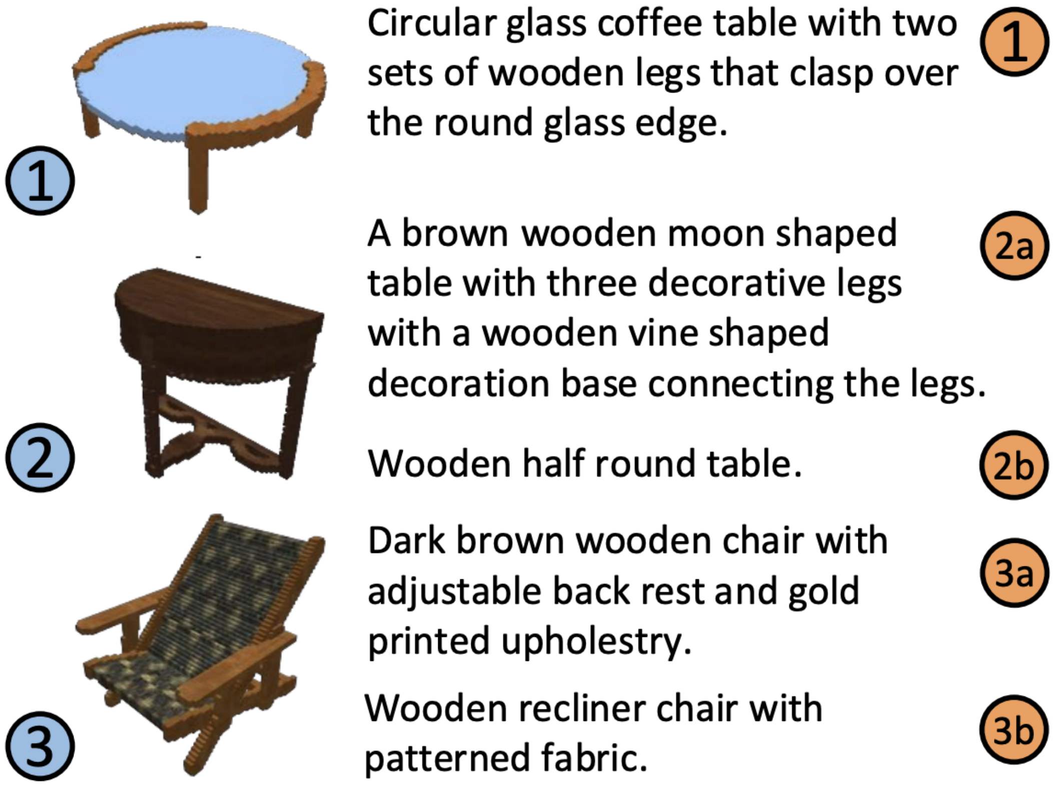
# Text and 3D Representation Learning

# Overview

1. Learning text and image representations
2. Extending to text and 3D
3. Efficient text and 3D
4. Applications

# Preliminary (CLIP)



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

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

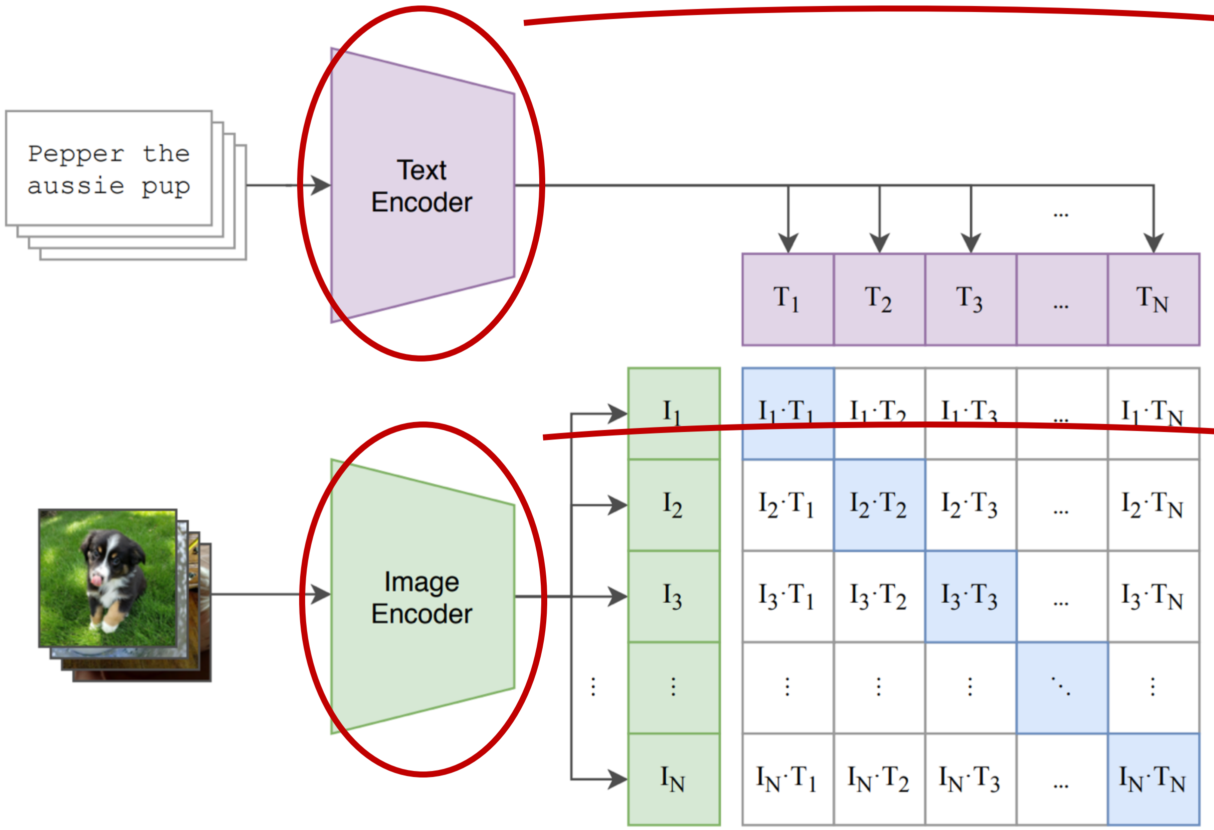
Wooden half round table.

