Tomer Biton

HF model of the weekend

openai/clip-vit-base-patch32

openai/clip-vit-base-patch32

Tomer Biton

WHAT CLIP ACTUALLY IS

Contrastive Language–Image Pretraining

If CNNs classify images
and LLMs understand text
CLIP connects both.

CLIP learns:

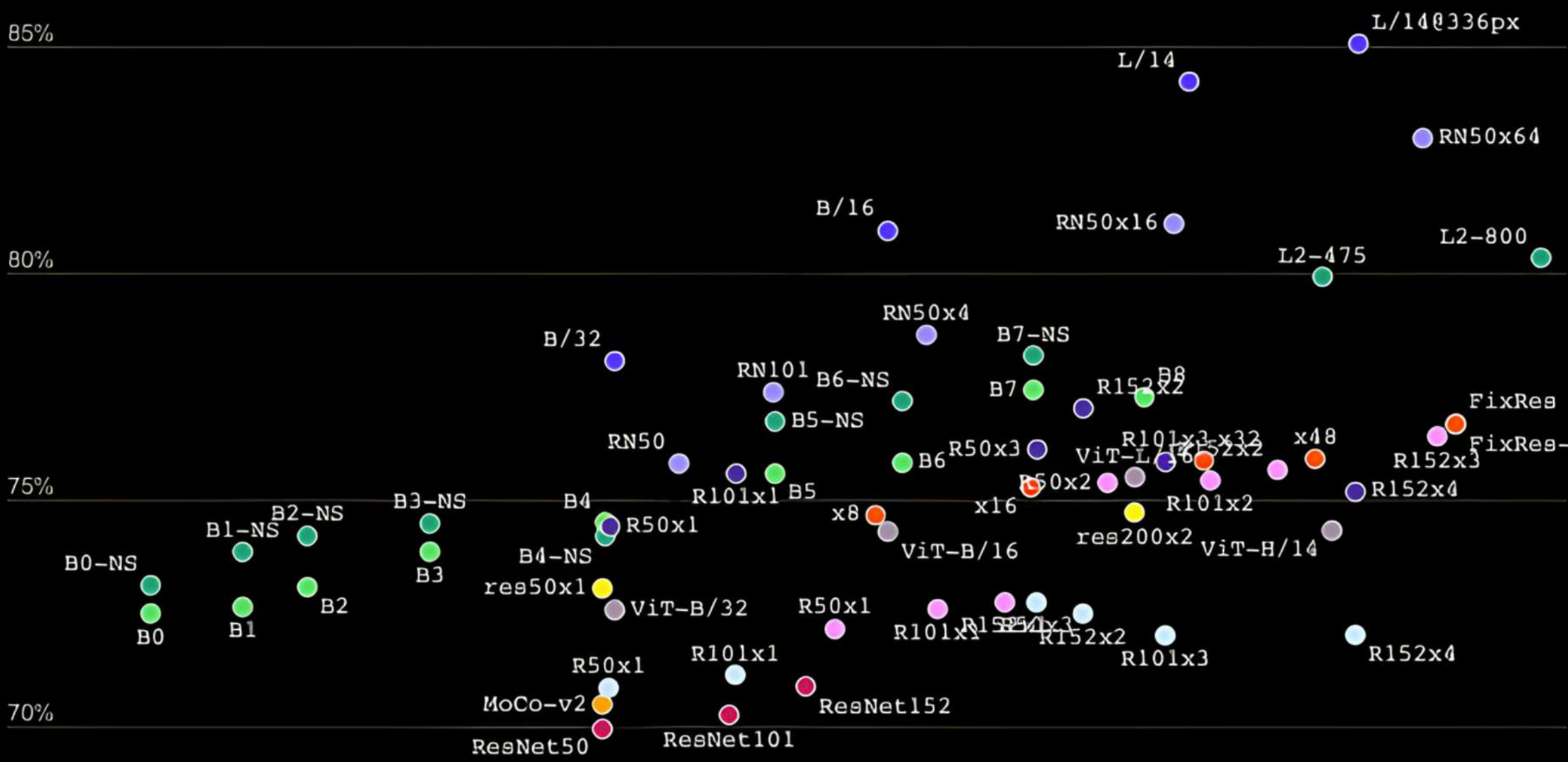
- What text describes an image
- And what image matches a sentence

