Foundation Models and Vision–Language Models

Computer Vision (CSCI 5520G)

Faisal Z. Qureshi

http://vclab.science.ontariotechu.ca

Week 11 at a glance

- **What "foundation model" means, and whether the term earns its keep**
- **CLIP** — language as supervision, and zero-shot recognition
- **Where CLIP fails**
- **From dual encoders to generative VLMs**
- **SAM** — promptable segmentation and the data engine
- **Adapting a foundation model** — probing, fine-tuning, LoRA
- **Evaluation, bias, cost, and consequence**

Reading: CLIP (Radford et al., ICML 2021), SAM (Kirillov et al., ICCV 2023).

The shape of the argument

Weeks 1–5: build a model for a task, from labelled data for that task.

The shape of the argument

Weeks 1–5: build a model for a task, from labelled data for that task.

Week 8: pretrain without labels, then adapt.

The shape of the argument

Weeks 1–5: build a model for a task, from labelled data for that task.

Week 8: pretrain without labels, then adapt.

This week: pretrain **once**, at enormous scale, on **broad** data, then adapt to many tasks — sometimes with no task-specific training at all.

The shape of the argument

Weeks 1–5: build a model for a task, from labelled data for that task.

Week 8: pretrain without labels, then adapt.

This week: pretrain **once**, at enormous scale, on **broad** data, then adapt to many tasks — sometimes with no task-specific training at all.

The term *foundation model* was coined for this pattern. It is descriptive, not technical: a model is a foundation model because of how it is **used**, not because of what it **is**.

The shape of the argument

Weeks 1–5: build a model for a task, from labelled data for that task.

Week 8: pretrain without labels, then adapt.

This week: pretrain **once**, at enormous scale, on **broad** data, then adapt to many tasks — sometimes with no task-specific training at all.

The term *foundation model* was coined for this pattern. It is descriptive, not technical: a model is a foundation model because of how it is **used**, not because of what it **is**.

Treat the term with mild suspicion. It bundles a real capability shift with a good deal of marketing.

Language as supervision

Classification forces a fixed, closed vocabulary. ImageNet has exactly 1,000 classes; a model trained on it cannot express anything else.

Language as supervision

Classification forces a fixed, closed vocabulary. ImageNet has exactly 1,000 classes; a model trained on it cannot express anything else.

But images on the web arrive **with text** — captions, alt-text, surrounding prose.

Language as supervision

Classification forces a fixed, closed vocabulary. ImageNet has exactly 1,000 classes; a model trained on it cannot express anything else.

But images on the web arrive **with text** — captions, alt-text, surrounding prose.

That text is noisy, incomplete, and unstructured. It is also:

- free, and available at the scale of hundreds of millions
- open-vocabulary — not restricted to a label set decided in advance
- semantically rich — relations, attributes, context, not just nouns

Language as supervision

Classification forces a fixed, closed vocabulary. ImageNet has exactly 1,000 classes; a model trained on it cannot express anything else.

But images on the web arrive **with text** — captions, alt-text, surrounding prose.

That text is noisy, incomplete, and unstructured. It is also:

- free, and available at the scale of hundreds of millions
- open-vocabulary — not restricted to a label set decided in advance
- semantically rich — relations, attributes, context, not just nouns

The bet: noisy supervision at enormous scale beats clean supervision at moderate scale. This is the same bet as ViT's, made again.

CLIP

Contrastive **L**anguage–**I**mage **P**retraining. 400 million image–text pairs.

CLIP

Contrastive **L**anguage–**I**mage **P**retraining. 400 million image–text pairs.

- An image encoder (ResNet or ViT) gives $\mathbf{z}^I$
- A text encoder (transformer) gives $\mathbf{z}^T$
- Both are projected to a shared embedding space and ℓ_2-normalised

CLIP

Contrastive **L**anguage–**I**mage **P**retraining. 400 million image–text pairs.

- An image encoder (ResNet or ViT) gives $\mathbf{z}^I$
- A text encoder (transformer) gives $\mathbf{z}^T$
- Both are projected to a shared embedding space and ℓ_2-normalised

For a batch of N pairs, form the $N \times N$ matrix of cosine similarities. The N correct pairs are on the diagonal.

