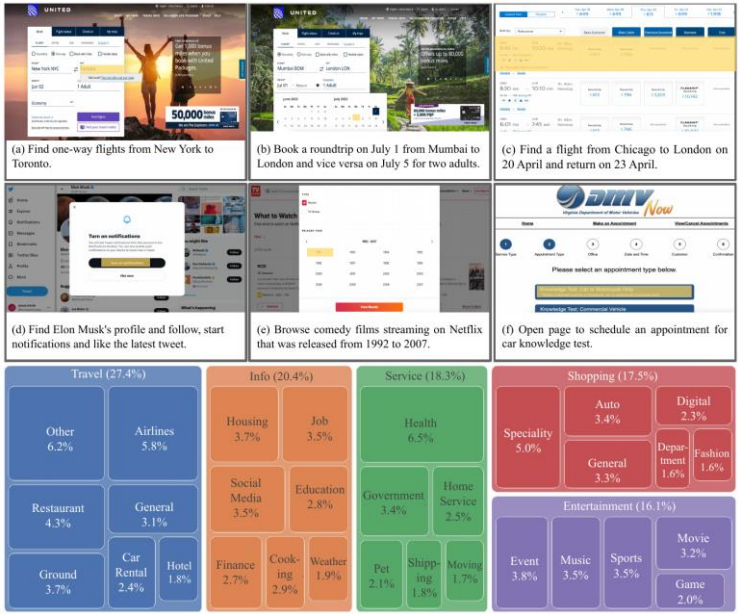


Surís et al., ViperGPT, ICCV 2023 — figure from the paper.

- A code LLM emits Python against a visual API; the interpreter composes detection, depth, cropping and search.
- Compositional and inspectable
- Limitations: brittle to API coverage, and errors compound along the call chain.
- Frontier: multi-agent systems that mix GUI actions and code execution in one policy.

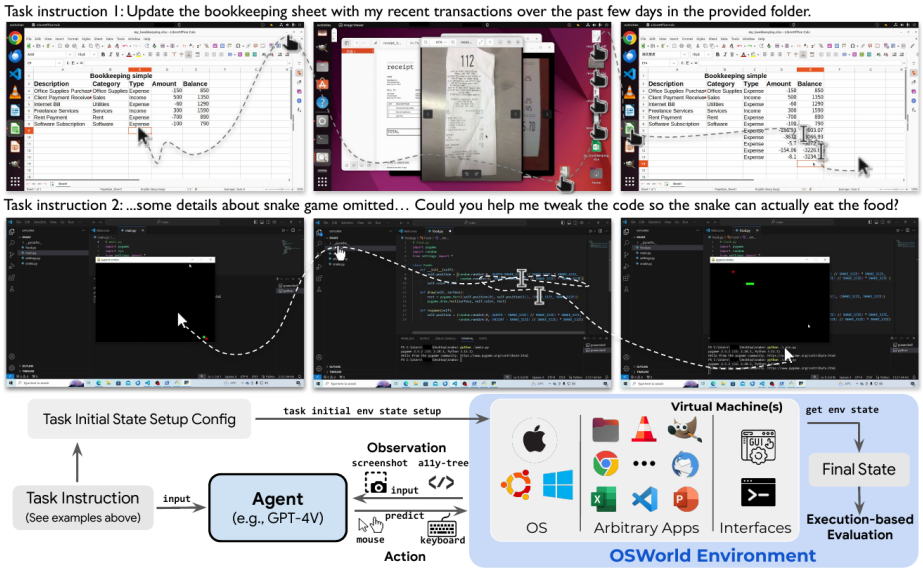
Real websites, real tasks, and an action space of clicks, types and selects.

- Thousands of tasks across 137 real websites in 31 domains — not a sandbox built for the benchmark.
- Challenge: Grounding an intent onto one element out of thousands in a live page.
- Splits are by unseen website and unseen domain
- Frontier: screenshot-only control with no DOM access, plus open web-agent data at scale.



Deng et al., Mind2Web, NeurIPS 2023 — figure from the paper.

Computer Use



Xie et al., OSWorld, NeurIPS 2024 — figure from the paper.

