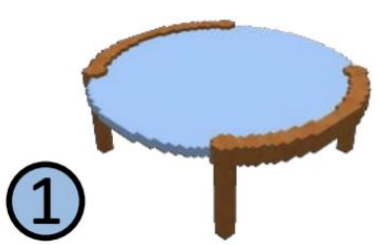
Text and 3D Representation Learning

Presenter: Qirui Wu
2026.3.30

Overview

- Learning text and image representations
- Extending to 3D
- Efficient text and 3D
- Applications

Align text and image representation (CLIP)



1

Circular glass coffee table with two sets of wooden legs that clasp over the round glass edge.

1

2

A brown wooden moon shaped table with three decorative legs with a wooden vine shaped decoration base connecting the legs.

2a

Wooden half round table.

2b

3

Dark brown wooden chair with adjustable back rest and gold printed upholestry.

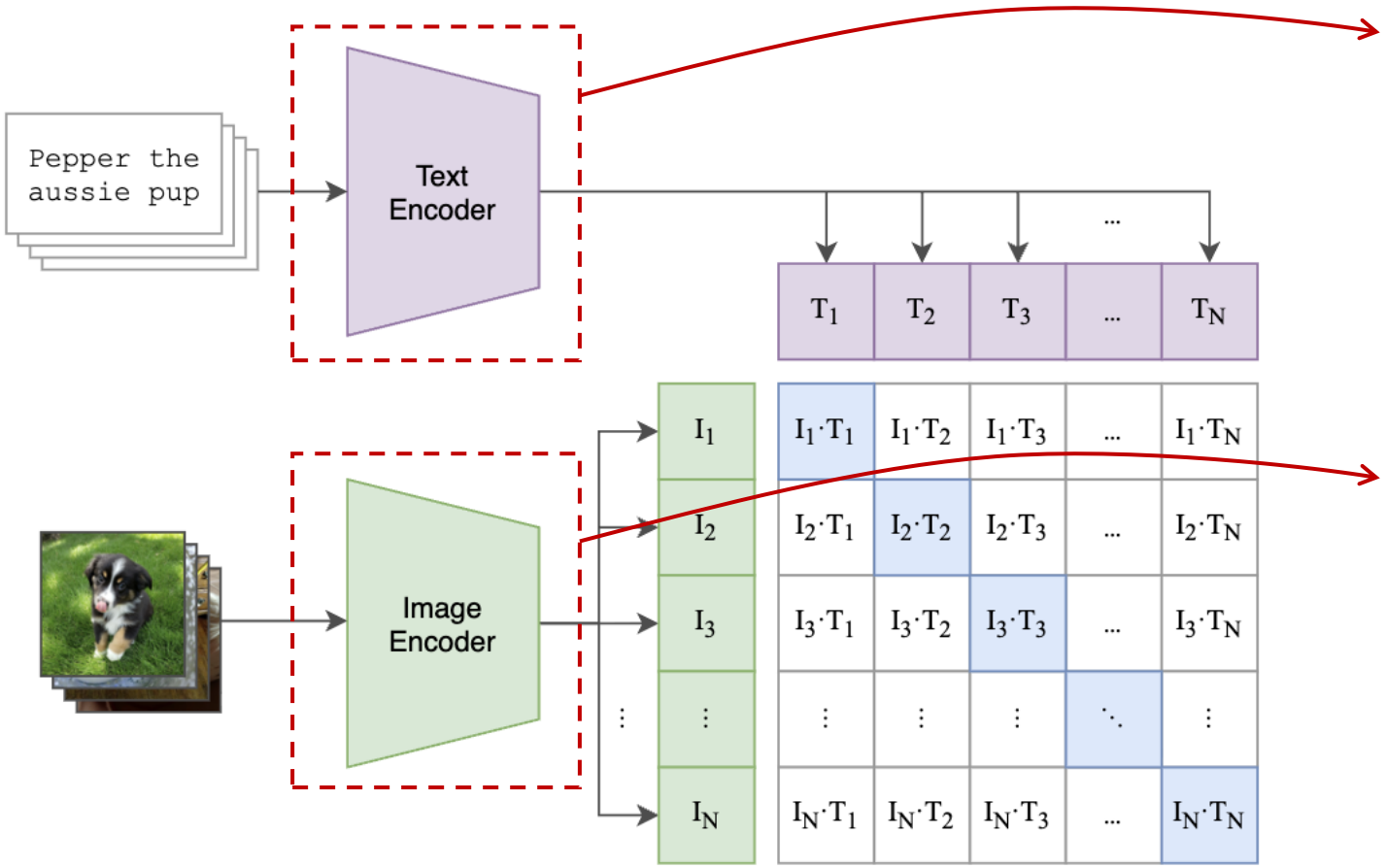
3a

Wooden recliner chair with patterned fabric.

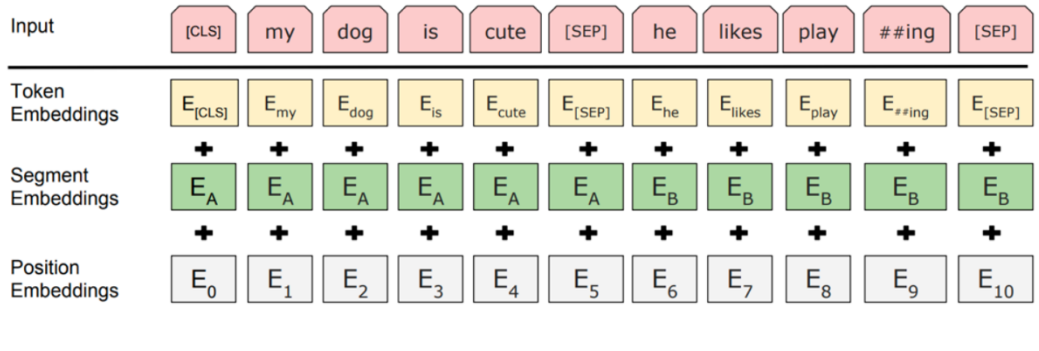
3b

1. Have a bunch of images as well as captions from the Internet
2. How to learn a model to align embeddings for images and text?

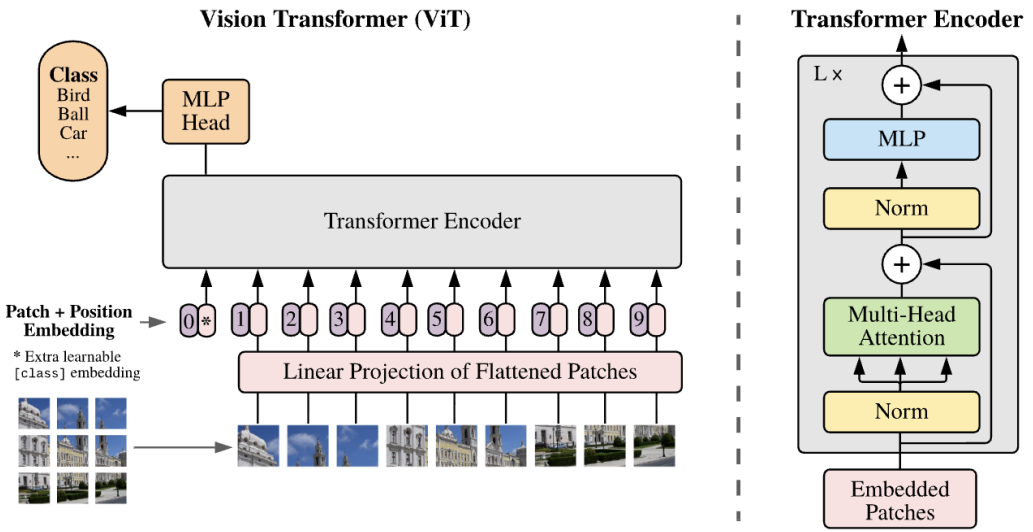
CLIP Preview



CLIP (Radford et al. 2021)

