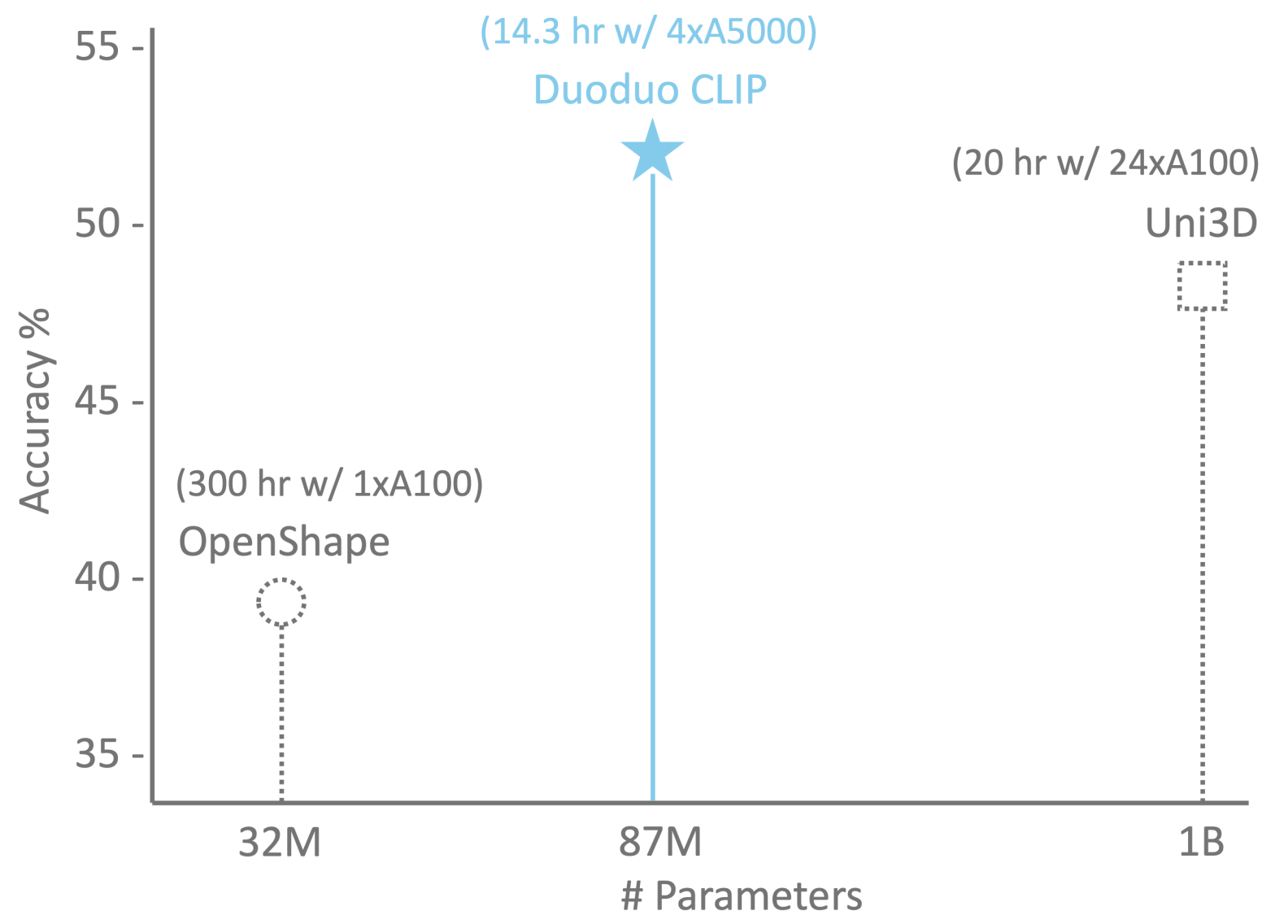
Duoduo CLIP: Efficient 3D Understanding with Multi-View Images

Han-Hung Lee[1,*], Yiming Zhang[1,*], Angel X. Chang[1,2]
[1]Simon Fraser University, [2]Canada CIFAR AI Chair, Amii