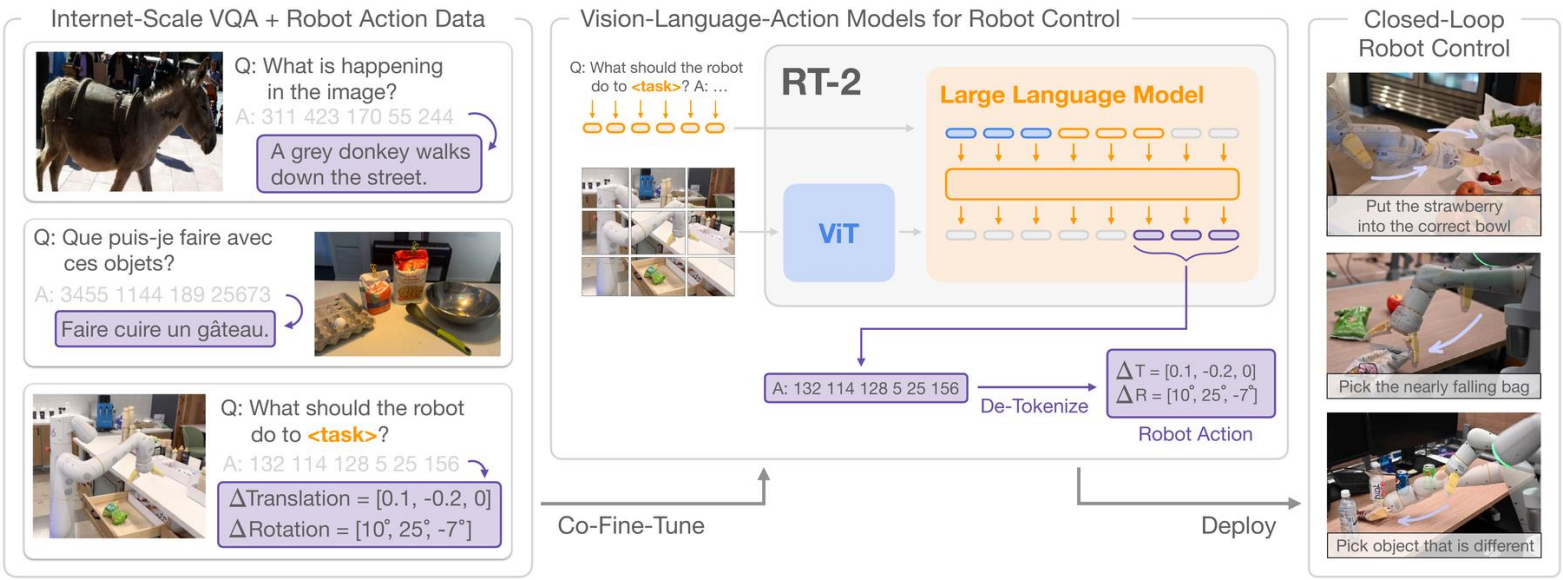
- A real operating system with real apps: file manager, browser, office suite, terminal
- Success is verified by running a script against the final machine state
- Humans clear most tasks; early agents cleared a small fraction.
- Frontier: better screenshot grounding, agent memory, and recovery after a wrong action.

PART 4

Acting in the World

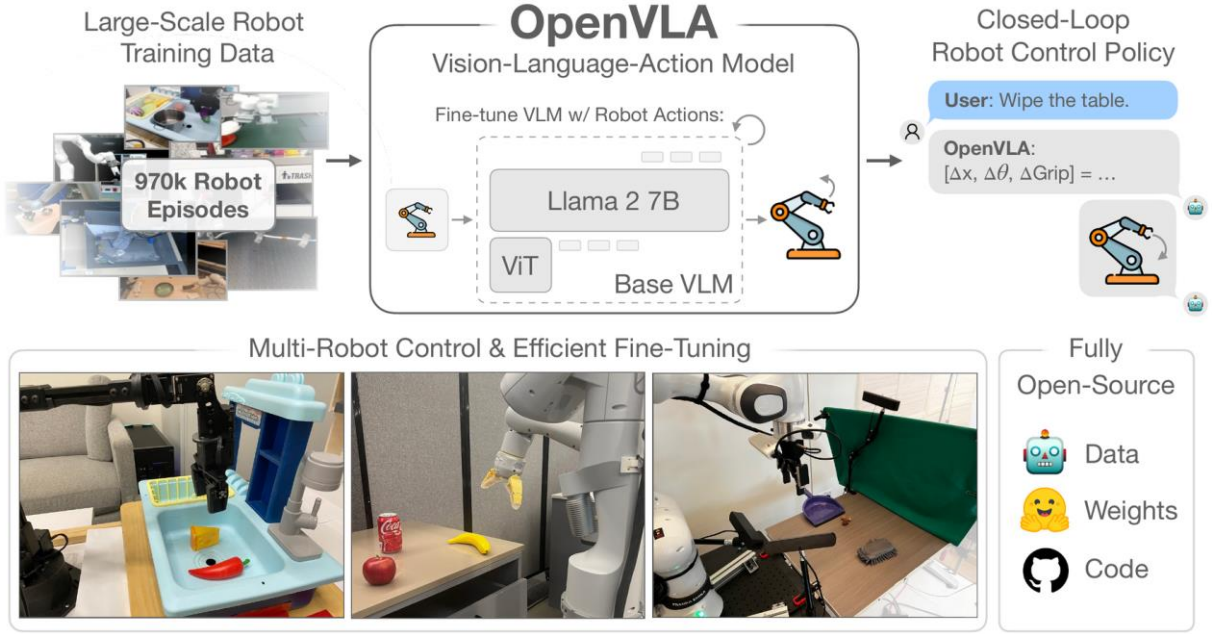
Vision-language-action models, real-time control, and world models — Weeks 11–13.