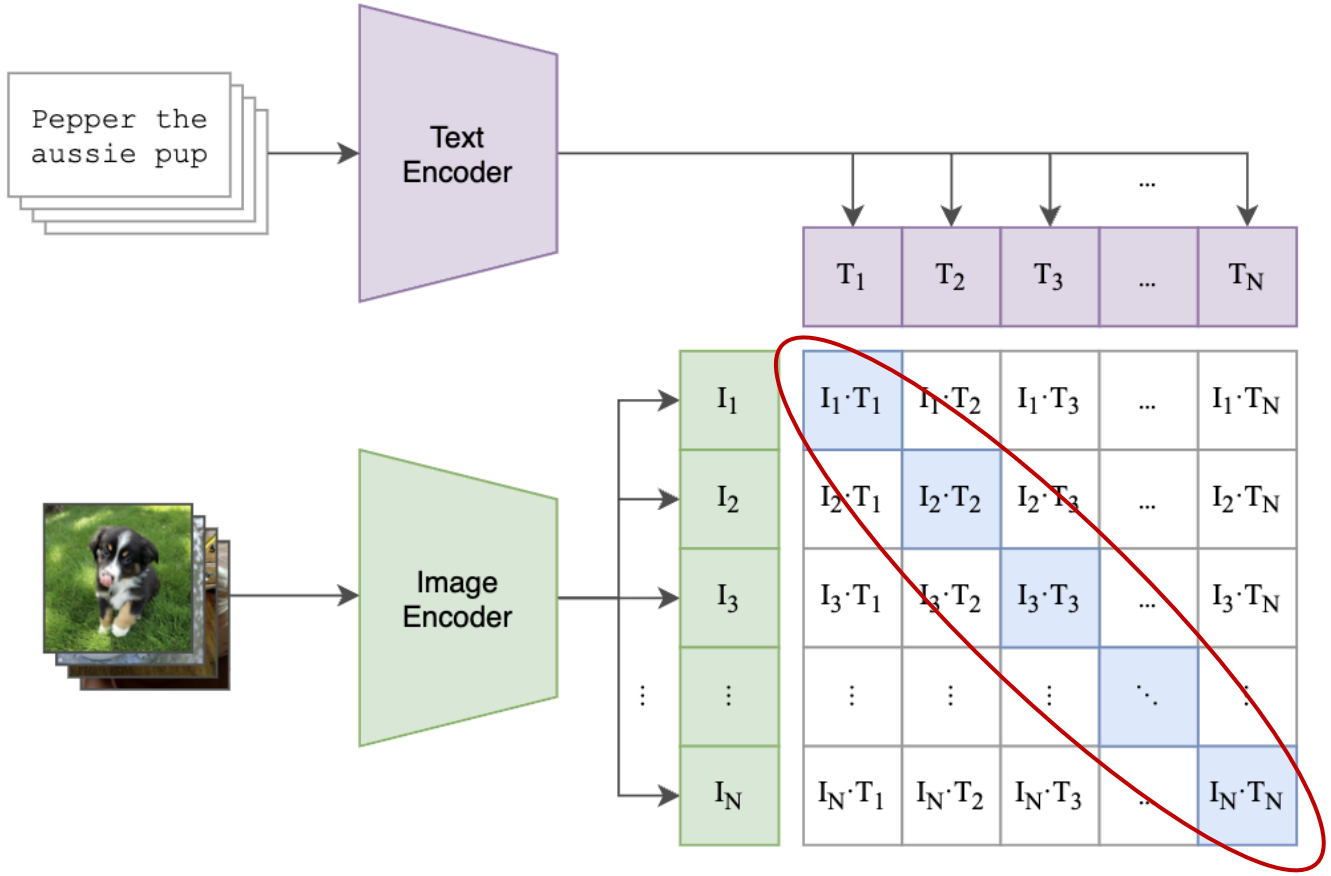


Decoder Only Transformer



ViT (Dosovitskiy et al. 2021)

CLIP Preview



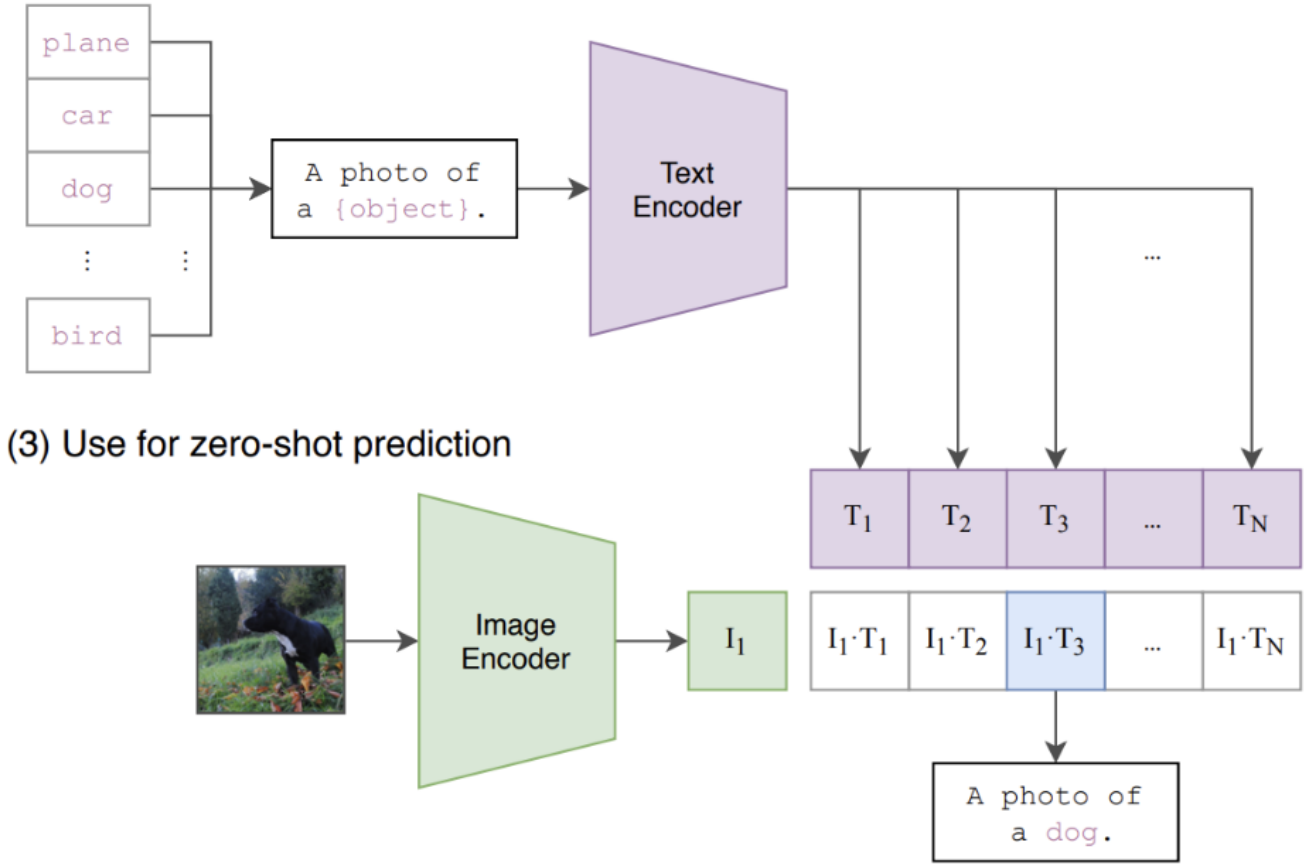
CLIP (Radford et al. 2021)