CLIP

Contrastive **L**anguage–**I**mage **P**retraining. 400 million image–text pairs.

- An image encoder (ResNet or ViT) gives $\mathbf{z}^I$
- A text encoder (transformer) gives $\mathbf{z}^T$
- Both are projected to a shared embedding space and ℓ_2-normalised

For a batch of N pairs, form the $N \times N$ matrix of cosine similarities. The N correct pairs are on the diagonal.

Train with a symmetric InfoNCE loss — cross-entropy across each row *and* each column.

CLIP

Contrastive **L**anguage–**I**mage **P**retraining. 400 million image–text pairs.

- An image encoder (ResNet or ViT) gives $\mathbf{z}^I$
- A text encoder (transformer) gives $\mathbf{z}^T$
- Both are projected to a shared embedding space and ℓ_2-normalised

For a batch of N pairs, form the $N \times N$ matrix of cosine similarities. The N correct pairs are on the diagonal.

Train with a symmetric InfoNCE loss — cross-entropy across each row *and* each column.

Same machinery as Week 8. The difference is that the positive pair is now **image and its caption**, not two crops of one image.

The CLIP objective

[Figure placeholder]

Image and text encoders projecting into a shared embedding space, and the $N \times N$ similarity matrix whose diagonal holds the correct pairs

Zero-shot classification

The consequence is what made the paper famous.

Zero-shot classification

The consequence is what made the paper famous.

To classify into arbitrary classes $c_1 \ldots c_K$ with no training:

1. Write each class as a sentence — "a photo of a c_k"
2. Encode all K sentences with the text encoder
3. Encode the image
4. Return the class whose text embedding is most similar

Zero-shot classification

The consequence is what made the paper famous.

To classify into arbitrary classes $c_1 \dots c_K$ with no training:

1. Write each class as a sentence — "a photo of a c_k"
2. Encode all K sentences with the text encoder
3. Encode the image
4. Return the class whose text embedding is most similar

The classifier is **constructed from language at inference time**. Change the class list and you change the task, with no gradient steps.

Zero-shot classification

The consequence is what made the paper famous.

To classify into arbitrary classes $c_1 \dots c_K$ with no training:

1. Write each class as a sentence — "a photo of a c_k"
2. Encode all K sentences with the text encoder
3. Encode the image
4. Return the class whose text embedding is most similar

The classifier is **constructed from language at inference time**. Change the class list and you change the task, with no gradient steps.

Zero-shot CLIP matches a fully supervised ResNet-50 on ImageNet — having seen none of its training set.

Prompts are not a detail

"a photo of a {}" outperforms the bare class name by several points.

Prompts are not a detail

"a photo of a {}" outperforms the bare class name by several points.

Why: the training captions are sentences, so a bare noun is off-distribution.

Prompts are not a detail

"a photo of a {}" outperforms the bare class name by several points.

Why: the training captions are sentences, so a bare noun is off-distribution.

It gets more specific — "a satellite photo of a {}", "a blurry photo of a {}", "a sketch of a {}" each help on the matching domain. Ensembling many prompts and averaging the embeddings helps again.

Prompts are not a detail

"a photo of a {}" outperforms the bare class name by several points.

Why: the training captions are sentences, so a bare noun is off-distribution.

It gets more specific — "a satellite photo of a {}", "a blurry photo of a {}", "a sketch of a {}" each help on the matching domain. Ensembling many prompts and averaging the embeddings helps again.

This should make you uncomfortable. Reported zero-shot accuracy depends on prompts chosen by the authors, tuned on the evaluation task.

Prompts are not a detail

"a photo of a {}" outperforms the bare class name by several points.

Why: the training captions are sentences, so a bare noun is off-distribution.

It gets more specific — "a satellite photo of a {}", "a blurry photo of a {}", "a sketch of a {}" each help on the matching domain. Ensembling many prompts and averaging the embeddings helps again.

This should make you uncomfortable. Reported zero-shot accuracy depends on prompts chosen by the authors, tuned on the evaluation task.

Is that still zero-shot? There is a real argument that prompt engineering is a form of supervision that the accounting does not capture.

Where CLIP fails

The failures are as instructive as the successes.

Where CLIP fails

The failures are as instructive as the successes.