1	2	3	4	5	6
Week 1	Week 2	Weeks 3–5	Weeks 6–9	Weeks 9–11	Weeks 11–13
Pixels & words	**Aligned representations**	**Vision-language models**	**Reasoning & reliability**	**Agents**	**Action & world models**
Patches and sub-words become one token stream.	Vision and language share an embedding space.	A language model that can see, and generate.	Think before answering; know when you are wrong.	Perceive, plan, and take actions in software.	Act in the physical world; imagine what follows.
Transformer · ViT	CLIP → SigLIP 2 · DINOv3	Flamingo · LLaVA → Qwen3-VL	MM-CoT · RLVR · POPE	ViperGPT · Mind2Web · OSWorld	RT-2 · OpenVLA · DreamZero



Discretize the end-effector delta, put it in the vocabulary, and co-fine-tune on web VQA.

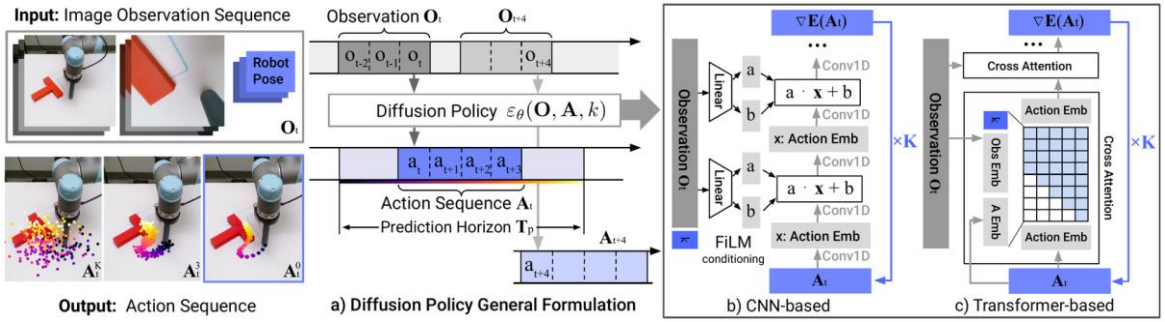
Brohan et al., RT-2, CoRL 2023 — figure from the paper.

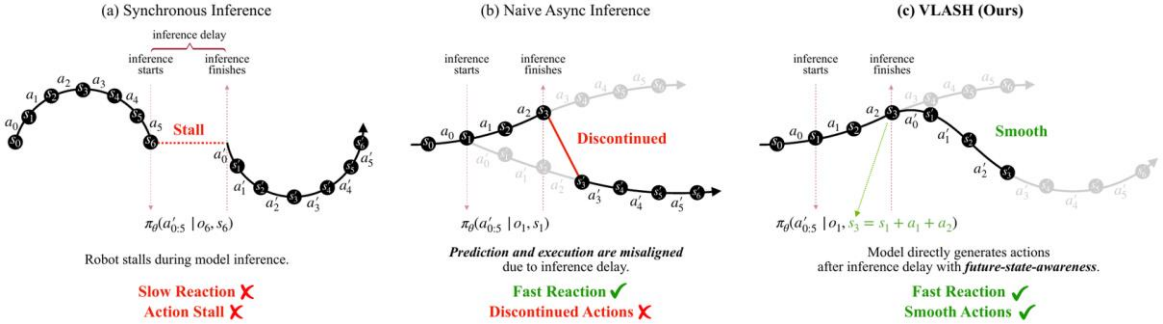


An open 7B VLA trained on ~1M cross-embodiment episodes — the bottleneck is trajectories, not parameters.

Kim et al., OpenVLA, CoRL 2024 — figure from the paper.

Discrete tokens are a poor fit for smooth motion at 30–50 Hz. Two fixes: predict chunks, and hide the latency.