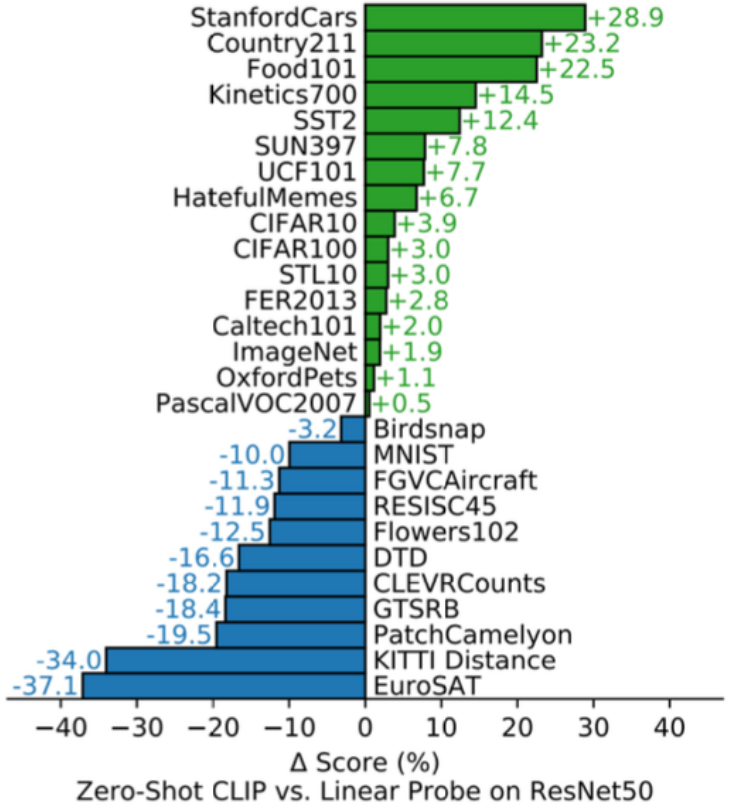
$$\mathcal{L}_{ITC} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(s\left(I_i,T_i\right)/\tau\right)}{\sum_{k\in\mathcal{B}}\exp\left(s\left(I_i,T_k\right)/\tau\right)} - \frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}\log\frac{\exp\left(s\left(I_j,T_j\right)/\tau\right)}{\sum_{k\in\mathcal{B}}\exp\left(s\left(I_k,T_j\right)/\tau\right)}$$

Contrastive Losses

CLIP Preview

Can perform zero-shot inference using text.



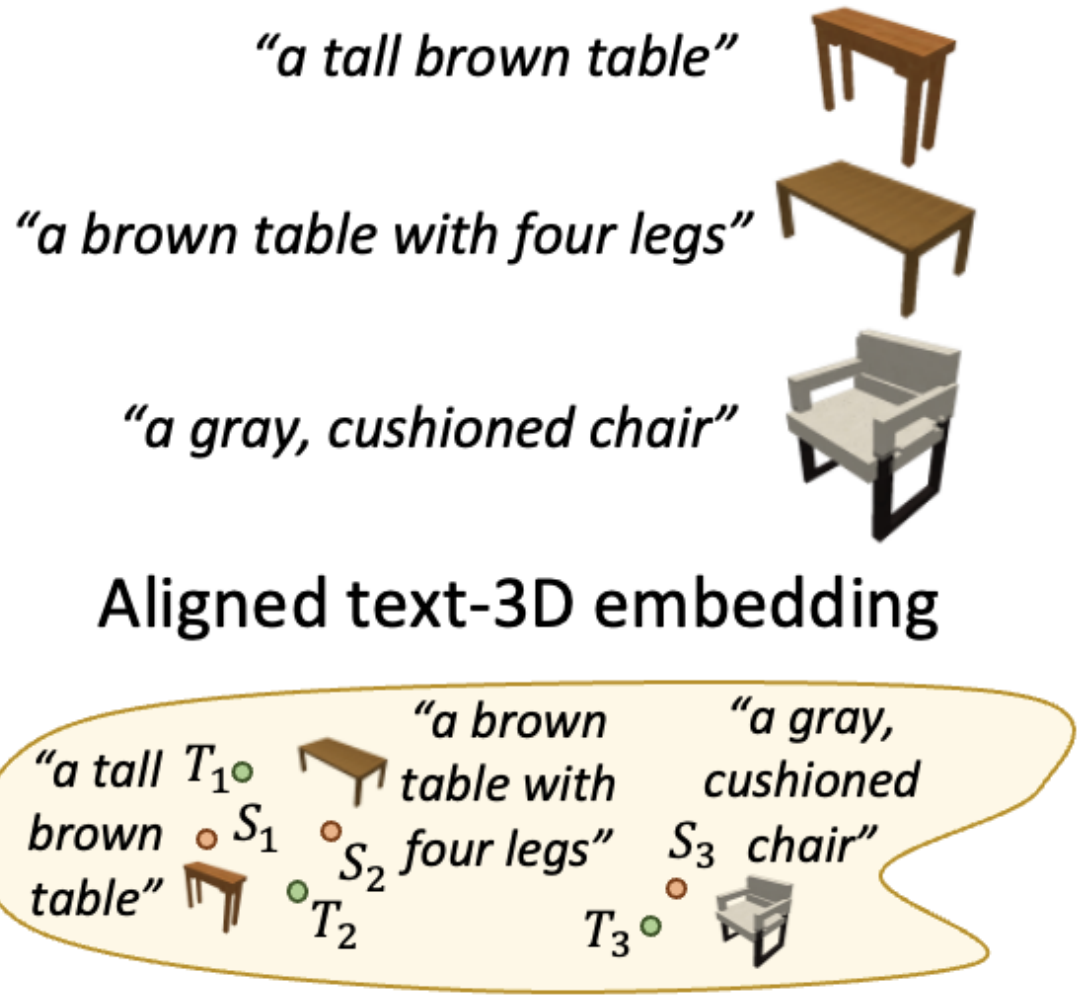


CLIP (Radford et al. 2021)

What about 3D?

Similar story

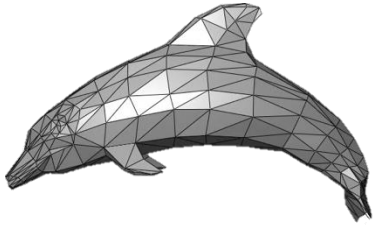
Learn a joint-embedding space between text and 3D.



What about 3D?

But, how to represent a 3D object?