1. Introduction

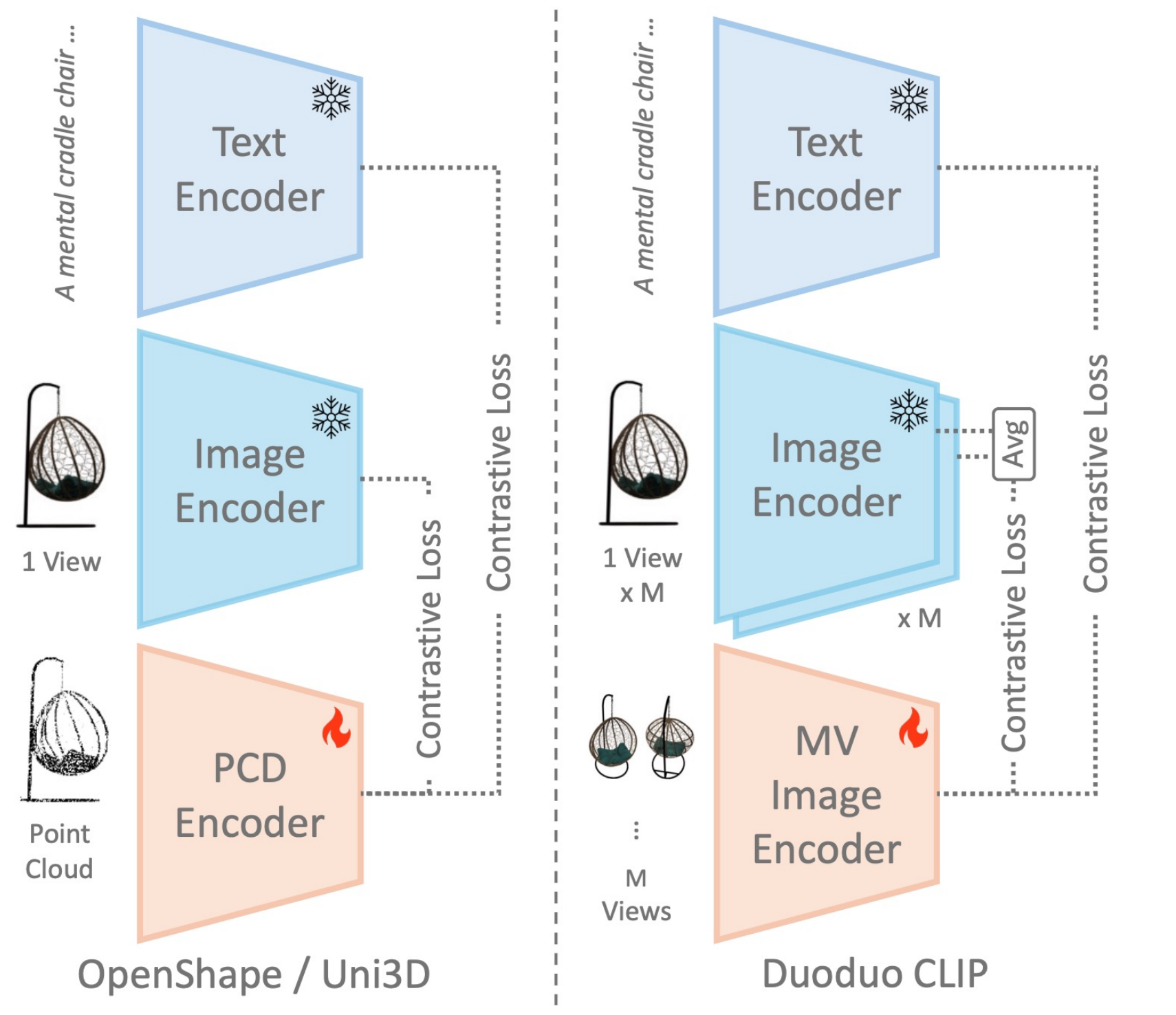
- **Task**: text and 3D alignment.
- **Motivation**: point clouds are harder to obtain and compute intensive to train.

Method	GPU	Time
OpenShape Liu et al. (2023a)	1×A100 (80GB)	300 hr
Uni3D Zhou et al. (2024)	24×A100 (40GB)	20 hr
RECON++ Qi et al. (2024)	8×A800 (80GB)	1 day
Ours (Full)	4×A40 (48GB)	14.3 hr
Ours (6 layers)	4×A5000 (24GB)	14.3 hr



- **Key insight**: utilize **multi-view images** to better leverage the priors from CLIP.
- **Contribution**: an efficient training framework for aligning text and 3D, offering better generalization on unseen shapes and more flexible inputs.