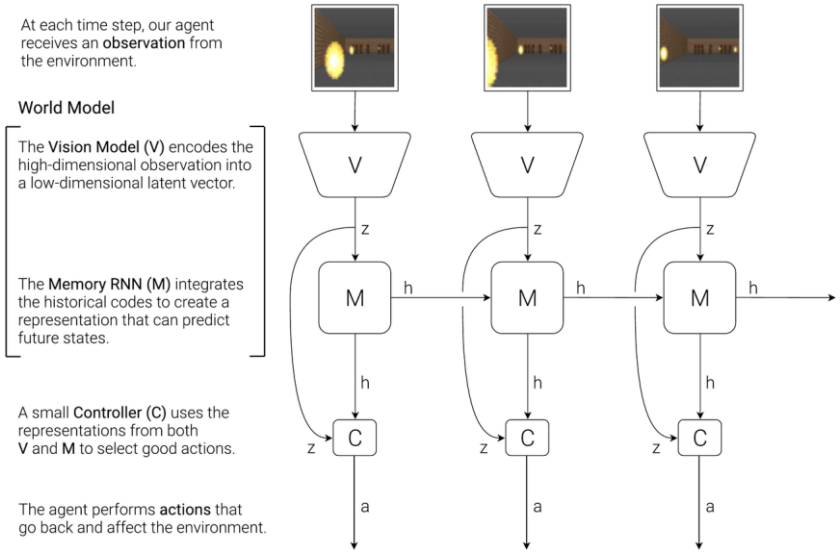
Chi et al., Diffusion Policy, RSS 2023 · Tang et al., VLASH, 2025/26 — figures from the papers.

Diffusion Policy denoises a whole action sequence conditioned on recent observations, instead of emitting one step at a time.

Asynchronous inference acts on predictions made for a future state, so inference delay stops turning into jerky motion.

Compress the observation, learn the dynamics, and train the controller inside the dream.

- V compresses the frame, M predicts the next latent, C is a tiny controller sitting on top of both.
- If your model of the world is good enough, you can plan inside it instead of in the world
- The 2018 version is a VAE plus an RNN. The 2026 version is a video generator.



Ha & Schmidhuber, World Models, NeurIPS 2018 — figure from the paper.

- Joint video-action modelling borrows supervision from videos alone (cheaper than robot teleop data).
- Rolling out an imagined future closes the loop: predict, act, re-observe, correct.
- The claimed payoff is zero-shot transfer to unseen tasks, scenes and embodiments.

- **Predict the future frames and the actions that get you there.**
- **Learned from video, not from teleoperation.**

Ye et al., DreamZero, 2026 — figure from the paper.

- Interface: tokenization, resolution vs. token budget, fusion (early / mid / late)
- Training: pre-training, instruction tuning, RLHF, verifiable rewards, distillation
- Data: captions, synthetic instructions, trajectories, video — and who generated them
- Alignment & grounding: open-vocabulary, dense prediction, spatial and temporal structure
- Reliability: hallucination, calibration, failure recovery, safe execution
- Outputs: text, masks, code and tool calls, GUI actions, continuous control, predicted futures
- Evaluation: static benchmark, live environment, or real hardware — and what that hides

- **For every paper: what are we improving, and how does the improvement come about?**

Important Concepts / Considerations

Georgia Tech

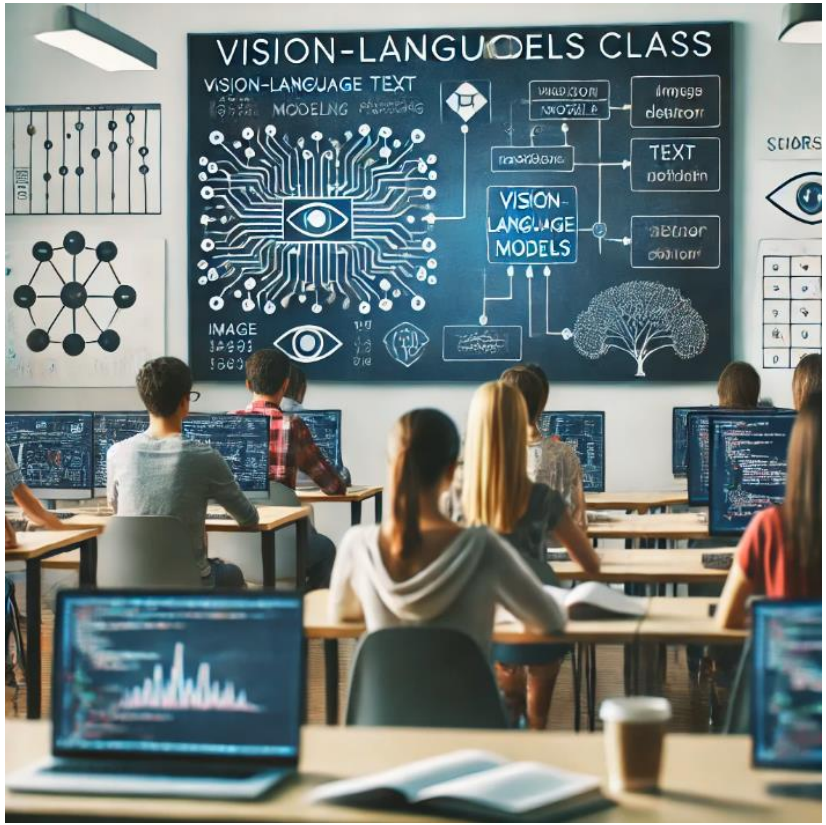
Bring these to your reviews, your presentation, and your project.