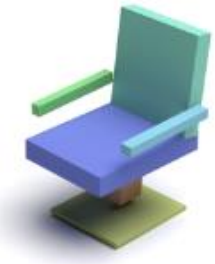
Explicit Representations:



Mesh

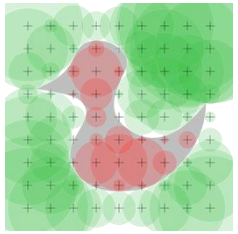
Point Cloud

Voxel

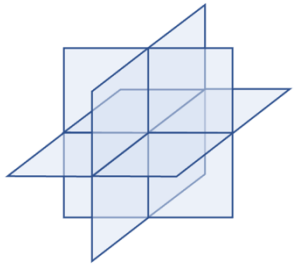


Primitives

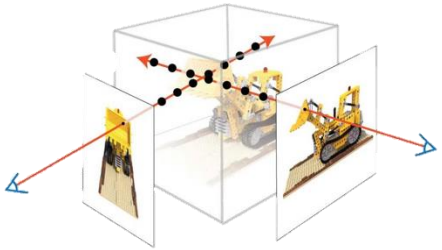
Implicit Representations:



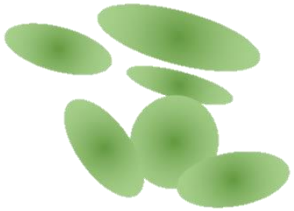
SDF



Tri-plane

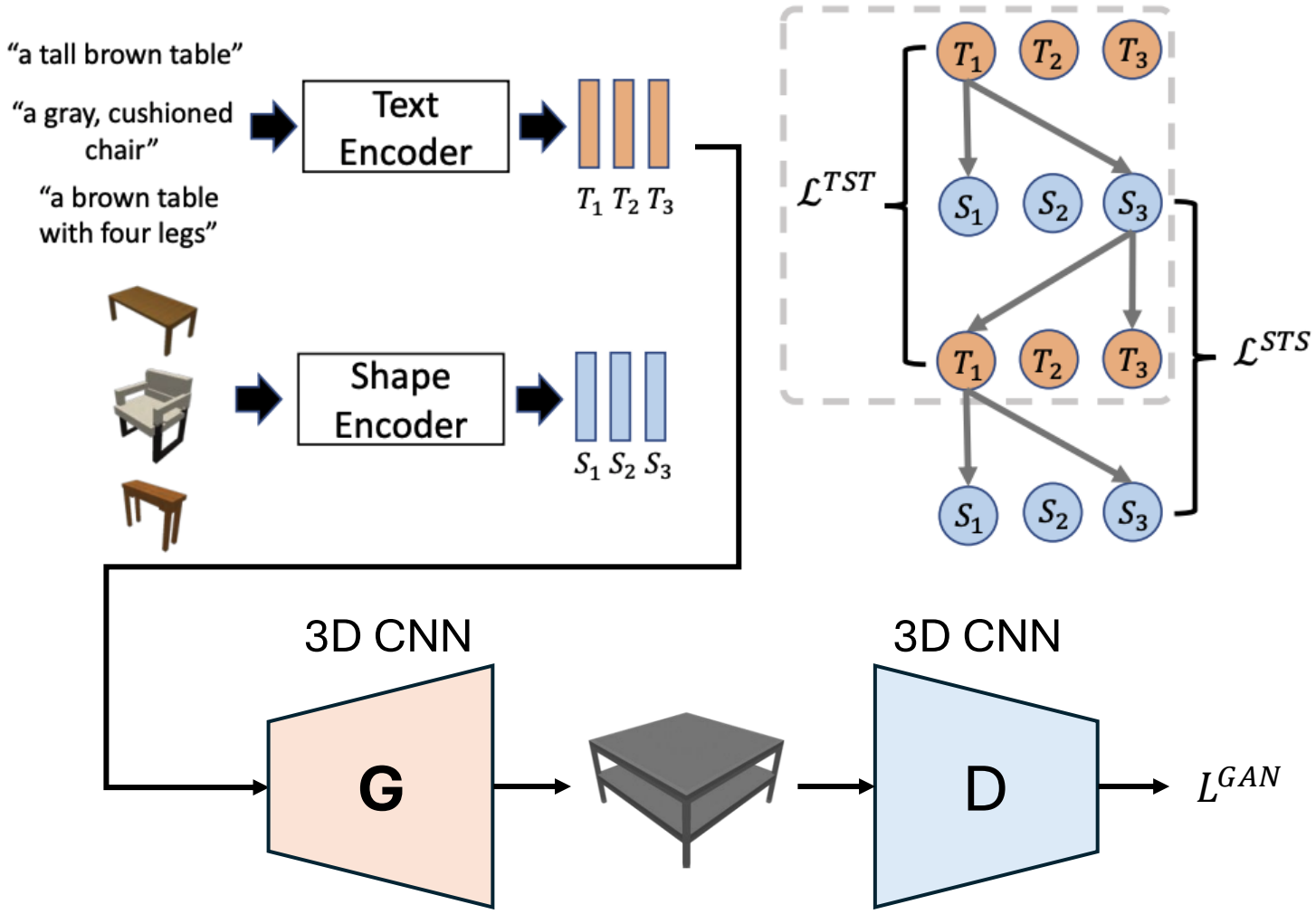


NeRF



3DGS

Early work in 2018



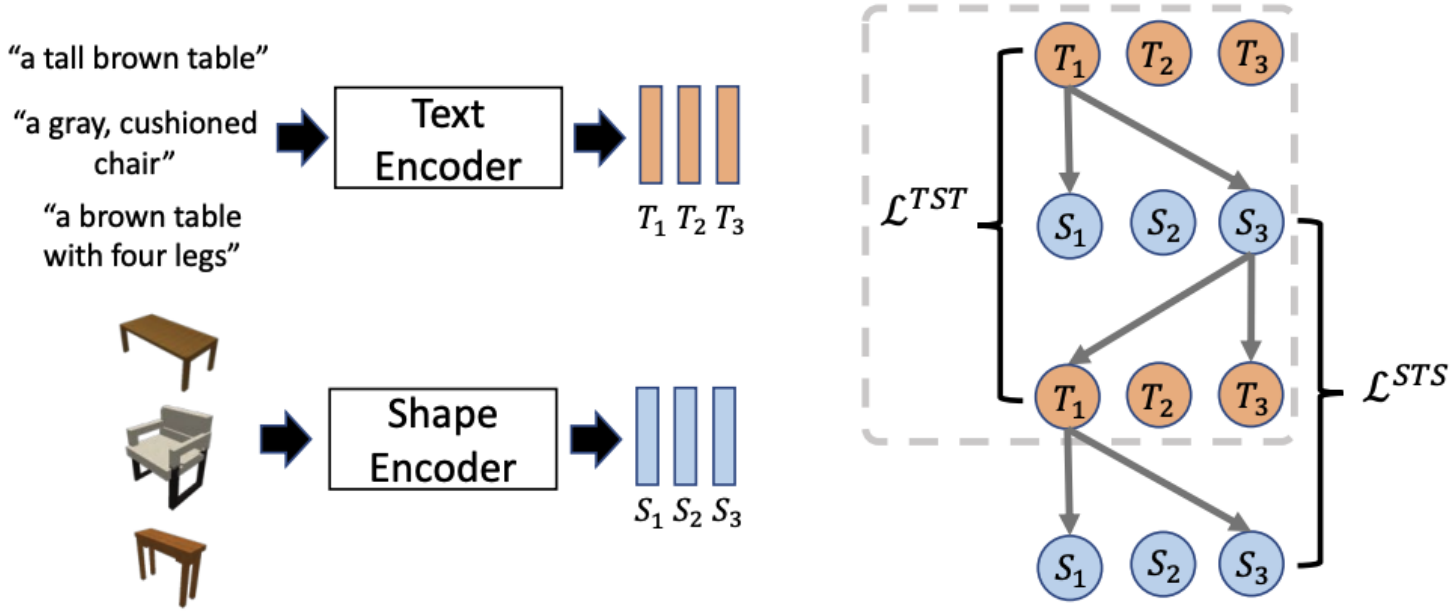
Trained in stages

Metric Learning to align text-shape in latent space

GANs for generating voxels from latent code

Text2Shape (Chen et al. 2018)

