Lang4D: Weakly Supervised Learning of 4D Language Splatting

Mana Masuda*[1,2], Taiki Sekii*[1]
[1]CyberAgent, [2]Keio University

WINTICKET

CyberAgent AI Lab

Abstract

✓ **We propose a technique to compute the abstract meaning of a scene as language embeddings at arbitrary positions in a time-varying 3D space (i.e., 4D).**

Background

Application Retrieve prompt-specified object state changes directly within 4D media represented with 4D-GS.

Task Learn a 4D language field that computes language embeddings at arbitrary 4D positions.

Related Work

3D language field: Distill language embeddings into 3D from multi-view static images.

LangSplat [Qin et al., CVPR'24]

4D novel view synthesis: Render the appearance of time-varying 3D scenes from arbitrary camera viewpoints.

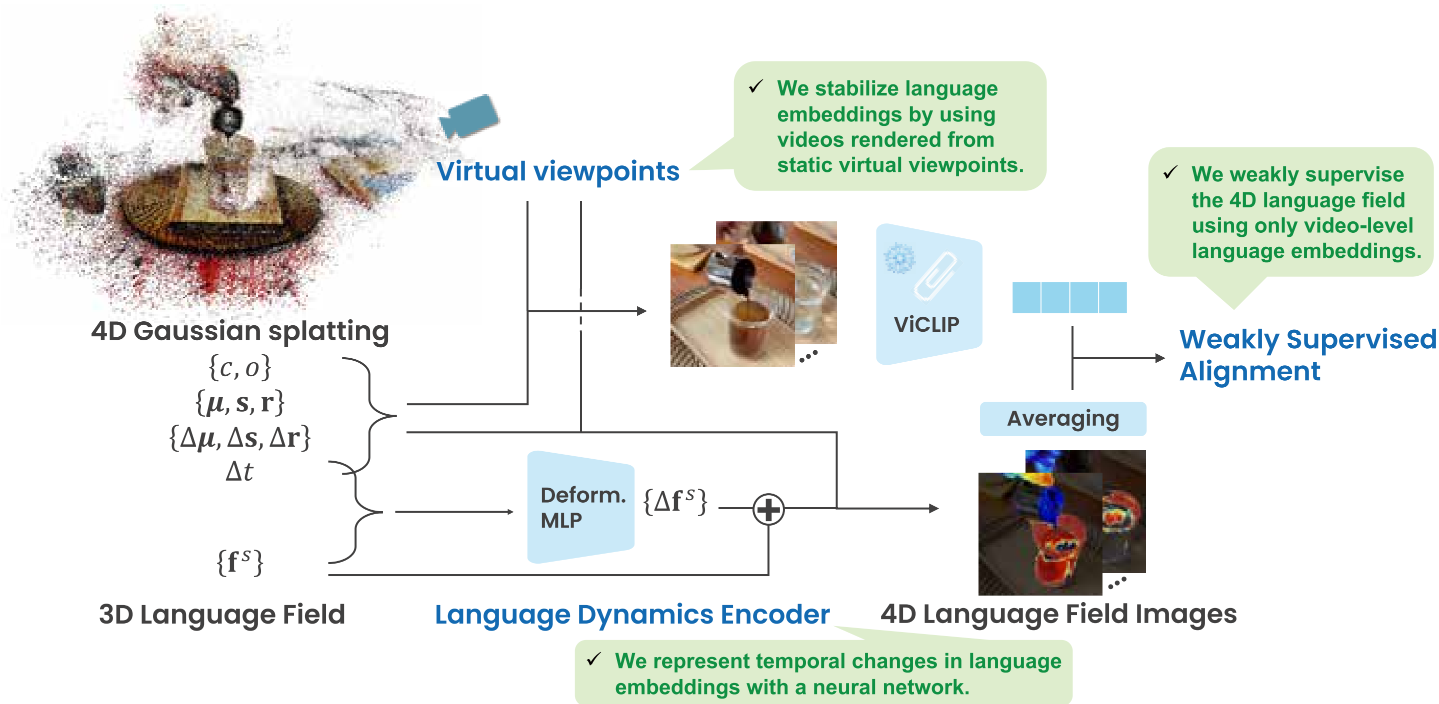
[Wu et al., CVPR'24]

Challenges

✓ **Establish a method for a 4D language field that models the temporal evolution of language embeddings beyond static 3D fields.**

✓ **Learn from video foundation models using only video-level language embeddings to supervise the 4D field.**

Proposed Method



Experimental Results

Method	mIoU
CLIP-based [17]	71.3
LangSplat-based [16]	16.6
GT-Based	89.4
Ours	80.5

Table 1: Accuracy comparison of 4D language grounding and segmentation.

Method	mIoU
Original	75.1
Virtual	80.5

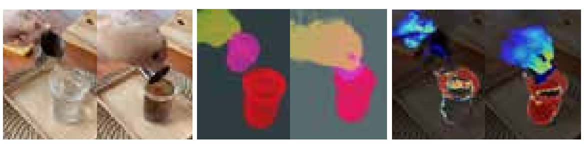
Table 2: Impact of different camera viewpoint sampling methods.

Method	mIoU
Ours w/o $\mathbf{f}^s$	4.9
Ours w/ $\mathbf{f}^s$	80.5

Table 3: Effect of dynamic language embedding design.

Query type	`Class`	`Temporal change`	`Long text`	`Multiple classes`
mIoU	74.1	78.2	79.5	78.5

Table 4: Ablation study on query sentence formulations. We employ four types of queries: describing the object class (`Class`), describing the object's change over time (`Temporal change`), providing a detailed object description (`Long text`), and describing a single object in the context of relationships among multiple objects (`Multiple classes`).



(a) Projected 4D Gaussians. (b) 3D language field. (c) 4D language field.