- What is the interface? What enters as tokens, what leaves, and who pays for the token budget?
- What is actually being trained — the encoder, the connector, the decoder, or only the data mixture?
- Where did the supervision come from: humans, another model, or a verifier that can check the answer?
- How is success measured — a static benchmark, a live environment, or real hardware? What does that measurement hide?
- What persisted from the seminal paper, what changed, and why did it have to change?

Learning Objectives

- Understand the architectural, training, dataset, and evaluation aspects of multi-modal models.
  - Be able to compare & contrast: what persisted, what changed, and why
  - Have a holistic picture: representations → VLMs → reasoning → agents → embodied action
- Develop skills to read papers, identify gaps/limitations, create novel methods, and rigorously evaluate them.
- Develop skills for presenting research results — and for defending them with evidence.

Georgia Tech



Zsolt Kira

Instructor

Teaching Assistants

Junhyun Kim

Brisa Maneechotesuwan

Office hours: TBA on Ed

Logistics

Georgia Tech

- A participation-driven, in-person research seminar
- In-person attendance is expected and **attendance taken**
  - No routine remote option
- No recordings except for approved accommodations
- Attendance, preparation, and engagement are part of the 15% participation grade
- Don’t come in sick please!!! Tell us early if you will miss a presentation or interview

Delivery Format

Georgia Tech

- It's all deep learning – We expect some deep learning background!
  - CS 7643 (Deep Learning) or similar

Survey:
- Deep Learning class/equivalent?
- Transformers?
- PyTorch/Tensorflow?
- Vision? Language? Audio? Other?
- ECCV? CVPR? EACL? CHI? ICRA? COLING? ICLR?

- What do you want to get out of the class?

- We do not expect deep experience in all modalities!

Recommended Background

- **15% Class Participation**
- **20% Weekly Paper Reviews**
- **15% Paired-Paper Presentation & Discussion**
- **50% Research Project**
  - 5% Proposal
  - 10% Midterm Presentation
  - 10% Rebuttal to the AI-assisted audit
  - 10% Final Presentation
  - 15% Final Report
  - Ungraded but required: code + audit packet, TA interview

Credits: This course was designed with materials and inspiration from great courses such as Learning from Limited Labels and Internet Data Science

Grading Components

Georgia Tech

- **Every research session pairs a seminal paper with a current frontier paper**
  - We study what persisted, what changed, and why
- Broader arc: reasoning, tool use, web/GUI agents, driving, VLA and world-action models
- **New project workflow with verification**
  - Code + audit packet → AI-assisted audit returned
  - Evidence-based rebuttal (10%) → TA project/code interviews
  - AGENT_USAGE.md required in the repository root
- Syllabus changes: You can propose replacement/addition of paper for your section (same topic)

- Each research meeting has two anchors
- **Seminal anchor**
  - The idea, abstraction, or architecture that established the line of work
- **Frontier anchor**
  - A recent paper that changes the scale, objective, architecture, evaluation, or deployment setting
  - The default core review paper for that week (unless website says otherwise)
- **Discussion asks: what persisted, what changed, and why?**

Wk	Date	Topic	Seminal → Frontier (core review)
W1	Aug 25 Tue	Intro, logistics & the VLM → agent arc	—
W1	Aug 27 Thu	**Lecture:** Transformers & multimodal tokenization	Attention Is All You Need (2017) → Survey (2026)
W2	Aug 31 Mon	**Paper presentation sign-up due (11:59 pm ET)**	**—**
W2	Sep 1 Tue	**Lecture:** Vision encoders as foundation models	ViT (2020) → DINOv3 (2025)
W2	Sep 3 Thu	Vision–language alignment & open-vocabulary	CLIP (2021) → SigLIP 2 (2025)
W3	Sep 8 Tue	From frozen backbones to large VLMs	Flamingo (2022) → InternVL3.5 (2025)
W3	Sep 10 Thu	**Instruction-tuned VLMs · team roster due**	**LLaVA (2023) → Qwen3-VL (2025)**
W4	Sep 15 Tue	Open-vocab detection, grounding & segmentation	OWL-ViT (2022) → Qwen3-VL-Seg (2026)
W4	Sep 17 Thu	**Project proposal presentations (slides due Sep 16)**	**—**
W5	Sep 22 Tue	Unified multimodal understanding + generation	Chameleon (2024) → BAGEL (2025)
W5	Sep 24 Thu	Grounded image generation & editing	ControlNet (2023) → Skywork UniPic 2.0 (2025)
W6	Sep 29 Tue	Video VLMs & spatiotemporal reasoning	TimeSformer (2021) → STORM (2026)
W6	Oct 1 Thu	Multimodal chain-of-thought & reasoning	Multimodal CoT (2023) → Phi-4-reasoning-vision (2026)
W7	Oct 6 Tue	Fall Break — no class	—
W7	Oct 8 Thu	Post-training & RL for multimodal reasoning	InstructGPT / RLHF (2022) → SOPHIA (2025)
W8	Oct 13 Tue	Hallucination, grounding & visual reliability	POPE (2023) → SAGE (2026)
W8	Oct 15 Thu	**Project midterm presentations (slides due Oct 14)**	**—**