Early work in 2018



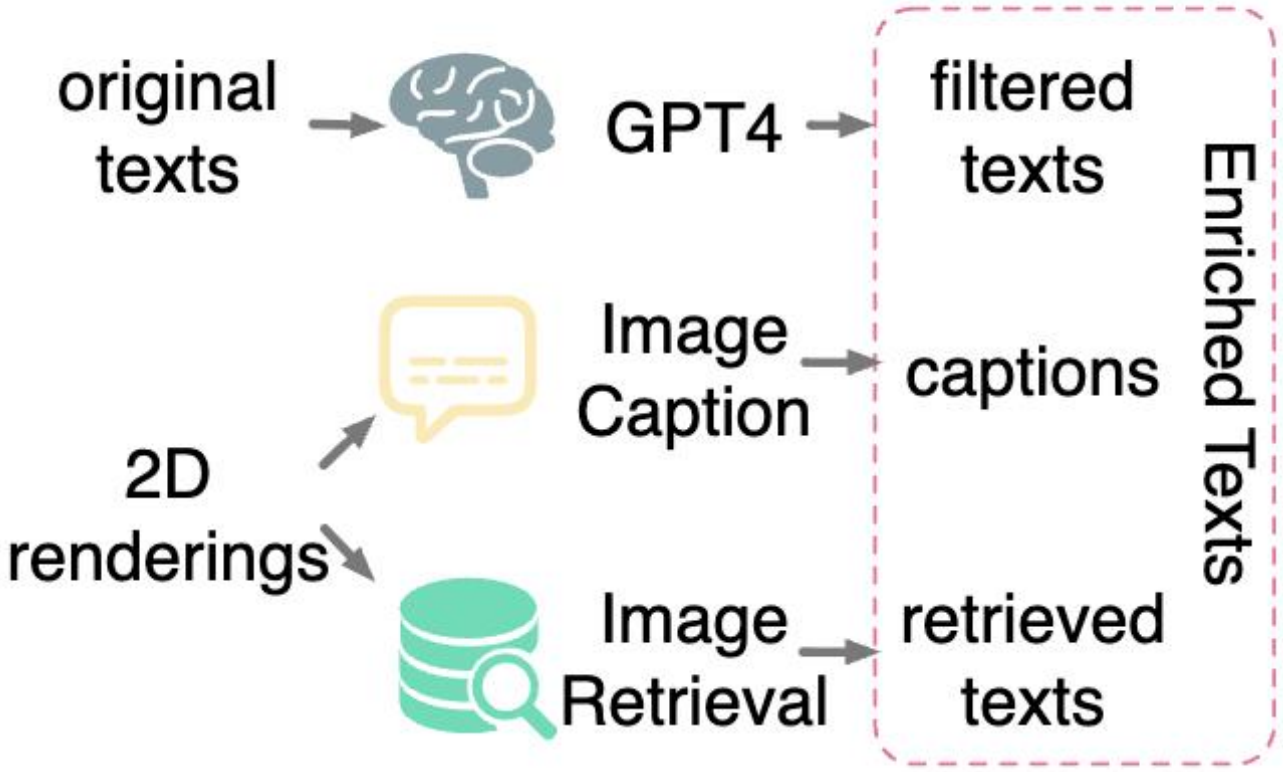
Trained on 75K natural language descriptions for 15K chair and table shapes.

Data scale is a key limitation.

Text2Shape (Chen et al. 2018)

How to scale up data?

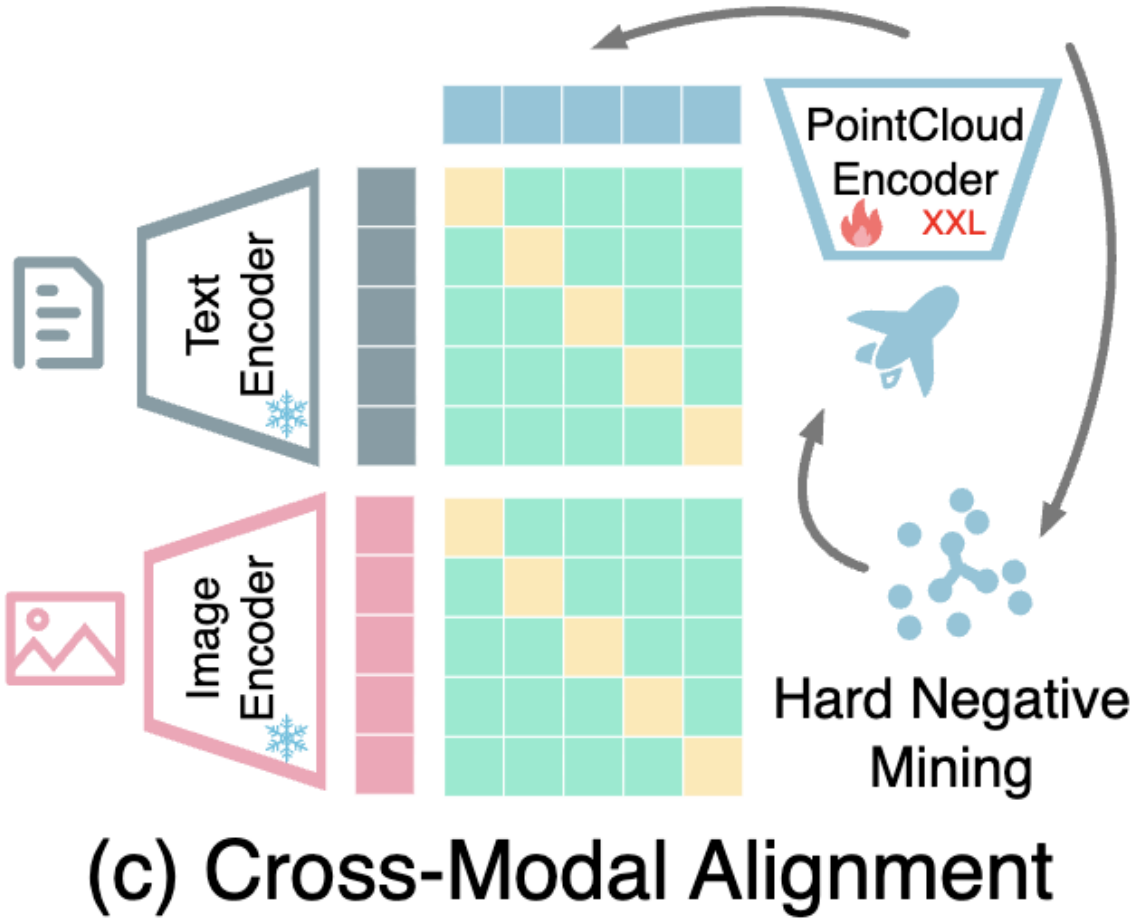
(a) Ensemble Datasets



(b) Text Filtering & Enrichment

OpenShape (Liu et al. 2023)

How to scale up data?