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

Wooden recliner chair with patterned fabric.

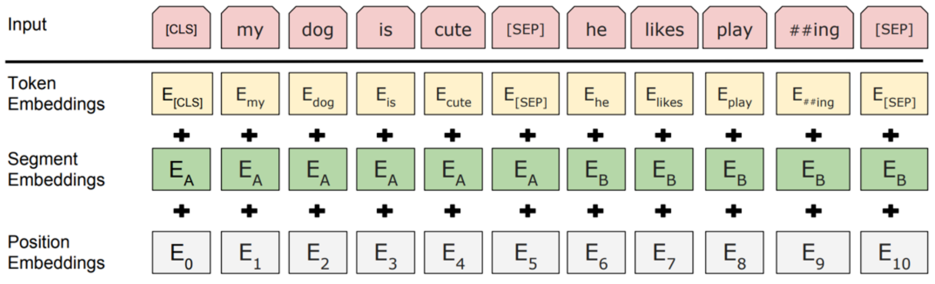
1. Have a bunch of images as well as captions from the Internet
2. How to learn a model to align embeddings for images and text?

# Preliminary (CLIP)



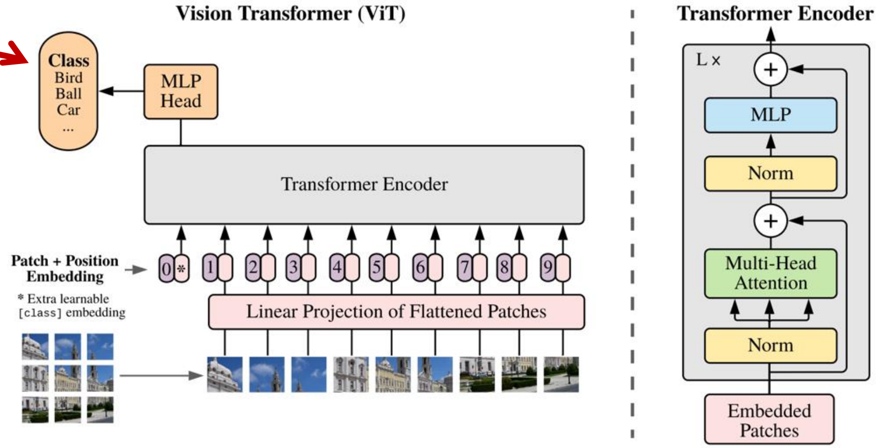
|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

CLIP (Radford et al. 2021)



