Promptable 3-D Object Localization with Latent Diffusion Models

Cheng-Yao Hong

Li-Heng Wang

Tyng-Luh Liu

Institute of Information Science, Academia Sinica

NeurIPS 2025

Outline

1 Introduction

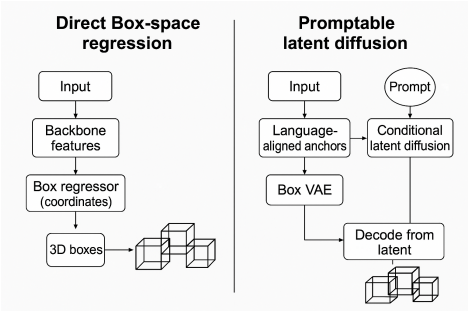
2 Method

3 Experiments

4 Conclusion

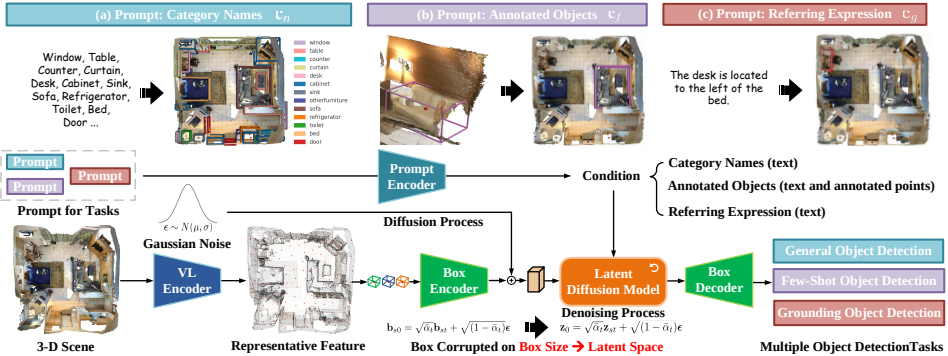
Motivation

- 3-D scenes are cluttered; direct coordinate regression is brittle.
- Many diffusion detectors still operate *in box space* $\Rightarrow$ limited flexibility.
- Goal: a **single** framework adaptable via **prompts** (class text, support examples, natural language).



Core idea

- **Box VAE** → latent box codes
- **Conditional latent diffusion** (prompt-guided)
- **Decode** latent → 3D boxes (DDIM)



Encode → diffuse (prompt) → decode for controllable 3-D boxes.

Outline

1 Introduction

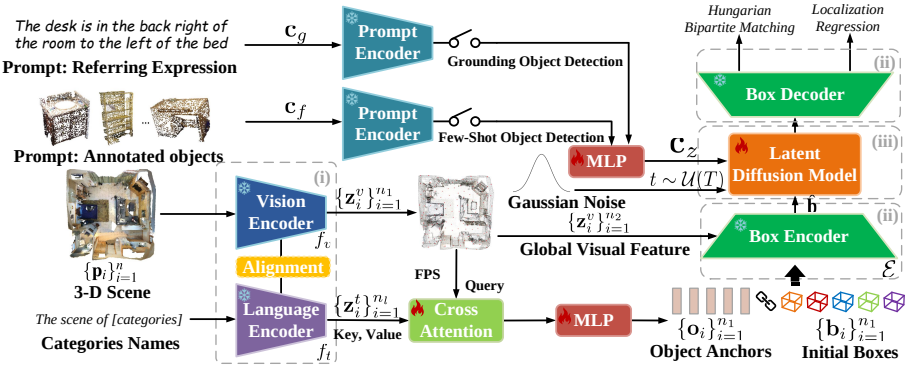
2 Method

3 Experiments

4 Conclusion

Overview of the model architecture

- **Language-aligned anchors**: cross-modal features seed semantically meaningful proposals.
- **Box VAE**: fuse anchors + scene features + coarse seeds ⇒ latent box codes.
- **Conditional diffusion**: classifier-free guidance; decoder → boxes; matching/NMS finalize outputs.