(c) Cross-Modal Alignment

$$l_i^{a \to b} = -\log \frac{\exp(\langle f_i^a, f_i^b \rangle)/\tau}{\Sigma_{k=1}^{N} \exp(\langle f_i^a, f_k^b \rangle)/\tau}$$

$$L_{CON} = \frac{1}{4N} \Sigma_{i=1}^{N} (l_i^{S \to T} + l_i^{T \to S} + l_i^{S \to I} + l_i^{I \to S})$$

OpenShape (Liu et al. 2023)

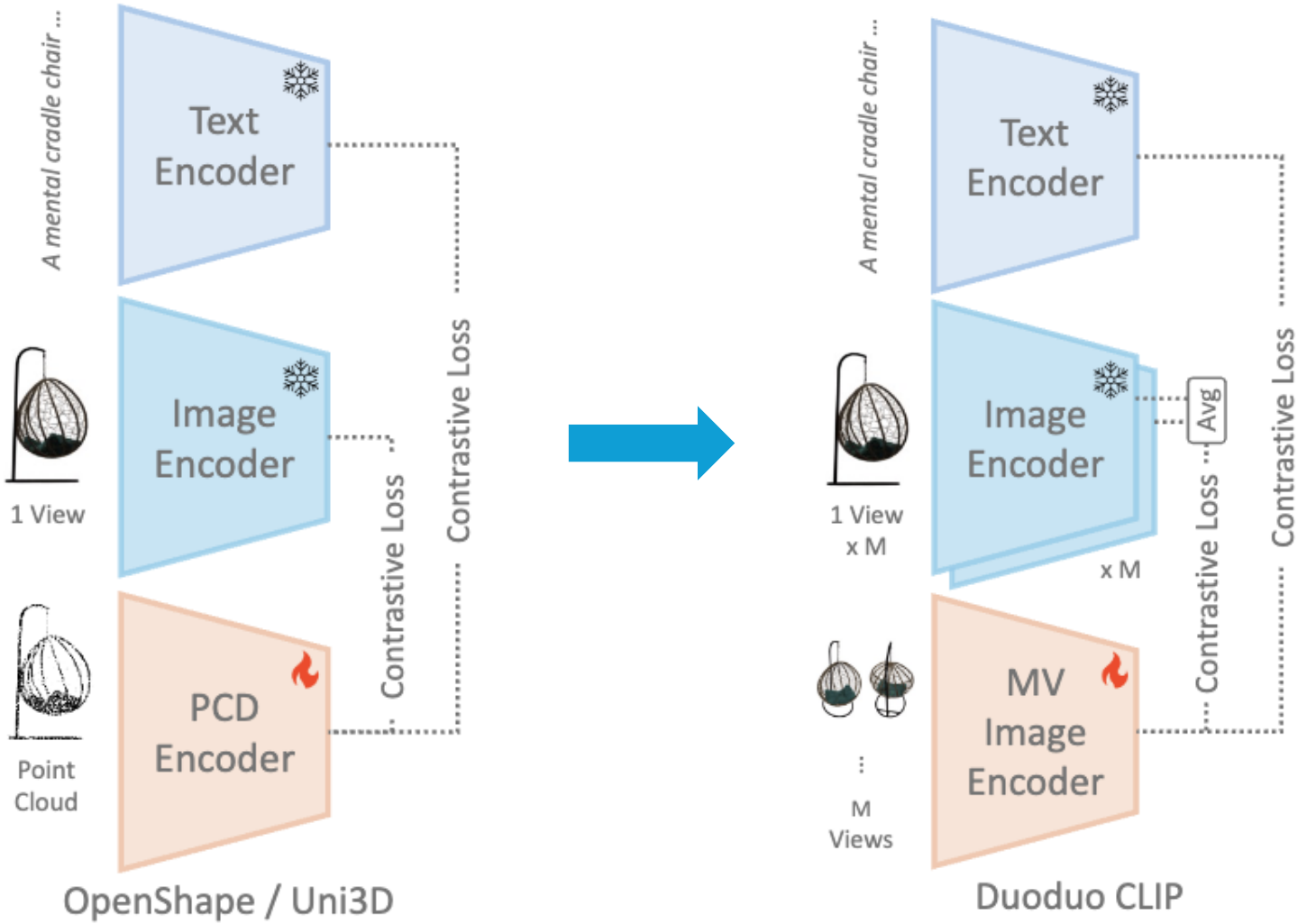
How to do better?

- Point clouds are harder to acquire for real world objects.
- Domain gap between images and point clouds.

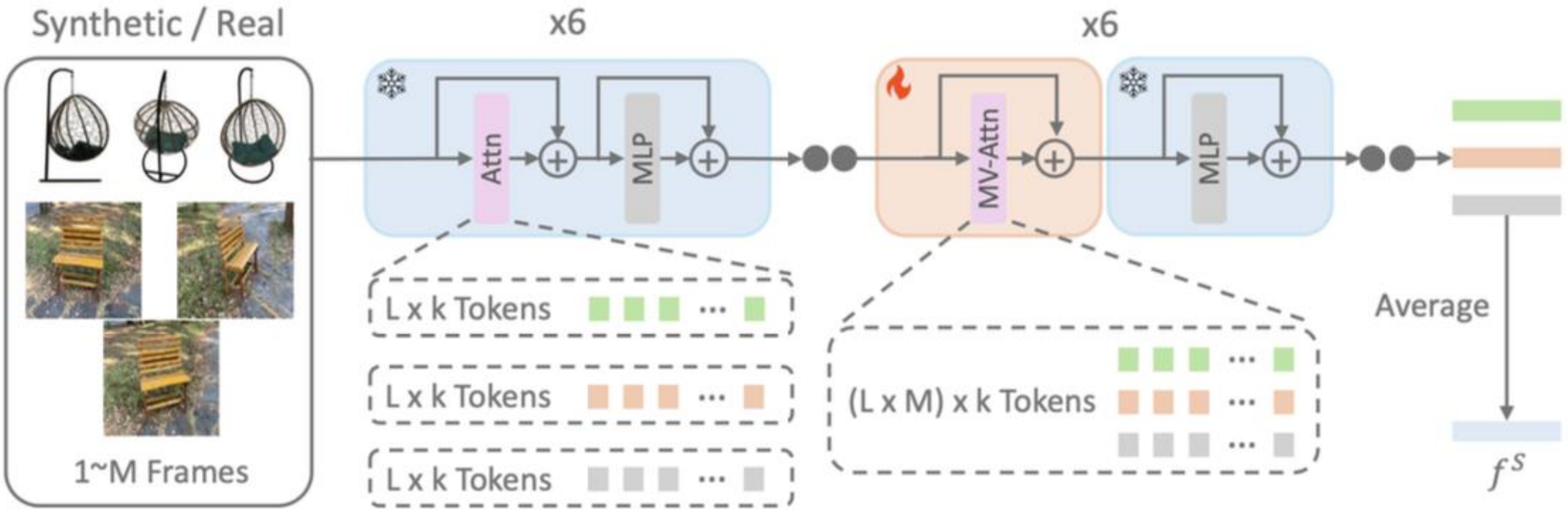
1~M Frames

Use multi-view images instead!

DuoduoCLIP


A mental cradle chair ...
Text Encoder
Image Encoder
1 View
PCD Encoder
Point Cloud
Contrastive Loss
Contrastive Loss
OpenShape / Uni3D
A mental cradle chair ...
Text Encoder
Image Encoder
1 View x M
x M
Avg
Contrastive Loss
MV Image Encoder
M Views
Contrastive Loss
Duoduo CLIP