- **Counting.** "three dogs" is close to "two dogs". Contrastive captions rarely hinge on number.
- **Spatial relations.** "the cup on the book" versus "the book on the cup" — near-identical embeddings. CLIP largely behaves like a bag of concepts.
- **Fine-grained and specialist domains.** Aircraft variants, tumour grading — poor, because the web text is not there.
- **Typographic attacks.** Tape a paper reading "iPod" onto an apple and CLIP says iPod. It reads text in the image, because captions correlate with text.

Where CLIP fails

The failures are as instructive as the successes.

- **Counting.** "three dogs" is close to "two dogs". Contrastive captions rarely hinge on number.
- **Spatial relations.** "the cup on the book" versus "the book on the cup" — near-identical embeddings. CLIP largely behaves like a bag of concepts.
- **Fine-grained and specialist domains.** Aircraft variants, tumour grading — poor, because the web text is not there.
- **Typographic attacks.** Tape a paper reading "iPod" onto an apple and CLIP says iPod. It reads text in the image, because captions correlate with text.

Every one of these traces back to what the training signal could reward. The failures are not random — they are the shape of the data.

Data, not just scale

The follow-on literature is largely about **data quality**.

Data, not just scale

The follow-on literature is largely about **data quality**.

- **LAION** reproduced CLIP-scale data openly — and thereby made its contents auditable. Audits found sexual, violent, and hateful material, and in one case child sexual abuse material, prompting a takedown.
- **DataComp** fixed the training recipe and varied only the data, showing filtering strategy matters more than raw count.

Data, not just scale

The follow-on literature is largely about **data quality**.

- **LAION** reproduced CLIP-scale data openly — and thereby made its contents auditable. Audits found sexual, violent, and hateful material, and in one case child sexual abuse material, prompting a takedown.
- **DataComp** fixed the training recipe and varied only the data, showing filtering strategy matters more than raw count.

The lesson: "trained on 400M pairs" tells you very little. *Which* pairs is the substantive question, and for proprietary models it is unanswerable.

Data, not just scale

The follow-on literature is largely about **data quality**.

- **LAION** reproduced CLIP-scale data openly — and thereby made its contents auditable. Audits found sexual, violent, and hateful material, and in one case child sexual abuse material, prompting a takedown.
- **DataComp** fixed the training recipe and varied only the data, showing filtering strategy matters more than raw count.

The lesson: "trained on 400M pairs" tells you very little. *Which* pairs is the substantive question, and for proprietary models it is unanswerable.

For a paper you present: if the dataset is undisclosed, say so explicitly as a limitation.

Beyond dual encoders

CLIP scores an (image, text) pair. It cannot *produce* text.

Beyond dual encoders

CLIP scores an (image, text) pair. It cannot *produce* text.

For captioning, VQA, and dialogue you need a generative decoder. The dominant pattern:

$$\text{frozen vision encoder} \rightarrow \text{projector} \rightarrow \text{LLM}$$

Beyond dual encoders

CLIP scores an (image, text) pair. It cannot *produce* text.

For captioning, VQA, and dialogue you need a generative decoder. The dominant pattern:

$$\text{frozen vision encoder} \;\rightarrow\; \text{projector} \;\rightarrow\; \text{LLM}$$

- **Frozen encoder** — often CLIP's own ViT
- **A small trained projector** mapping visual features into the LLM's token space
- **A large language model**, frozen or lightly tuned

Beyond dual encoders

CLIP scores an (image, text) pair. It cannot *produce* text.

For captioning, VQA, and dialogue you need a generative decoder. The dominant pattern:

$$\text{frozen vision encoder} \;\rightarrow\; \text{projector} \;\rightarrow\; \text{LLM}$$

- **Frozen encoder** — often CLIP's own ViT
- **A small trained projector** mapping visual features into the LLM's token space
- **A large language model**, frozen or lightly tuned

Only the projector needs training. This is why capable VLMs appeared so quickly after capable LLMs — most of the capability was already paid for.

Beyond dual encoders

CLIP scores an (image, text) pair. It cannot *produce* text.