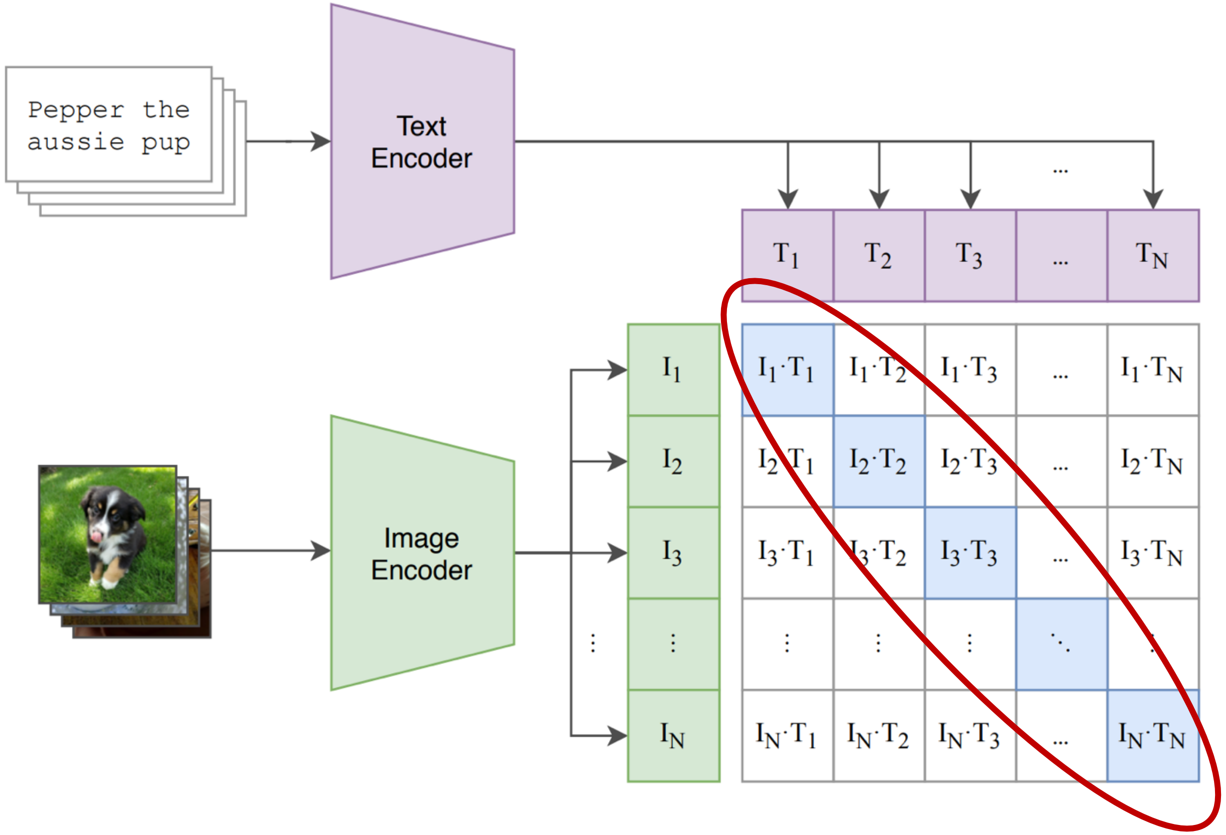
| Input | [CLS] | my | dog | is | cute | [SEP] | he | likes | play | ##ing | [SEP] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Token Embeddings | $E_{[CLS]}$ | $E_{my}$ | $E_{dog}$ | $E_{is}$ | $E_{cute}$ | $E_{[SEP]}$ | $E_{he}$ | $E_{likes}$ | $E_{play}$ | $E_{\#\#ing}$ | $E_{[SEP]}$ |
|  | + | + | + | + | + | + | + | + | + | + | + |
| Segment Embeddings | $E_A$ | $E_A$ | $E_A$ | $E_A$ | $E_A$ | $E_A$ | $E_B$ | $E_B$ | $E_B$ | $E_B$ | $E_B$ |
|  | + | + | + | + | + | + | + | + | + | + | + |
| Position Embeddings | $E_0$ | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ | $E_6$ | $E_7$ | $E_8$ | $E_9$ | $E_{10}$ |

Decoder Only Transformer



ViT (Dosovitskiy et al. 2021)

# Preliminary (CLIP)



|  | $T_1$ | $T_2$ | $T_3$ | ... | $T_N$ |
|---|---|---|---|---|---|
| $I_1$ | $I_1 \cdot T_1$ | $I_1 \cdot T_2$ | $I_1 \cdot T_3$ | ... | $I_1 \cdot T_N$ |
| $I_2$ | $I_2 \cdot T_1$ | $I_2 \cdot T_2$ | $I_2 \cdot T_3$ | ... | $I_2 \cdot T_N$ |
| $I_3$ | $I_3 \cdot T_1$ | $I_3 \cdot T_2$ | $I_3 \cdot T_3$ | ... | $I_3 \cdot T_N$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| $I_N$ | $I_N \cdot T_1$ | $I_N \cdot T_2$ | $I_N \cdot T_3$ | ... | $I_N \cdot T_N$ |

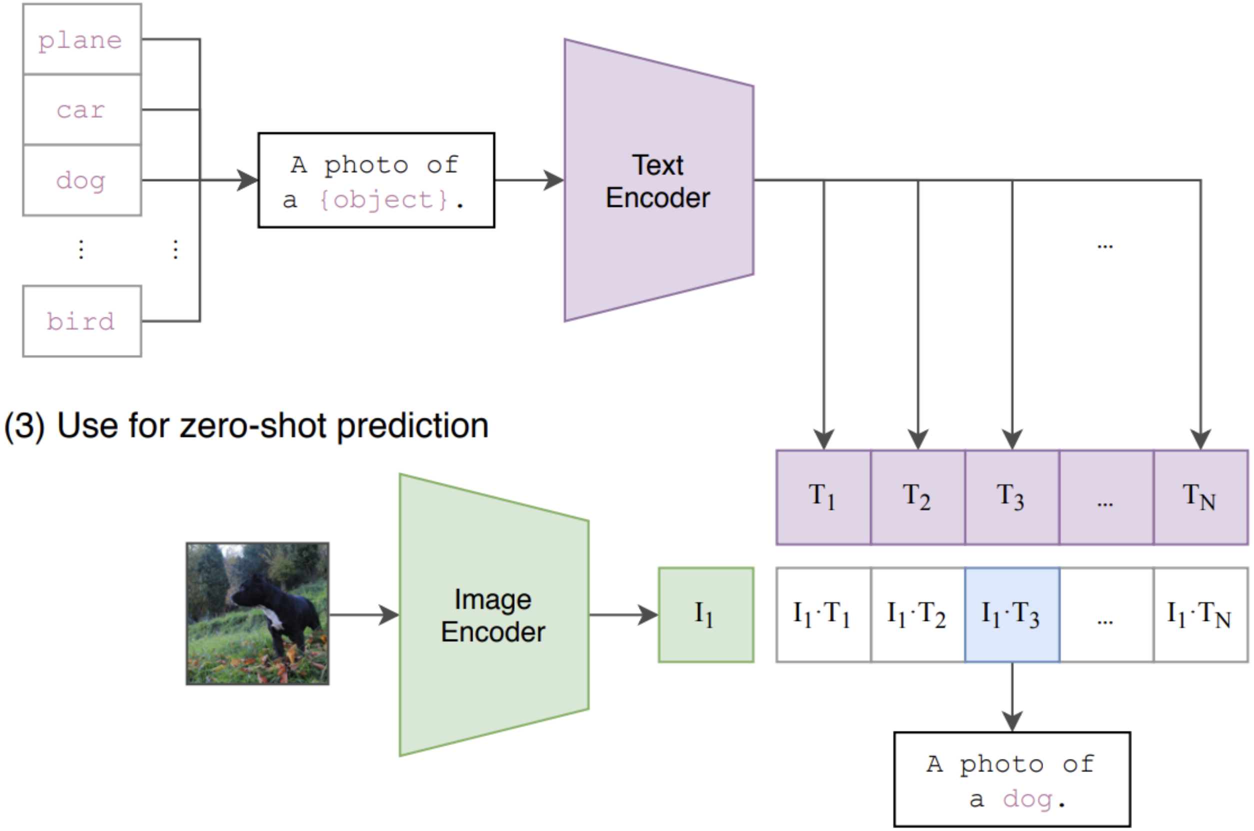
CLIP (Radford et al. 2021)