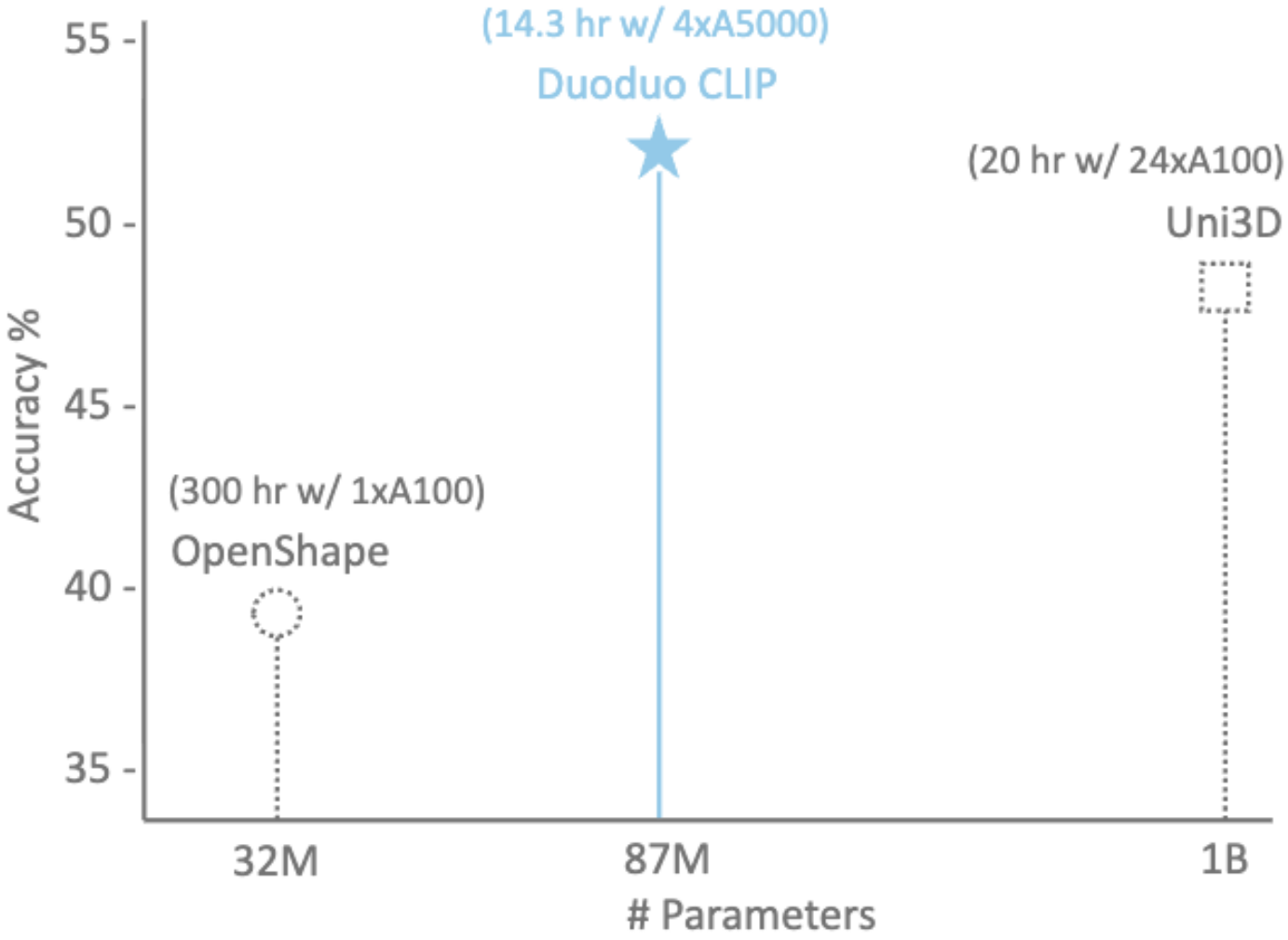
DuoduoCLIP



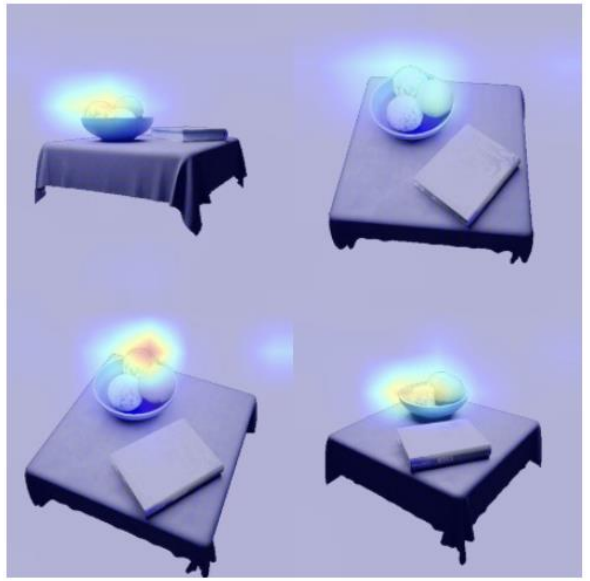
$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$Q,K,V \in R^{L\times k} \qquad Q^{\text{MV}},K^{\text{MV}},V^{\text{MV}} \in R^{(L\times m)\times k}$$

DuoduoCLIP



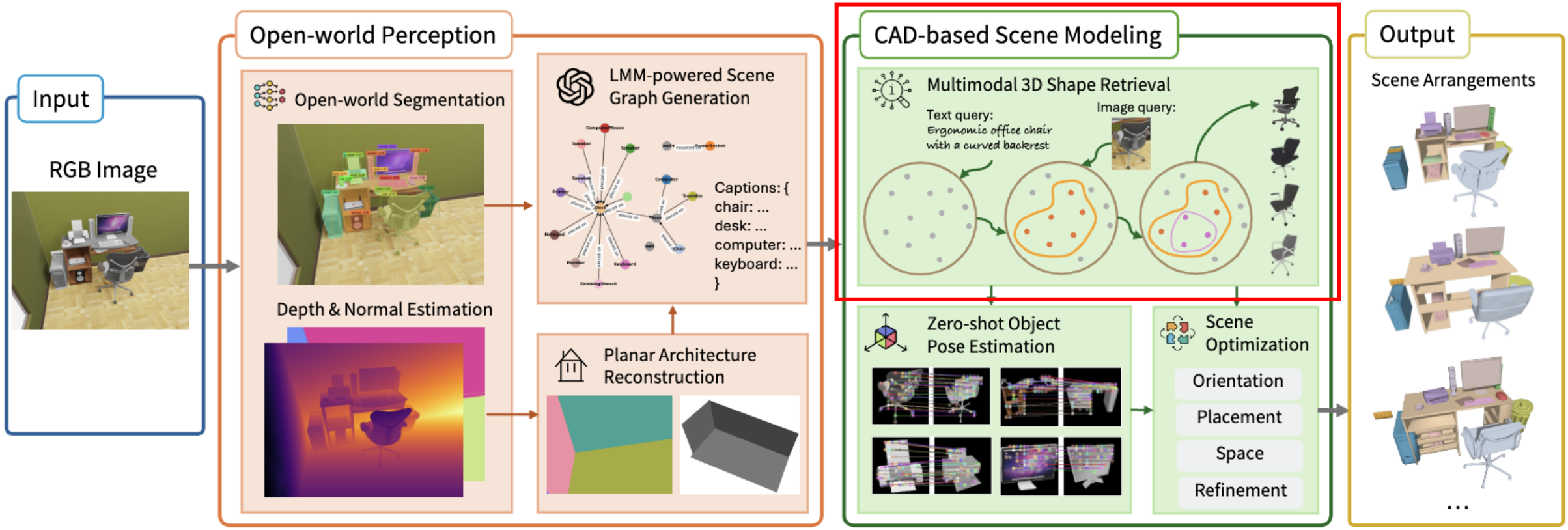
Better generalization on unseen shapes!