2. Training Framework



Initialize shape encoder with CLIP

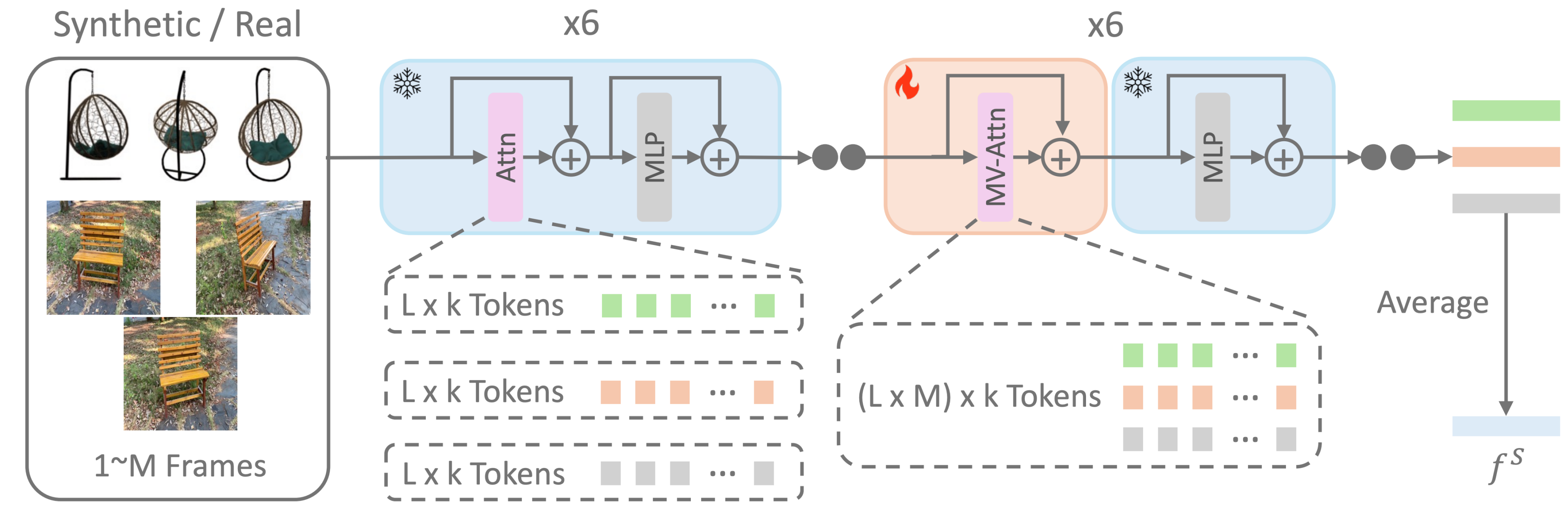
Contrastive loss to distill text and image knowledge from CLIP

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

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

3. Model Architecture

- ❄ Layers are frozen for training efficiency and preserve generalization.
- 🔥 Modified with trainable **multi-view attention** to learn 3D context.



Flexible encoding of arbitrary M views!

References

[1] Liu, M, et al. "Openshape: Scaling up 3d shape representation towards open-world understanding." NeurIPS 2023

[2] Zhou, J, et al. "Uni3d: Exploring unified 3d representation at scale." *ICLR 2024*

[3] Radford, A, et al. "Learning transferable visual models from natural language supervision." PmLR, 2021.

Acknowledgement

This work was funded by a CIFAR AI Chair, an NSERC Discovery grant, and a CFI/BCKDF JELF grant.