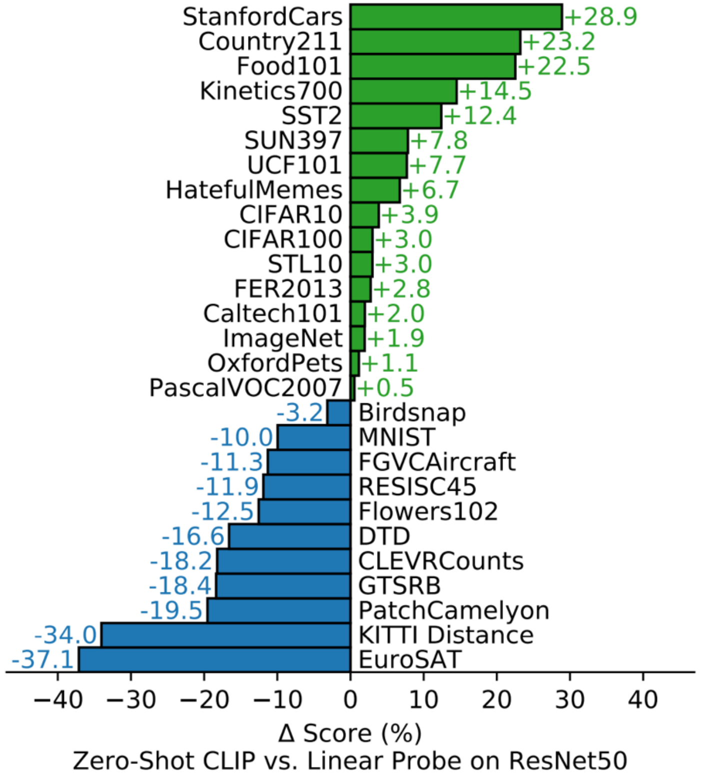
$$
\begin{aligned}
\mathcal{L}_{ITC} = &-\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(s\left(I_i, T_i\right)/\tau\right)}{\sum_{k\in\mathcal{B}}\exp\left(s\left(I_i, T_k\right)/\tau\right)} \\
&-\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}\log\frac{\exp\left(s\left(I_j, T_j\right)/\tau\right)}{\sum_{k\in\mathcal{B}}\exp\left(s\left(I_k,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

Contrastive Losses

# Preliminary (CLIP)

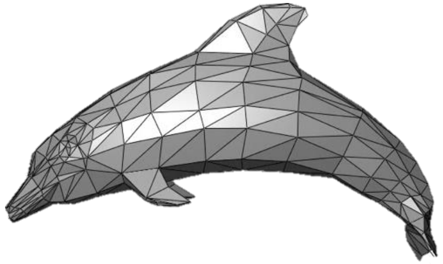
Can perform zero-shot inference using text.





CLIP (Radford et al. 2021)