Applications (Digital Twin)

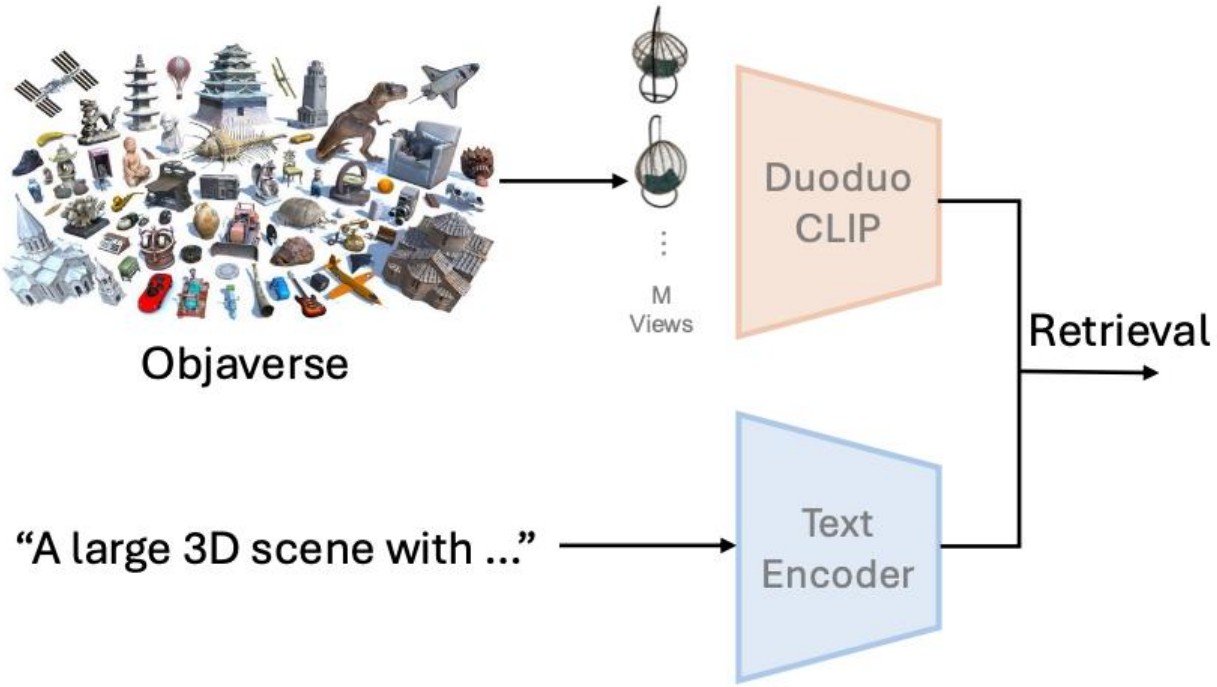
Diorama (Wu et al. 2025)

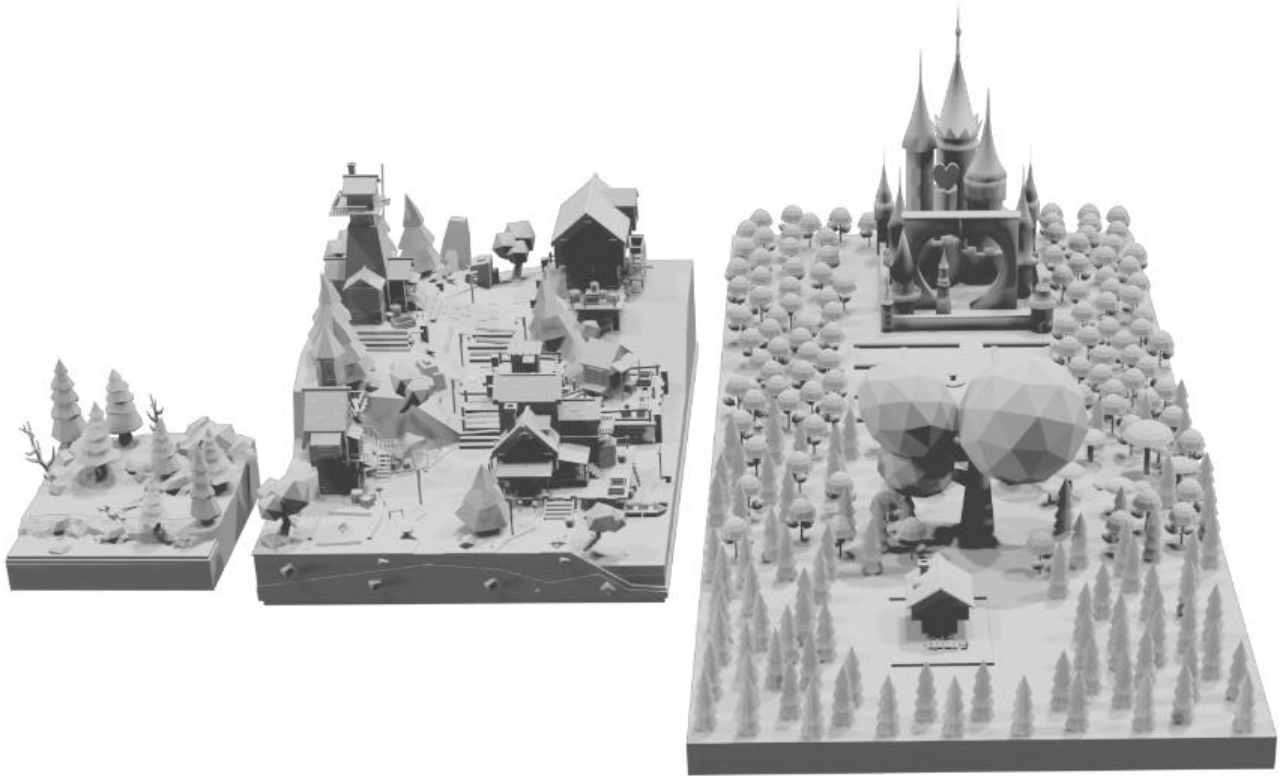
Applications (Digital Twin)



Diorama (Wu et al. 2025)

Applications (Dataset Filtering)





NuiScene (Lee et al. 2025)

Summary

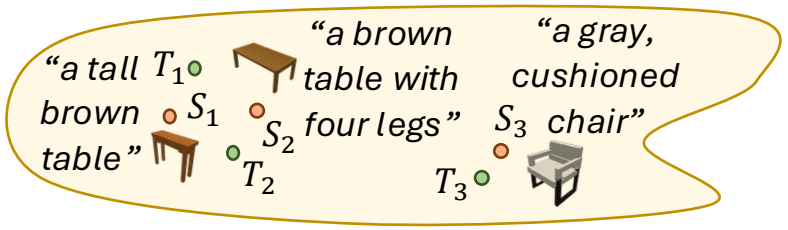
3D with Paired Text (3DPT)

"a tall brown table"

"a brown table with four legs"

"a gray, cushioned chair"

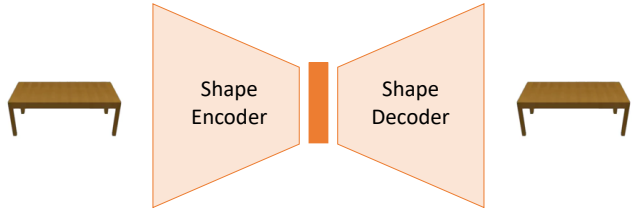
Aligned text-3D embedding



3D with Unpaired Text (3DUT)

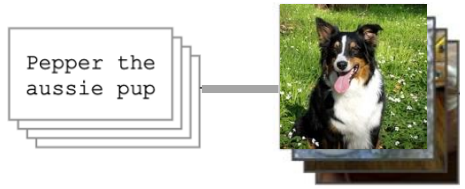
3D shape corpus

Learned 3D shape priors



No 3D data (No3D)

Large text-image corpus → Large vision-language model



CLIP | DALLE-2 | ALIGN | IMAGEN

Prompt-based optimization of **differentiable** 3D representation

NeRF

DMTet