For captioning, VQA, and dialogue you need a generative decoder. The dominant pattern:

$$\text{frozen vision encoder} \rightarrow \text{projector} \rightarrow \text{LLM}$$

- **Frozen encoder** — often CLIP's own ViT
- **A small trained projector** mapping visual features into the LLM's token space
- **A large language model**, frozen or lightly tuned

Only the projector needs training. This is why capable VLMs appeared so quickly after capable LLMs — most of the capability was already paid for.

Flamingo, BLIP-2 and LLaVA are variations on where the adapter sits and what is frozen.

Segment Anything

A different kind of foundation model: not open-vocabulary *recognition*, but open-ended *segmentation*.

Segment Anything

A different kind of foundation model: not open-vocabulary *recognition*, but open-ended *segmentation*.

The task is redefined. SAM does not segment classes. It takes a **prompt** — a point, a box, a rough mask — and returns the corresponding region. Ambiguity is handled by returning several valid masks.

Segment Anything

A different kind of foundation model: not open-vocabulary *recognition*, but open-ended *segmentation*.

The task is redefined. SAM does not segment classes. It takes a **prompt** — a point, a box, a rough mask — and returns the corresponding region. Ambiguity is handled by returning several valid masks.

The data engine is the real contribution:

1. Annotators label with model assistance
2. The model retrains on that data
3. It pre-segments more, so annotation gets faster
4. Repeat, ending largely automatic

Segment Anything

A different kind of foundation model: not open-vocabulary *recognition*, but open-ended *segmentation*.

The task is redefined. SAM does not segment classes. It takes a **prompt** — a point, a box, a rough mask — and returns the corresponding region. Ambiguity is handled by returning several valid masks.

The data engine is the real contribution:

1. Annotators label with model assistance
2. The model retrains on that data
3. It pre-segments more, so annotation gets faster
4. Repeat, ending largely automatic

Result: 1.1 billion masks over 11 million images. The model made its own training set feasible.

Segment Anything

A different kind of foundation model: not open-vocabulary *recognition*, but open-ended *segmentation*.

The task is redefined. SAM does not segment classes. It takes a **prompt** — a point, a box, a rough mask — and returns the corresponding region. Ambiguity is handled by returning several valid masks.

The data engine is the real contribution:

1. Annotators label with model assistance
2. The model retrains on that data
3. It pre-segments more, so annotation gets faster
4. Repeat, ending largely automatic

Result: 1.1 billion masks over 11 million images. The model made its own training set feasible.

Note what SAM does not do: it produces no labels. It finds boundaries, and you must supply the semantics.

Adapting a foundation model

In rough order of cost:

Adapting a foundation model

In rough order of cost:

- **Zero-shot.** Prompt it. No training, no data.
- **Linear probe.** Freeze the backbone, train a linear head. Minutes; needs little data.
- **Prompt / prefix tuning.** Learn the prompt embeddings, not the weights.
- **LoRA.** Freeze W; learn a low-rank update $\Delta W = BA$ with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times d}$, $r \ll d$. Typically under 1% of the parameters, and mergeable at inference.
- **Full fine-tuning.** Best when you have data; risks catastrophic forgetting and needs real compute.

Adapting a foundation model

In rough order of cost:

- **Zero-shot.** Prompt it. No training, no data.
- **Linear probe.** Freeze the backbone, train a linear head. Minutes; needs little data.
- **Prompt / prefix tuning.** Learn the prompt embeddings, not the weights.
- **LoRA.** Freeze W; learn a low-rank update $\Delta W = BA$ with $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times d}$, $r \ll d$. Typically under 1% of the parameters, and mergeable at inference.
- **Full fine-tuning.** Best when you have data; risks catastrophic forgetting and needs real compute.

Default advice: start with a linear probe. It is a strong baseline and it tells you how much of your problem the representation already solves.

Evaluation is in trouble

Evaluation is in trouble

- **Contamination.** When pretraining data is the web, and benchmarks are on the web, "held-out" is not a claim anyone can verify.
- **Benchmarks measure what is easy to measure.** Multiple-choice VQA rewards plausible guessing; many questions are answerable from the question alone.
- **Aggregate numbers hide disparity.** A model can gain overall accuracy while losing accuracy for a subgroup.
- **Undisclosed data makes comparison impossible.** Two models with different data and different recipes differing by 0.4 points tells you nothing.