# Extend to 3D and Text?



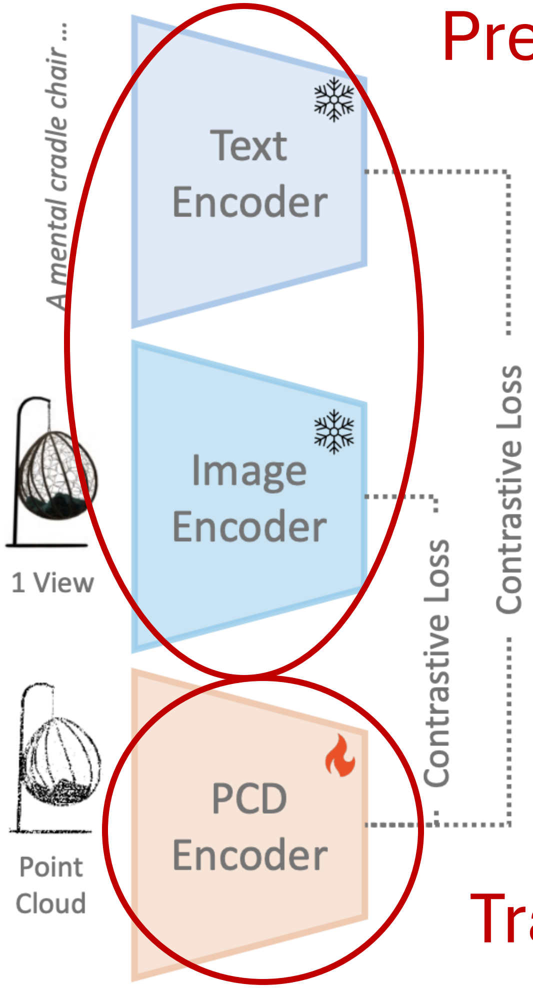
Mesh

Point Cloud

Voxel

Primitives

# Extend to 3D and text?



OpenShape / Uni3D

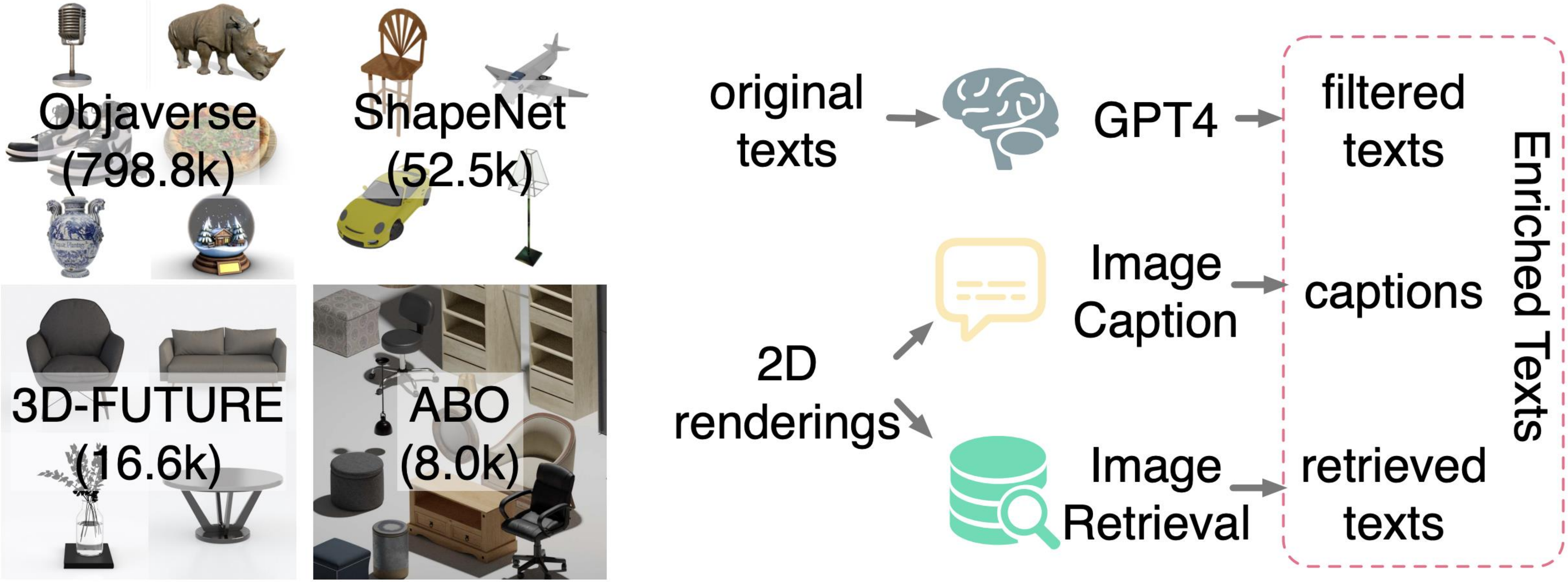
Pretrained CLIP Models

Trainable Point Bert Model

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

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

# Data



(a) Ensemble Datasets (b) Text Filtering & Enrichment

OpenShape (Liu et al. 2023)