Fall 2026 Schedule — Foundations to Reasoning

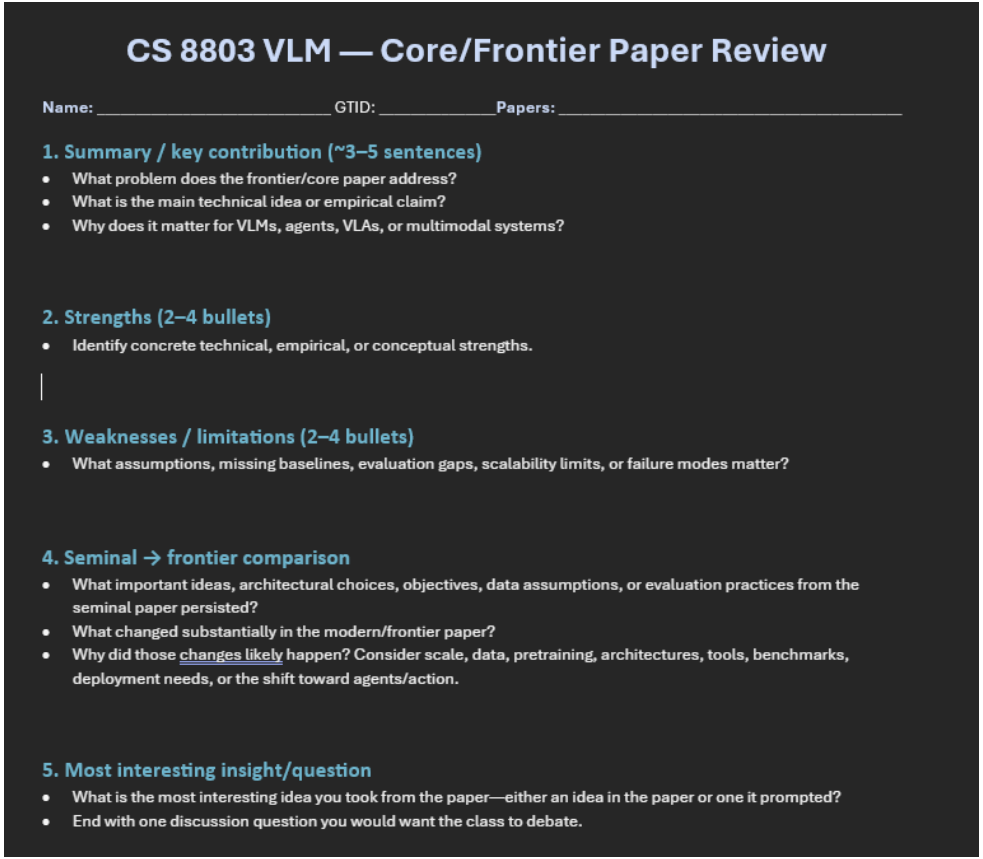
Georgia Tech

Wk	Date	Topic	Seminal → Frontier (core review)
W9	Oct 20 Tue	Spatial / 3D reasoning & geometric structure	SpatialVLM (2024) → CRISP (2026)
W9	Oct 22 Thu	Multimodal tool use & modular agents	ViperGPT (2023) → CoAct-1 (2025)
W10	Oct 27 Tue	Web agents: structured pages → visual policies	Mind2Web (2023) → MolmoWeb (2026)
W10	Oct 29 Thu	Computer-use / GUI agents	OSWorld (2024) → OSWorld 2.0 (2026)
W11	Nov 3 Tue	Agent memory, planning & safe computer use	ReAct (2022) → Action-Grounded Visual Memory (2026)
W11	Nov 5 Thu	Autonomous driving as multimodal decision-making	DriveLM (2023) → VLGA (2026)
W12	Nov 10 Tue	Embodied VLMs → Vision-Language-Action	RT-2 (2023) → MolmoAct2 (2026)
W12	Nov 12 Thu	Generalist VLAs, action reps & data scaling	OpenVLA (2024) → Xiaomi-Robotics-1 (2026)
W13	Nov 17 Tue	**Fast, continuous VLA control · code + audit packet due**	**Diffusion Policy (2023) → VLASH (2025/26)**
W13	Nov 19 Thu	World models + action: world-action models	World Models (2018) → DreamZero (2026)
W13	Nov 20 Fri	**AI-assisted project audit / review returned**	**—**
W14	Nov 24 Tue	**Project clinic · rebuttal due (11:59 pm ET)**	**—**
W14	Nov 26 Thu	Thanksgiving — no class	—
W15	Nov 30–Dec 4	**TA project / code interviews (verification gate)**	**—**
W15	Dec 1 Tue	**Final project presentations I (slides due Nov 30)**	**—**
W15	Dec 3 Thu	**Final project presentations II (slides due Dec 2)**	**—**
W16	Dec 8 Tue	**Final presentations III + synthesis (slides due Dec 7)**	**—**
Final	Canvas	**Final project report — deadline posted on Canvas**	**—**