4. Ablation

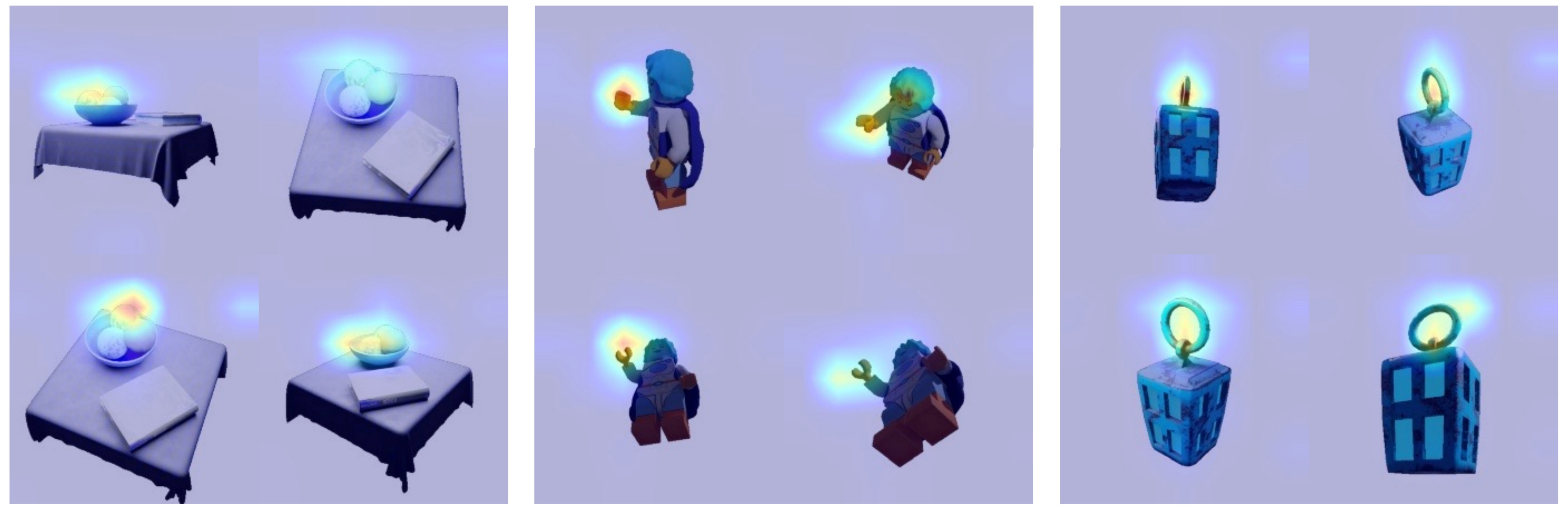
- **Takeaway**: frozen layers prevent overfitting, while MVA provides better generalization.

Table 5: Ablation on the number of layers for accuracy on Objaverse-LVIS using 12 input frames. Default model highlighted in gray.

Method	Top 1	Top 3	Top 5
3 layers	53.77	75.8	82.41
6 layers	55.24	77.08	83.43
12 layers (full)	**55.32**	**77.08**	**83.49**

Table 6: Multi-view attention (MVA) ablation of accuracy on Objaverse-LVIS (O-LVIS), MVPNet and ScanObjectNN with 12 frames. Default model highlighted in gray.