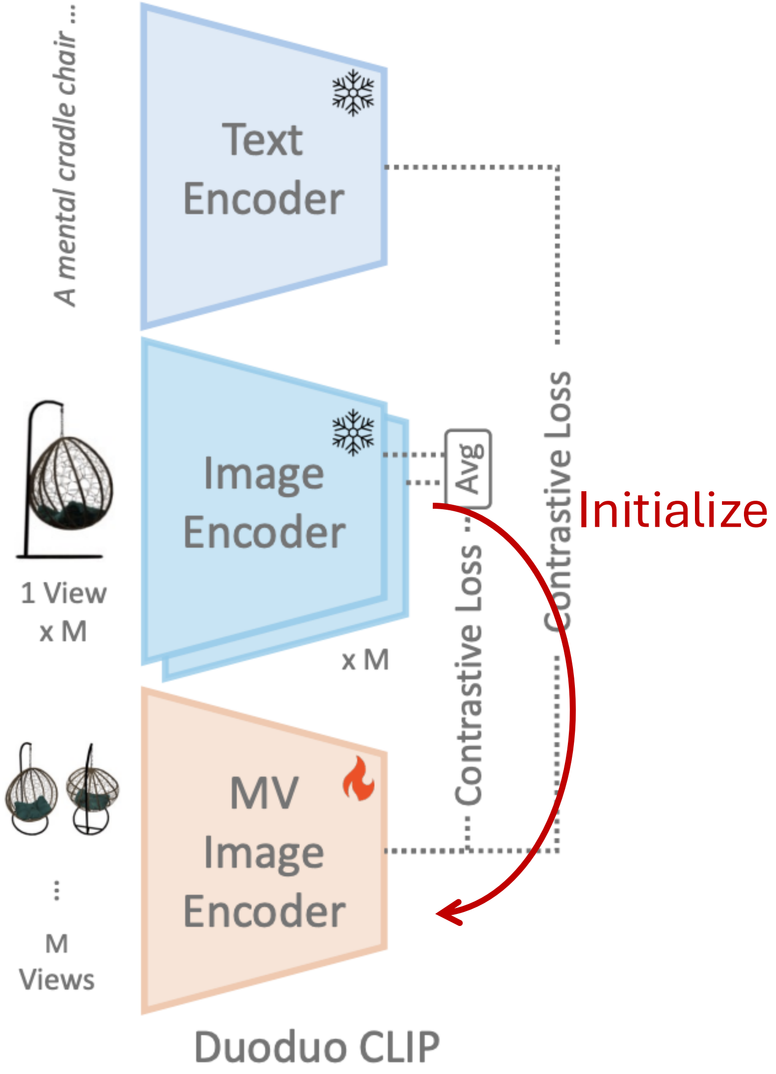
# More Efficient Training?

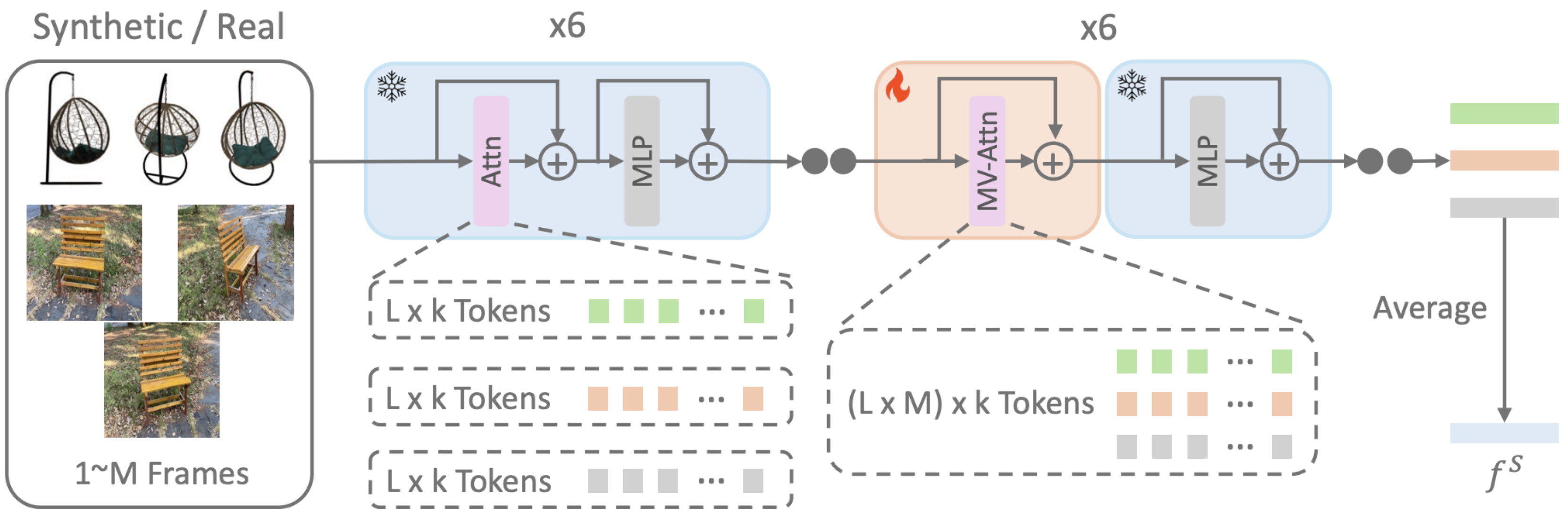
1. Point clouds are harder to acquire for real world objects
2. Domain gap between images and point clouds

1~M Frames

Use multi-view images instead!

# DuoduoCLIP





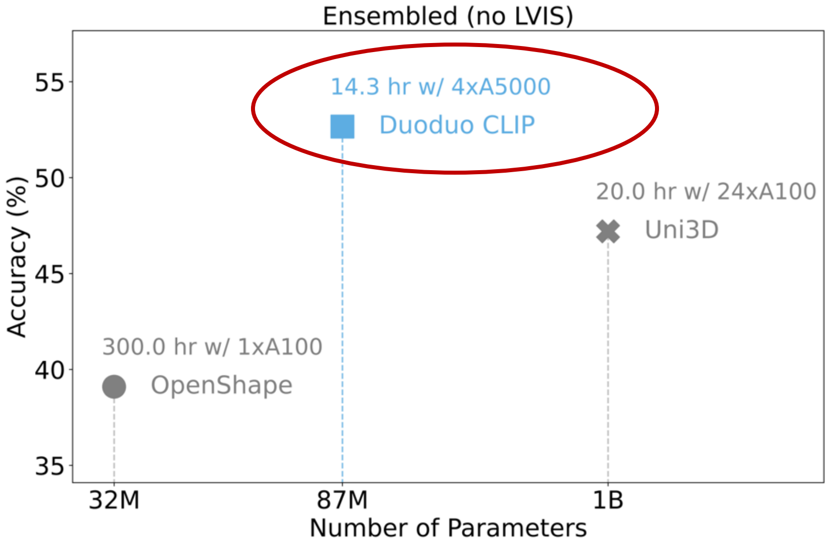
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$Q, K, V \in R^{L\times k}$$