Fall 2026 Schedule — Agents to Embodied Action

Georgia Tech

- Template (instructions are in Word comments)
- 1-page review of the designated frontier / core paper
- Presenters do not review their own session
- **Due: 11:59 pm ET on the day before class. Late reviews will not be accepted.**
- We will drop your lowest 3 eligible scores.
- Limit reviews to 1 page using our review template
- **Write in your own words and proof read — individual work, no generative AI!**



CS 8803 VLM — Core/Frontier Paper Review

Name: _____ GTID: _____ Papers: _____

1. Summary / key contribution (~3–5 sentences)
   - What problem does the frontier/core paper address?
   - What is the main technical idea or empirical claim?
   - Why does it matter for VLMs, agents, VLAs, or multimodal systems?

2. Strengths (2–4 bullets)
   - Identify concrete technical, empirical, or conceptual strengths.

3. Weaknesses / limitations (2–4 bullets)
   - What assumptions, missing baselines, evaluation gaps, scalability limits, or failure modes matter?

4. Seminal → frontier comparison
   - What important ideas, architectural choices, objectives, data assumptions, or evaluation practices from the seminal paper persisted?
   - What changed substantially in the modern/frontier paper?
   - Why did those changes likely happen? Consider scale, data, pretraining, architectures, tools, benchmarks, deployment needs, or the shift toward agents/action.

5. Most interesting insight/question
   - What is the most interesting idea you took from the paper—either an idea in the paper or one it prompted?
   - End with one discussion question you would want the class to debate.

- Template — once per assigned group, lead ~45 minutes on the paired seminal + frontier papers as one intellectual story:
  - **Jointly - slides covering problem, methods, evidence, limitations, and the seminal → frontier evolution**

- After (or during) the presentation we will have in-depth discussions!

- Add any additional resources for students to explore if background is needed or they would like to learn more

- Submit draft slides:
  - Friday before week of presentation (Tuesday presentation)
  - Monday on week of presentation (Thursday presentation)

- **We will give thorough feedback for iteration**

- Submit final form the night before class presentation **11:59 pm ET**
  - Option of a practice presentation during office hours (optional).

- Full paired-paper schedule is on the course website (and the two schedule slides)
- Feel free to suggest papers!
- Frontier papers subject to updating, but similar topic outline
- **Will put sign-up sheet tomorrow**
  - **Sign up on by Monday, August 31, 11:59 pm ET**

Let me know if you may drop the course, before signing up.

- **50% Research Project**
  - 5% Proposal
  - 10% Midterm Presentation
  - 10% Rebuttal to the AI-assisted audit
  - 10% Final Presentation
  - 15% Final Report
  - Required gates: code + audit packet (Nov 17) and TA project/code interview

- **Goal is to develop a novel approach to a multi-modal problem**
- **Shooting for publications at top venues highly encouraged!**

- Possibilities
  - Design and evaluate a novel approach
  - A novel application, use case
  - Extension of a technique studied in class
  - Be creative!

- Target is a research paper at a good conference

- Work in teams of 3-4 (varies by enrollment)

- Team roster due Thursday, September 10, 11:59 pm ET (submit on Canvas)

- Three in-class presentations, plus an audit/rebuttal cycle and a final report

- Project proposal [5%] — slides due Sep 16, present Sep 17
- Project midterm [10%] — slides due Oct 14, present Oct 15
- Rebuttal to audit [10%] — due Nov 24
- Final presentation [10%] — Dec 1 / 3 / 8, each team presents once
  - Slides due 11:59 pm ET the night before your assigned date
- Final report [15%] — deadline posted on Canvas

- **Nov 17 — Code + audit packet (required)**
  - Repo link + immutable commit/tag; README for setup, training, inference, evaluation
  - Dependencies, data/checkpoint instructions, repository-root AGENT_USAGE.md, attribution
  - Team-contribution statement + 2–3 page summary of method, claims, experiments, results
- **Nov 20 — AI-assisted audit + research critique returned**
  - Advisory evidence to evaluate, not an autonomous grade — it may contain mistakes
- **Nov 24 — Rebuttal due [10%]**
  - Per finding: agree / partly agree / disagree, with code pointers, experiments, citations
- **Nov 30 – Dec 4 — TA project/code interviews (verification gate)**
  - Individual understanding of the system, experiments, audit/rebuttal, and contributions
  - Insufficient demonstrated understanding may downweight your entire 50% project component