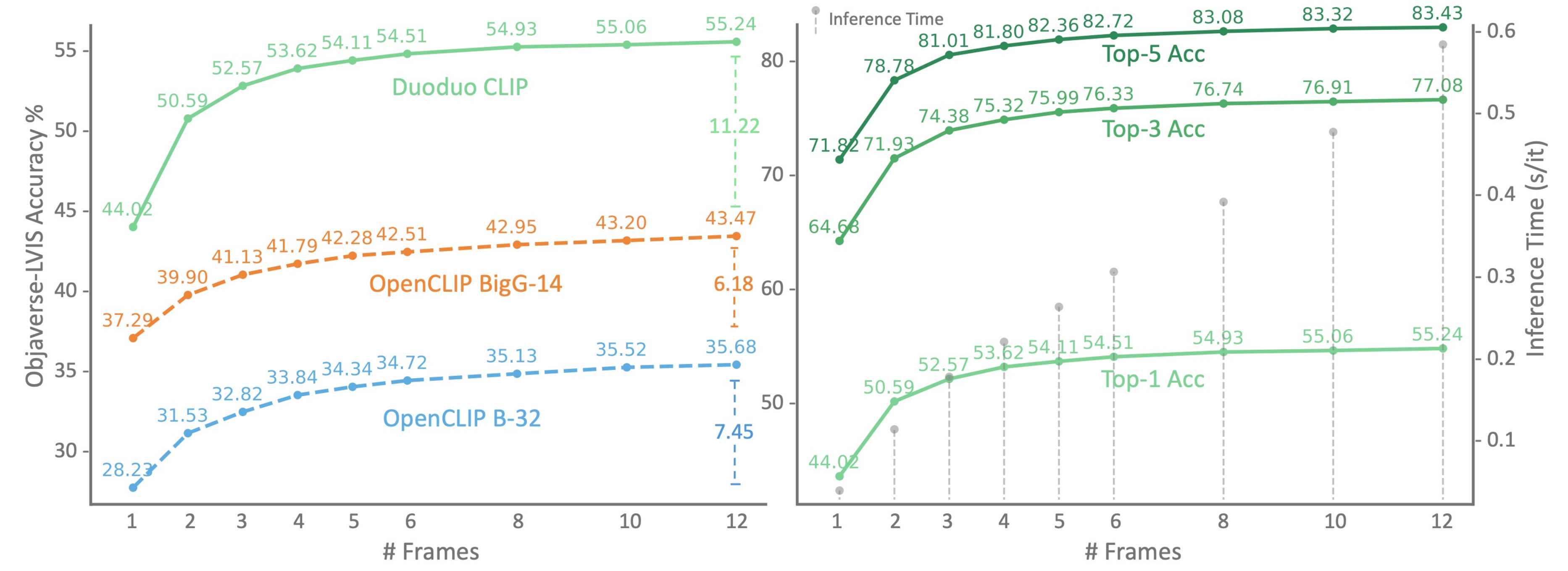
Method	Layers	O-LVIS	MVPNet	ScanObjectNN
-MVA	6	54.61	47.87	58.77
+MVA	6	55.24	**49.16**	**66.32**
-MVA	12	55.02	43.75	56.83
+MVA	12	**55.32**	44.42	64.15



MVA layers learn 3D correspondences.

5. Synthetic Dataset Results

- **Dataset**: ensemble of 4 synthetic datasets (874k shapes).
- **Evaluation**: Objaverse LVIS (46k) with 1156 categories.



Better performance scaling with more views

Pretrain Dataset			Ensembled (no LVIS)			Ensembled					
			O-LVIS			O-LVIS			ScanObjectNN		
Method	Rep	Enc	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
ZS B-32 (12F)	MV	Avg	35.7	54.8	62.1	35.7	54.8	62.1	53.9	73.5	81.2
ZS BigG-14 (12F)	MV	Avg	43.5	64.2	71.3	43.5	64.2	71.3	56.7	78.2	85.8
FT B-32 (12F)	MV	Avg	50.1	72.0	79.2	53.0	74.7	81.4	55.1	75.6	83.9
OpenShape (Liu et al., 2023a)	PC	PointBERT	39.1	60.8	68.9	46.8	69.1	77.0	52.2	79.7	88.7
TAMM (Zhang et al., 2024)	PC	PointBERT	42.0	63.6	71.7	50.7	73.2	80.6	55.7	80.7	88.9
MixCon3D Gao et al. (2024)	PC + MV	PointBERT	47.5	69.0	76.2	52.5	74.5	81.2	58.6	80.3	89.2
Uni3D (Zhou et al., 2024)	PC	3D VIT	47.2	68.8	76.1	**55.3**	76.7	82.9	65.3	85.5	**92.7**
ShapeLLM (Qi et al., 2024)	PC	RECON++	–	–	–	53.7	75.8	82.0	65.4	84.1	89.7
VIT-LENS (Lei et al., 2024)	PC	VIT-LENS$_G$	50.1	71.3	78.1	52.0	73.3	79.9	60.1	81.0	90.3
Duoduo CLIP (5F)	MV	MVA	51.3	73.1	79.9	54.1	76.0	82.4	60.7	82.4	88.5
Duoduo CLIP (12F)	MV	MVA	**52.7**	**74.5**	**81.3**	55.2	**77.1**	**83.4**	**66.3**	**85.5**	90.2