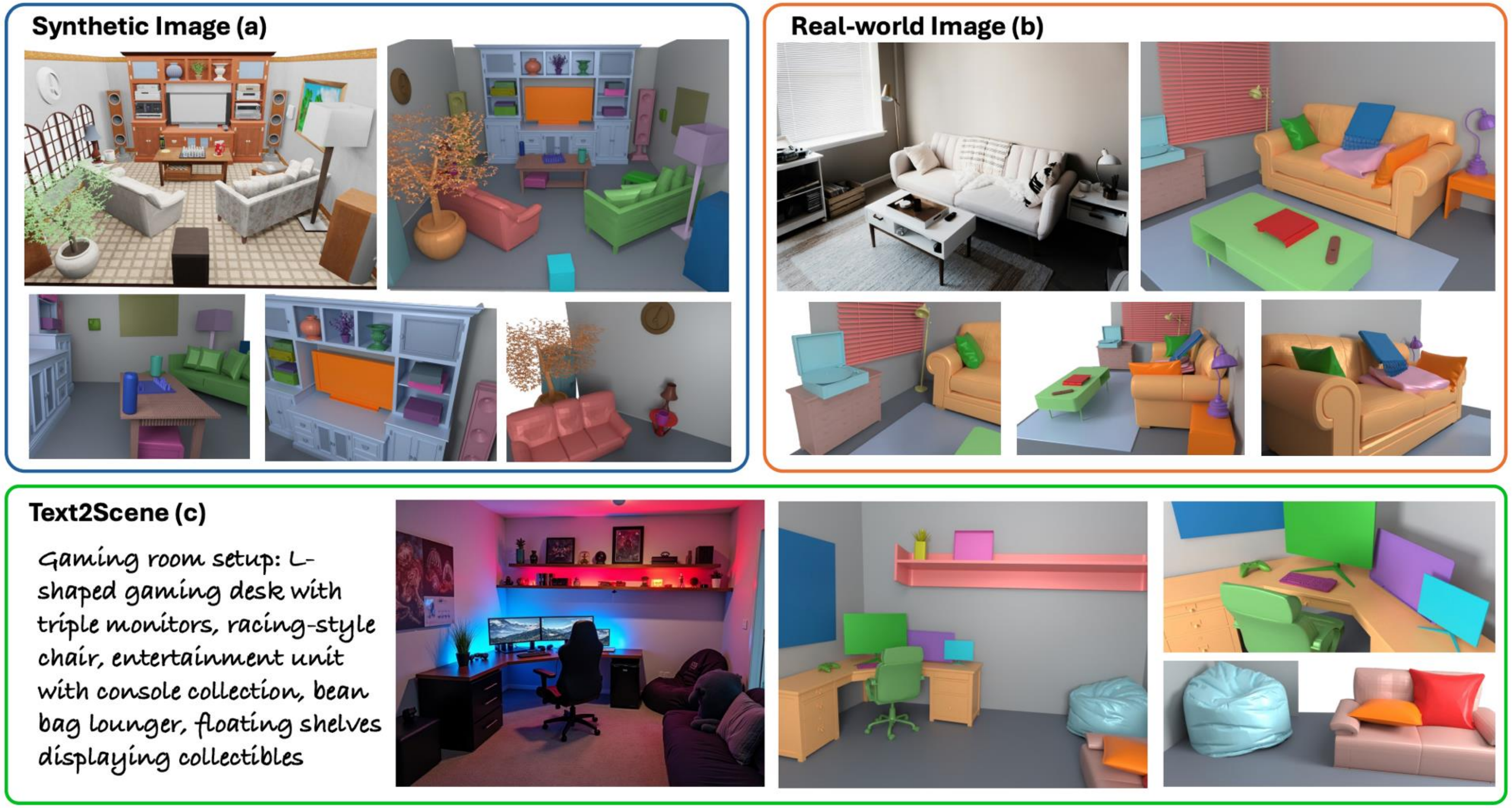
$$Q^{\text{MV}}, K^{\text{MV}}, V^{\text{MV}} \in R^{(L\times m)\times k}$$

# DuoduoCLIP

Better generalization on unseen shapes!