- Use any language / platform / package you like
- No support for code / implementation issues will be provided — this is not a programming assignment!
- AI assistants / agents OK for implementation
  - But you are the driver of the research question, creativity, novelty, and approach!
  - Material AI/agent use must be documented in a repository-root AGENT_USAGE.md: what was used, for what purpose, which files/experiments it affected, and how you validated the output

Grade	Percentage Range
A	90% and above
B	80% - 89%
C	70% - 79%
D	60% - 69%
F	Below 60%

Grading Scale

Georgia Tech

1. *Read first, Write later*. Read the ENTIRE set of posts/comments before posting.
2. *Avoid language that may come across as strong or offensive*. Language can be easily misinterpreted.
3. *Follow the language rules of the Internet*. Do not write using all capital letters, use emoticons to temper.
4. *Consider the privacy of others.* Ask permission prior to giving out a classmate's email address or other information.
5. *inappropriate material*. Do not forward virus warnings, chain letters, jokes, etc. to classmates or instructors. The sharing of pornographic material is forbidden.

Be professional, objective, and kind!

NOTE: The instructor reserves the right to remove posts that are not collegial in nature and/or do not meet the Online Student Conduct and Etiquette guidelines listed above.

What is the collaboration policy?

- **Collaboration**
  - Encouraged for presentation and project (team members)
  - Can also discuss anything with classmates, but reviews must all be written by you!
    - Cite all resources, including collaborators, used

- **AI Assistants:**
  - Paper reviews, presentations, Ed posts: NO generative AI to summarize, draft, rewrite, or critique — entirely your own work.
  - Project: external codebases, pretrained models, tools, and AI coding assistants/agents are allowed but must be attributed. Document material use in repository-root AGENT_USAGE.md.
  - Ownership: the research question, experimental decisions, execution, interpretation, and understanding must be the team's own.
- **Zero tolerance on plagiarism**
  - Always credit your sources. Don't cheat
  - The TA code interview verifies authorship and understanding.

How do I get in touch?

- Primary means of communication -- **Ed**
  - No direct emails to Instructor unless private information
  - Instructor/TAs can provide answers to everyone on forum
  - Class participation credit for answering questions!
  - No posting answers. We will monitor.
  - Stay respectful and professional

GT Resources for Mental Health

Georgia Tech Police Department
Emergency: Call 911 | 404-894-2500

Dean of Students Office
404-894-2565 | studentlife.gatech.edu
Afterhours Assistance Line & Dean on Call: 404-894-2204

Center for Assessment, Referral and Education (CARE)
404-894-3498 | care.gatech.edu

Collegiate Recovery Program
404-894-2575 | counseling.gatech.edu

Counseling Center
404-894-2575 | counseling.gatech.edu

Health Initiatives
404-894-9980
healthinitiatives.gatech.edu

LGBTQIA Resource Center
404-385-4780 | lgtbqia.gatech.edu

Stamps Psychiatry Center
404-894-1420

VOICE
404-385-4464 | 404-385-4451
24/7 Info Line: 404-894-9000 | voice.gatech.edu

Women's Resource Center
404-385-0230 | womenscenter.gatech.edu

Veterans Resource Center
404-894-4953 | veterans.gatech.edu

Georgia Crisis and Access Line
1-800-715-4225
The crisis line is staffed with professional social workers and counselors 24 hours per day, every day, to assist those with urgent and emergency needs.

Trevor Project
1-866-488-7386
Trained counselors are available to support anyone in need.

National Suicide Prevention Hotline
1-800-273-8255
A national network of local crisis centers that provides free and confidential emotional support to people in suicidal crisis or emotional distress 24/7.

Georgia State Psychology Clinic
404-413-2500
The clinic offers high quality and affordable psychological services to adults, children, adolescents, families and couples from the greater Atlanta area.

Computing

- Major bottleneck: GPUs
- Options
  - Your own / group / advisor's resources
  - PACE-ICE (w/ some H100s)
  - Google Colab
    - jupyter-notebook + free GPU instance
  - Google Cloud credits
    - Tutorial on setting up gloud: https://github.com/cs231n/gcloud

FAQ

- Can I audit? **No** ☹
- Will I get a seat?
  - We can’t expand class size. Waitlist is only hope.

- Read the class website thoroughly after my initial Ed post
- Sign up to lead a paired-paper discussion by Monday, August 31 (11:59 pm ET)
- Start forming project teams — roster due September 10 (11:59 pm ET)
- Thursday: Attention Is All You Need → Qwen3-VL Technical Report
- Probability of dropping class? Talk to me first.

Coming Up

Georgia Tech

- **Have fun!**