A. Additional Experiments

A.1. Generalization to Non-Q-Former Architectures

To validate that our alignment regularization is architecture-agnostic and not limited to Q-Former-based models, we apply our method to 3D-LLaVA [5], which uses Omni Superpoint Transformer (OST) with MLP instead of Q-Former, on the Scan2Cap [2] dataset for scene-level 3D dense captioning. As shown in Table 9, our method yields consistent improvements across all four metrics, with CIDEr@0.5 improving by 2.7, demonstrating effectiveness beyond object-level tasks and Q-Former architectures.

Table 9. Scene-level 3D dense captioning on Scan2Cap [2]. Our alignment regularization is applied to 3D-LLaVA [5], which does not use Q-Former. (* denotes reproduced results.)

Method	C@0.5	B-4@0.5	M@0.5	R@0.5
3D-LLaVA*	76.1	36.3	27.0	57.3
+ Ours	**78.8**	**37.1**	**27.2**	**57.6**

A.2. Generalization to Different LLM Backbones

To verify that the middle-layer alignment strategy generalizes across LLM architectures, we conduct layer selection experiments on Phi-3 (32 layers). As shown in Table 10, middle layers (12–20) generally outperform both early and late layers. This aligns with our Phi-2 ablation (Table 5) where Layer 16 achieves optimal results, demonstrating that the strategy is robust across different LLM backbones.

Table 10. Layer selection ablation on Phi-3 backbone. Middle layers generally achieve better performance, aligning with the Phi-2 results in Table 5.

Layer	ModelNet40		Objaverse		Average
	(I)	(C)	(I)	(C)	
MiniGPT-3D	62.50	**56.50**	57.13	48.03	56.04
8	65.00	**56.50**	57.18	**49.59**	57.07
12	**66.00**	56.00	**58.74**	49.35	**57.52**
16	65.50	56.00	58.69	48.66	57.21
20	**66.00**	56.00	58.51	49.27	57.45
24	65.00	55.00	57.82	49.31	56.78
28	63.50	54.00	58.10	48.95	56.14