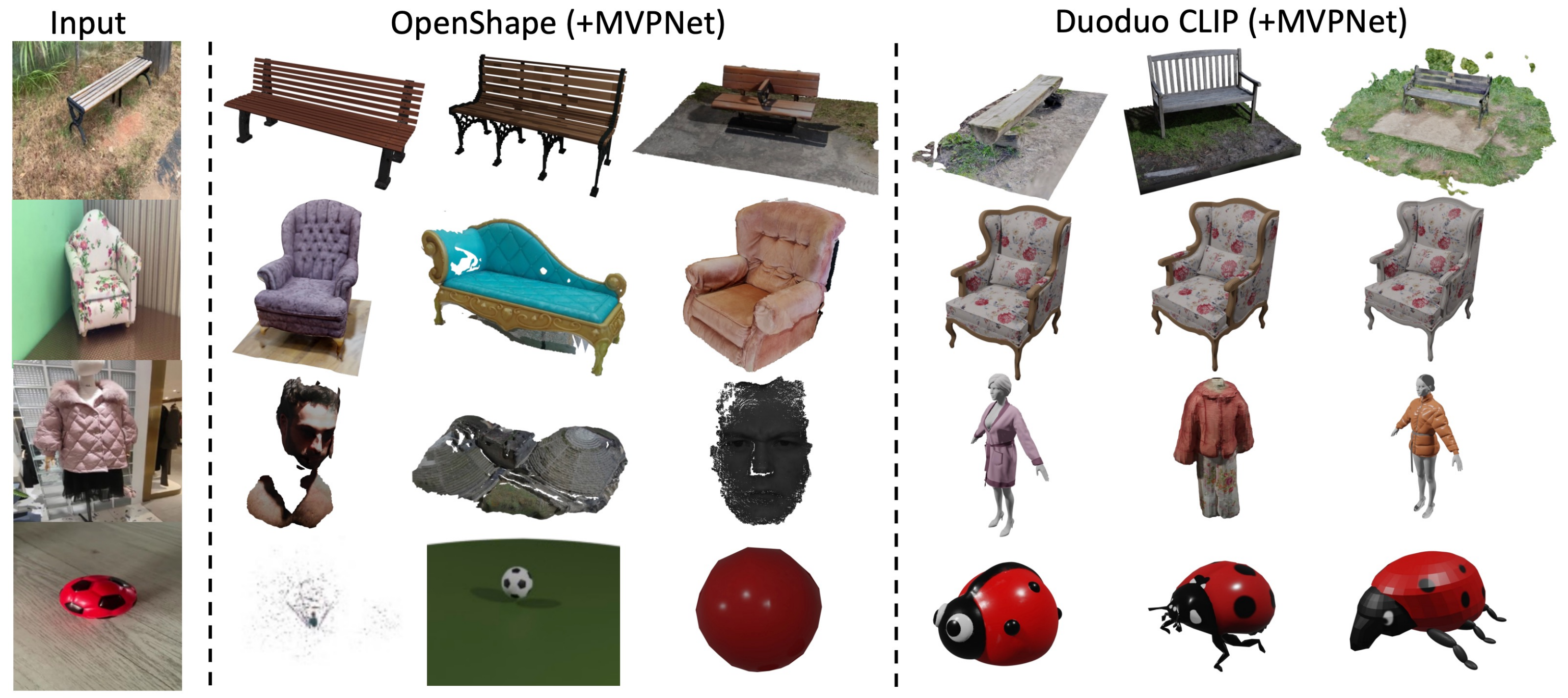
5 views match most methods; 12 views achieves SOTA

6. Real Dataset Results

- **Dataset**: MVImgNet (220k) multi-view images of real objects.
- **Evaluation**: MVPNet (87k) with 180 classes and point clouds.
- **Takeaway**: strong performance with just 1 view, and scales to data where point cloud isn't available.

Table 3: MVPNet classification comparison.

Method	Top 1	Top 3	Top 5
ZeroShot B-32 (12F)	52.68	70.99	77.22
FT B-32 (12F)	44.43	63.12	70.24
OpenShape†	10.80	19.62	25.20
OpenShape† (+MVPNet)	54.59	72.66	78.61
Ours (12F)	49.16	66.96	74.12
Ours (+MVPNet) (1F)	59.23	76.12	81.74
Ours (+MVPNet) (12F)	64.44	81.11	85.97
Ours (+MVImgNet) (12F)	**66.06**	**82.72**	**87.21**



7. Applications