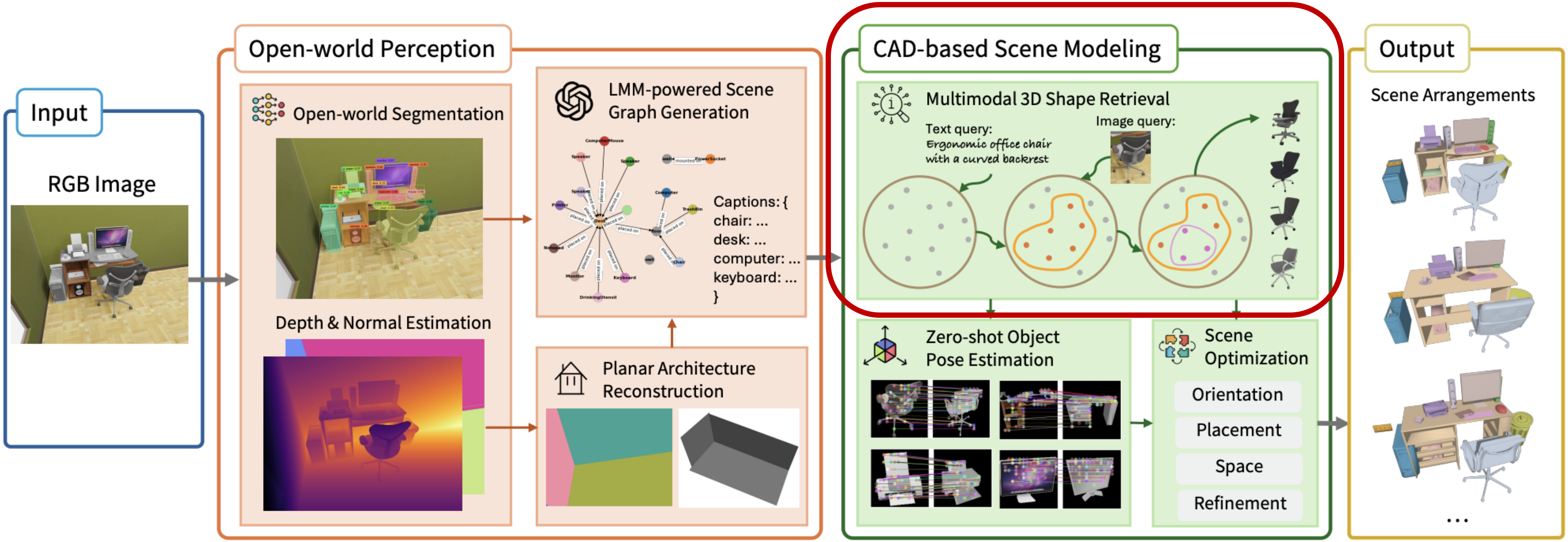


# Applications (Digital Twin)


Synthetic Image (a)
Real-world Image (b)
Text2Scene (c)
Gaming room setup: L-shaped gaming desk with triple monitors, racing-style chair, entertainment unit with console collection, bean bag lounger, floating shelves displaying collectibles

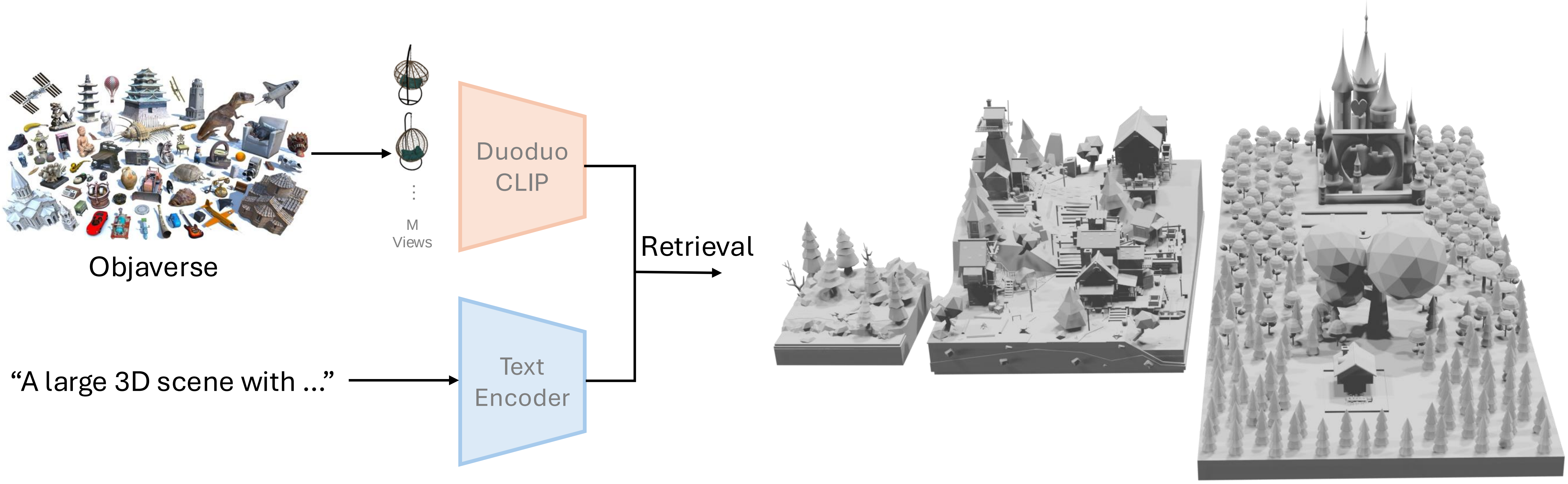
Diorama (Wu et al. 2025)

# Applications (Digital Twin)



Diorama (Wu et al. 2025)

# Applications (Dataset Filtering)



Objaverse

M
Views

Duoduo
CLIP

Retrieval

“A large 3D scene with …”

Text
Encoder

NuiScene (Lee et al. 2025)

# Applications (Dataset Filtering)

NuiScene (Lee et al. 2025)

# Check out the projects!

DuoduoCLIP

Diorama

NuiScene