Training objective: Latent modeling

Latent diffusion (prompt-conditioned)

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\, x_0,\, \epsilon}\Big[\, \| \,\epsilon - \epsilon_\theta(x_t, t, c)\, \|_2^2 \,\Big]$$

Classifier-Free Guidance (used at sampling)

$$\hat{\epsilon}_\theta(x_t \,|\, c) \;=\; (1 + w)\, \epsilon_\theta(x_t, t, c) \;-\; w\, \epsilon_\theta(x_t, t, \varnothing)$$

Notes: diffusion in **latent box space**; prompts c (class names, support exemplars, text) condition the denoiser. VAE is used to obtain latent boxes; it is *pretrained/frozen* at first stage (no explicit VAE loss term reported here).

Training objective: Detection & Total

Detection loss (DETR-style with Hungarian matching)

$$\mathcal{L}_{\text{det}} = \sum_{(i,j)\in\pi} \Big(\lambda_{L1} \, \| b_i - \hat{b}_j \|_1 + \lambda_{\text{IoU}} \, (1 - \text{IoU}(b_i, \hat{b}_j)) + \lambda_{\text{cls}} \, \ell_{\text{cls}}(y_i, \hat{y}_j) \Big)$$

Total loss

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}} \, \mathcal{L}_{\text{diff}} + \lambda_{\text{det}} \, \mathcal{L}_{\text{det}}$$

π: Hungarian bipartite matching. ℓ_{cls}: focal / asymmetric classification loss as used in the paper. $\lambda_{\text{diff}}, \lambda_{\text{det}}$: dynamically scheduled.

Outline

1 Introduction

2 Method

3 Experiments

4 Conclusion

Promptable tasks

- **General 3-D detection**: optional class-name prompts sharpen semantics.
- **Few-shot detection**: condition on **support examples** + class text.
- **3-D grounding**: condition on **natural-language** descriptions; single/multi-target.

Experiment: General object detection

Method	**SUN RGB-D**		**ScanNetV2**	
	mAP@25	mAP@50	mAP@25	mAP@50
VoteNet	57.9	29.3	57.8	36.0
3DETR	59.1	32.7	65.0	47.0
Group-Free	63.0	45.2	69.1	52.8
Uni3DETR	67.0	50.3	71.7	58.3
V-DETR	67.5	50.4	77.4	65.0
Diffusion-based detector				
Diffusion-SS3D	–	–	64.1	43.2
Diff3DETR	–	–	65.7	44.9
CaTFree3D	–	–	/ 52.0†	–
Ours	**67.4**	**50.2** / 54.5†	**72.8**	**60.3**

† Training on 31 categories (including background) and testing on the other 7.

Experiment: Few-shot object detection (FS-SUNRGBD)

Method	1-shot		3-shot		5-shot	
	mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50
VoteNet	5.46	0.22	13.73	2.20	22.99	5.90
GeneralizedFS3D	6.81	1.58	17.52	4.69	22.84	6.76
PointContrast-VoteNet	7.03	1.17	20.32	4.19	21.03	6.71
Fractal-VoteNet	7.54	1.39	21.08	4.25	22.01	6.77
Meta-Det3D	6.77	0.73	15.37	2.99	24.22	5.68
Prototypical-VoteNet	12.39	1.52	21.51	6.13	29.95	8.16
Prototypical-VAE	14.36	2.42	27.70	8.73	33.21	13.98
Ours	**20.69**	**6.52**	**34.72**	**13.52**	**40.52**	**20.25**

Experiment: Few-shot object detection (FS-ScanNet, Split-1)

Method	1-shot		3-shot		5-shot	
	mAP@25	mAP@50	mAP@25	mAP@50	mAP@25	mAP@50
VoteNet	11.72	8.02	21.13	9.57	28.63	15.69
GeneralizedFS3D	12.03	8.19	24.90	10.26	29.29	16.67
PointContrast-VoteNet	12.59	8.52	20.12	11.16	25.83	15.49
Fractal-VoteNet	11.81	7.57	21.38	10.11	24.66	14.73
Meta-Det3D	10.28	4.03	23.42	10.64	25.65	13.88
Prototypical-VoteNet	15.34	8.25	31.25	16.01	32.25	19.52
Prototypical-VAE	16.00	10.22	31.60	19.37	32.84	22.39
Ours	**20.34**	**13.64**	**36.75**	**24.42**	**37.45**	**26.54**