Evaluation is in trouble

- **Contamination.** When pretraining data is the web, and benchmarks are on the web, "held-out" is not a claim anyone can verify.
- **Benchmarks measure what is easy to measure.** Multiple-choice VQA rewards plausible guessing; many questions are answerable from the question alone.
- **Aggregate numbers hide disparity.** A model can gain overall accuracy while losing accuracy for a subgroup.
- **Undisclosed data makes comparison impossible.** Two models with different data and different recipes differing by 0.4 points tells you nothing.

When you present a paper: ask what the evaluation would look like if the claim were false. If you cannot answer, the evaluation is weak.

Consequences

These systems are deployed on people. That is not a footnote.

Consequences

These systems are deployed on people. That is not a footnote.

- **Consent and provenance.** Web-scraped images include people who never agreed to be training data. Artists' work is included and reproduced in style.
- **Bias.** Documented racial and gender disparities in retrieval and classification — including CLIP associating some demographic groups with non-human categories in the original paper's own audit.

Consequences, continued

- **Surveillance.** Open-vocabulary recognition means a watchlist is now a text file, not a retraining job. That lowers the barrier enormously.

Consequences, continued

- **Surveillance.** Open-vocabulary recognition means a watchlist is now a text file, not a retraining job. That lowers the barrier enormously.

- **Synthetic imagery.** Realistic generation of real people, including non-consensual sexual imagery, is a present harm, not a hypothetical.

Consequences, continued

- **Surveillance.** Open-vocabulary recognition means a watchlist is now a text file, not a retraining job. That lowers the barrier enormously.
- **Synthetic imagery.** Realistic generation of real people, including non-consensual sexual imagery, is a present harm, not a hypothetical.
- **Concentration.** Training runs of this scale are available to a handful of organisations, and they set the defaults everyone inherits.

Consequences, continued

- **Surveillance.** Open-vocabulary recognition means a watchlist is now a text file, not a retraining job. That lowers the barrier enormously.
- **Synthetic imagery.** Realistic generation of real people, including non-consensual sexual imagery, is a present harm, not a hypothetical.
- **Concentration.** Training runs of this scale are available to a handful of organisations, and they set the defaults everyone inherits.

If your project builds on a foundation model, you inherit all of this. Say so in your report.

Summary

- Foundation model describes a usage pattern: pretrain broadly once, adapt many times
- CLIP uses captions as open-vocabulary supervision; the same InfoNCE machinery as Week 8
- Zero-shot classification builds the classifier from text at inference time
- Prompt sensitivity complicates the zero-shot claim
- Failures — counting, spatial relations, typographic attacks — follow from what the training signal rewards
- Data composition matters more than data volume
- Generative VLMs bolt a frozen encoder to an LLM through a small projector
- SAM redefines segmentation as promptable, and bootstraps its own data
- Adaptation ranges from prompting to LoRA to full fine-tuning
- Evaluation is contaminated and disclosure is poor; the harms are concrete

For discussion next week

For discussion next week

1. CLIP's zero-shot number depends on prompts the authors tuned. Propose an evaluation protocol that would make the claim honest. Would it still be impressive?

For discussion next week

1. CLIP's zero-shot number depends on prompts the authors tuned. Propose an evaluation protocol that would make the claim honest. Would it still be impressive?

2. SAM segments without naming. Is a model that finds boundaries but no semantics a foundation model, or a very good edge detector?

For discussion next week

1. CLIP's zero-shot number depends on prompts the authors tuned. Propose an evaluation protocol that would make the claim honest. Would it still be impressive?

2. SAM segments without naming. Is a model that finds boundaries but no semantics a foundation model, or a very good edge detector?

3. LAION's value was that it could be audited — and auditing found serious harms. Does openness that reveals harm make things better or worse?

For discussion next week

1. CLIP's zero-shot number depends on prompts the authors tuned. Propose an evaluation protocol that would make the claim honest. Would it still be impressive?

2. SAM segments without naming. Is a model that finds boundaries but no semantics a foundation model, or a very good edge detector?

3. LAION's value was that it could be audited — and auditing found serious harms. Does openness that reveals harm make things better or worse?

4. Your project will use a pretrained backbone with undisclosed training data. What can you legitimately claim in your report, and what can you not?

Copyright and License