Key idea:
Images and text are mapped into the same vector space

visual in next slide →

Tomer Biton

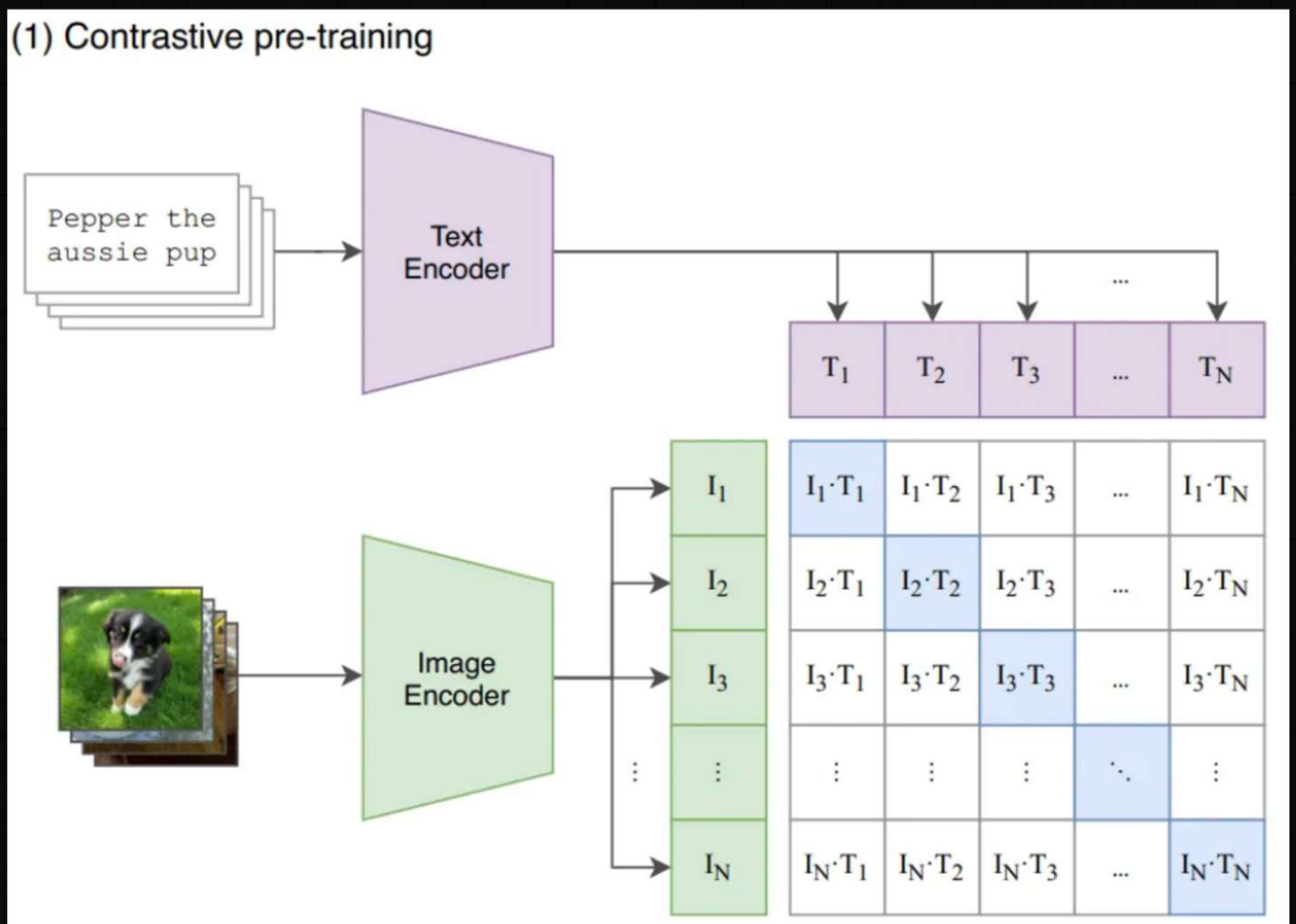
Model comparison chart



CLIP gets more useful representations per unit of compute

Tomer Biton

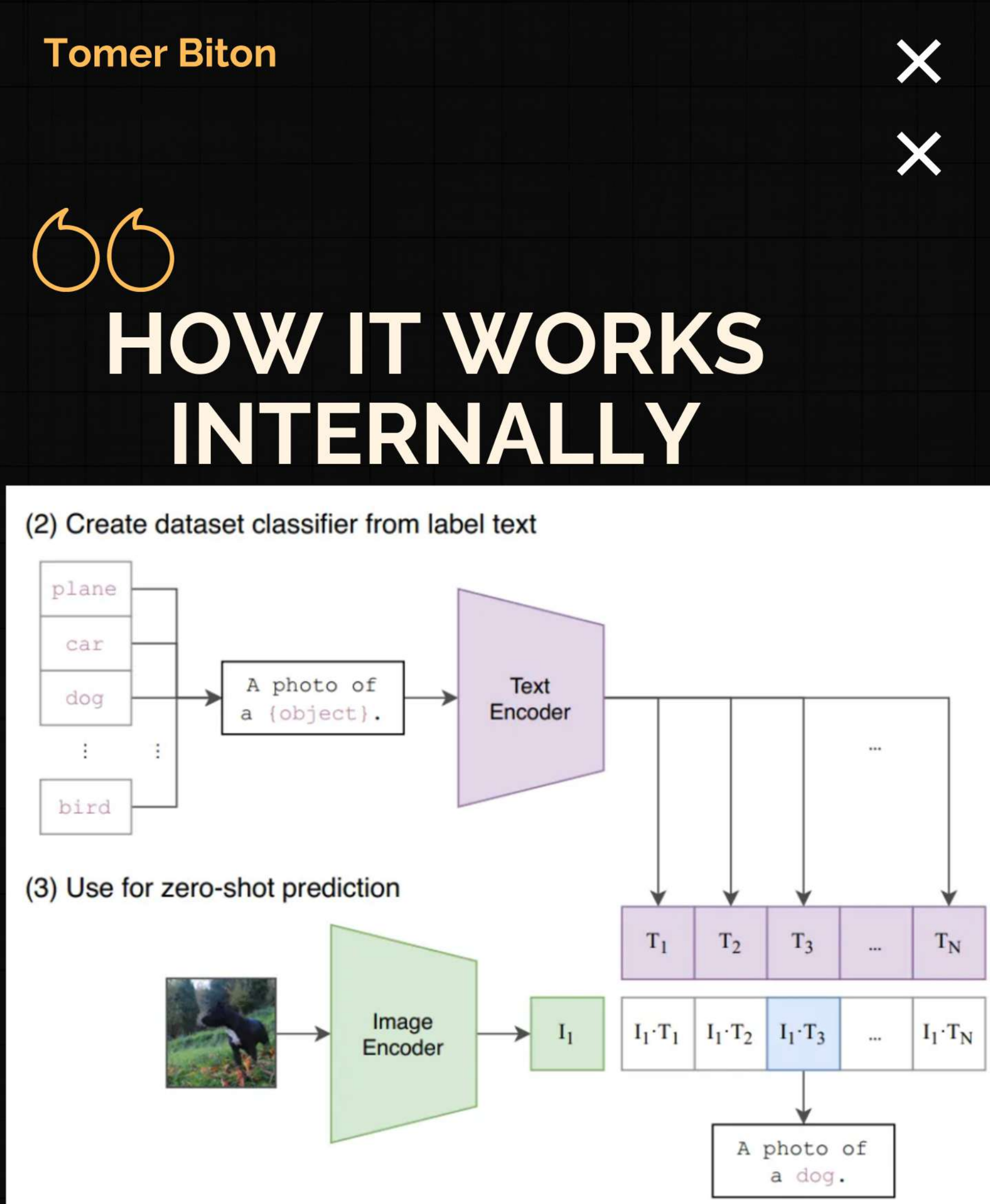
HOW IT WORKS INTERNALLY



CLIP is trained on image + text pairs from the internet.
Then CLIP compares every image with every text. →
Push matching image–text pairs closer
Push non-matching pairs apart

Tomer Biton

HOW IT WORKS INTERNALLY



You take a class name e.g “dog”
it go through a text encoder.
the an image goes through the image encoder
and CLIP compares it to all the text vectors.
the highest. match wins so the model answer
“Which description best matches this image?”

Tomer Biton

WHY THIS IS A BIG DEAL

Why CLIP Changed Vision Models

Before CLIP:

- Fixed labels
- Dataset-specific classifiers
- Retraining for every task

With CLIP:

- Zero-shot classification
- Natural language as the interface
- One model → many tasks

Tomer Biton

What I Learned This Weekend

Small model, big takeaways.

What makes CLIP so unique

- Vision + language trained together - not separately
- No fixed labels - matches images to descriptions
- Zero-shot by design - no retraining per task
- One embedding space for text and images