Experiment: Grounding object detection

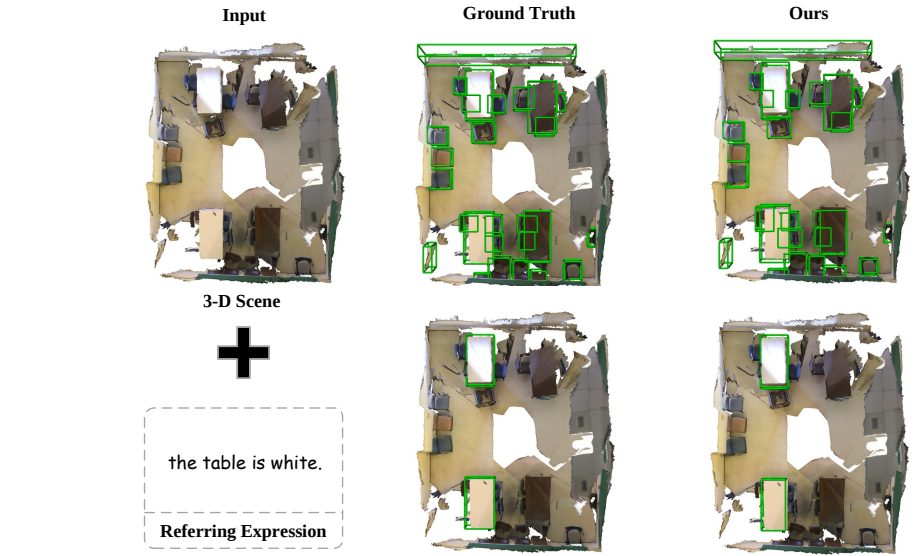
Method	ScanRefer		Multi3DRefer		ViGiL3D	
	Acc@25	Acc@50	F1@25	F1@50	Acc@25	Acc@50
ScanRefer	37.3	24.3	–	–	–	–
Multi3DRefer	51.9	44.7	42.8	38.4	–	–
ConcreteNet	46.5	46.5	–	–	–	–
D-LISA	–	46.9	–	51.2	–	–
Chat-Scene	55.5	50.2	57.1	52.4	11.0†	9.7†
PQ3D	57.0	51.2	–	50.1	10.8	10.8
Ours	**59.5**	**52.7**	**59.4**	**53.8**	**15.7**	**13.3**

† reproduced by our evaluation using released code.

Experiment: Zero-shot (OpenLex3D)

Method	Replica		ScanNet++		HM3D	
	Acc@25	Acc@50	F1@25	F1@50	Acc@25	Acc@50
OpenMask3D	21.5	15.1	9.8	4.2	8.2	5.3
ConceptGraphs	19.4	16.2	11.5	5.4	9.4	6.6
Ours	**19.5**	**17.9**	**11.3**	**5.4**	**9.9**	**7.6**

Qualitative results



Outline

1 Introduction

2 Method

3 Experiments

4 Conclusion

Conclusion

This work presents a unified, **promptable** framework for 3-D object localization that covers general detection, few-shot detection, and language grounding within a single model.

- We reformulate 3-D detection in **latent space**: language-aligned anchors and a Box VAE produce latent box codes, and **prompt-conditioned latent diffusion** refines them to final boxes.
- The same network seamlessly switches tasks by **changing prompts** (class names, support exemplars, free-form text) without retraining; few DDIM steps in latent space enable **efficient inference**.
- Extensive experiments on SUN RGB-D, ScanNetV2, *FS-SUNRGBD*/*FS-ScanNet*, and *ScanRefer*/*Multi3DRefer*/*ViGiL3D*/*OpenLex3D* show **strong mAP**, consistent **1/3/5-shot** gains, and competitive open-vocabulary grounding.

Thanks for your attention!