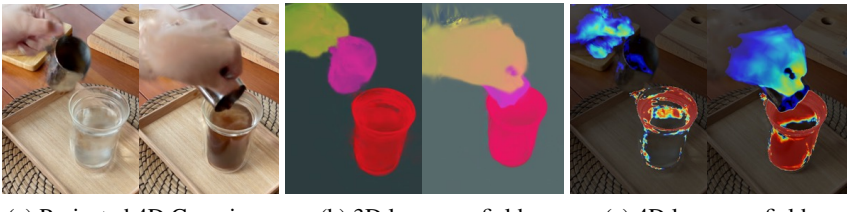
Lang4D: Weakly Supervised Learning of 4D Language Splatting

Mana Masuda[1,2,*,†]
mana.smile@keio.jp
Taiki Sekii[1,*,‡]
taiki.sekii@gmail.com

[1] CyberAgent
[2] Keio University

Abstract

Techniques for obtaining the abstract meaning of scenes as language embeddings at arbitrary positions in time-varying 3D spaces, *i.e.*, in 4D spaces, are now possible. These techniques are expected to enable applications such as content search and prediction in digital archives and digital twins. In this study, to acquire such a *4D language field*, we focus on two key challenges when applying existing methods designed for static scenes: (1) expanding the previous techniques to 4D space and (2) leveraging video-level language embeddings obtained from vision-language video foundation models. To address these challenges simultaneously, we propose *Lang4D*, a novel method that models the temporal changes of language embeddings in 3D space using video foundation models in a data-driven manner. Lang4D employs a video foundation model that inputs virtual viewpoint images, where recognition accuracy is stable, to obtain a common language embedding across all times and pixels for each video. Subsequently, it weakly supervises the learning of fluctuations in the language embeddings projected onto virtual viewpoints. In our experiments, we constructed a dataset specifically to evaluate the novel 4D language grounding and segmentation task. We verified the effectiveness of the proposed method in addressing these two challenges through quantitative evaluations and ablation studies.



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

Figure 1: Rendering results of the 4D language field estimated by the proposed method for novel viewpoints. We set a query to detect pouring coffee. The cosine similarity between the query and projected language embeddings varies over time in (c).

[*] Equal contribution.
[†] Work done during internship at CyberAgent, Inc.
[‡] Corresponding author.

1 Introduction

The technology of 4D reconstruction, which represents dynamically changing 3D scenes as a 4D medium, has garnered significant attention for its role in the geometrically precise recording and referencing of the real world as it evolves over time. Among existing approaches, *4D Gaussian splatting* [25, 27] is particularly promising for representing content such as digital archives and digital twins as 4D media, owing to its computational efficiency and high-fidelity reconstruction of real-world geometry. As such 4D media become more widespread, there is an increasing demand for techniques that enable retrieval and prediction directly within 4D space via open-ended natural-language instructions and queries from users—paralleling interaction paradigms already common for image and video media [1, 5]. Language models [12, 20, 23], now a mainstream technology in natural language processing, can capture semantic relationships and similarities between sentences and words through vector operations. Their language embeddings, produced from user-provided text, offer a quantitative representation of a user's abstract and complex intentions [3].

In the domain of language-based scene understanding, prior work in 3D language fields [8, 16] has distilled language embeddings into 3D space.[1] This enables the direct matching of user queries with individual objects for various tasks, such as 3D object detection and semantic segmentation. Notably, LERF [8] and LangSplat [16] utilize novel view synthesis techniques, such as NeRF [11] and 3D Gaussian splatting [7], to create language fields that project language embeddings onto arbitrary viewpoints, akin to radiance fields. In principle, a language field allows language embeddings to be computed at any point within 3D space. By comparing these embeddings with user queries, open-vocabulary 3D object detection becomes possible.

Motivated by these recent advancements, we address the challenge of modeling temporal changes in 3D language fields, thereby creating 4D language fields. Fig. 1 illustrates the 4D language field estimated by our proposed method. By directly representing a scene in 4D space, we can localize continuous and seamless temporal changes in objects within 3D space. This property has garnered significant attention. Recent studies have proposed novel view synthesis methods that model temporal changes in a scene's appearance within 4D space [13, 15, 19, 25, 27], while other methods can generate objects and motion in 4D space from user-input sentences [9]. Through data-driven modeling of temporal changes in 3D scenes, these studies have achieved more diverse and visually seamless high-quality representations than previous methods. Similarly, by extending the 3D language field to 4D, evolving object properties and attributes can be computed simultaneously within a single field, enabling the seamless analysis of 4D media in various applications.

In this study, we address two main challenges that emerge when extending conventional 3D language field methods [8, 16] to 4D space.

1. **Modeling the temporal evolution of language embeddings in 3D space.** Extending conventional 3D language field methods to 4D space is not straightforward, as they are based on deep neural networks (DNNs) and learning frameworks that assume 3D scenes composed of static objects. Their core components, such as NeRF [8, 11] and 3D Gaussian splatting [7, 16], are inherently designed for static scenes (*cf*. Tab. 1).

2. **Leveraging video-level language embeddings from vision-language video foundation models.** Conventional methods learn a 3D language field by applying the vision-

[1] We refer to learning a field from videos that can compute embeddings at arbitrary positions as *distillation*.

language foundation model CLIP [17] to every training image (or patch) and assigning the resulting language embeddings to individual pixels. Following the same rationale, it would be ideal to leverage state-of-the-art vision-language video foundation models trained on large-scale datasets. Unfortunately, these models are optimized with vast collections of video-sentence pairs and, at inference time, return just a single sparse language embedding[2] for the entire spatiotemporal volume. Consequently, they cannot readily provide a separate language embedding for each pixel.

Building on the above concepts, we propose Lang4D, a novel framework that learns a 4D language field using vision-language video foundation models to simultaneously address both challenges. The proposed method first visually models the appearance of scenes using 4D Gaussian splatting [25, 27]. To address the first challenge, we draw inspiration from 4D Gaussian splatting and introduce a new module named *Language Dynamics Encoder* (LDE) to model the temporal changes in language embeddings in a data-driven manner. In particular, the proposed method learns language embeddings in 3D space for each Gaussian at a reference time, which can then be projected onto arbitrary viewpoints. Subsequently, the LDE computes the variations in language embeddings from this reference time, efficiently distilling the language embeddings into 4D space.

To address the second challenge, we propose a weakly supervised learning task. Our solution uses a single-language embedding—common across all times and pixels of the input video and obtained from the video foundation model—to supervise the learning of language embeddings and their variations assigned to numerous Gaussians. Specifically, we exploit the property of language embeddings that allows new meanings to be represented through vector operations. This enables us to simultaneously align multiple embeddings in the language field with the language embeddings of the input video. Furthermore, to stabilize the estimation of language embeddings using a video foundation model, we propose a method that synthesizes and learns input videos from virtual static viewpoints from which more accurate language embeddings can be obtained.

In our experiments, we constructed a dataset to evaluate the novel 4D language grounding and segmentation task by annotating the descriptions and 2D masks of target objects in the publicly available HyperNeRF dataset [24]. Using this dataset, we performed quantitative evaluations and ablation studies to verify the effectiveness of the proposed method in addressing these two challenges.

In summary, our contributions are as follows: (i) We introduce the concept of a 4D language field to represent temporal changes in 3D language fields. (ii) We propose the LDE, which enables the modeling and learning of temporal changes in 3D language embeddings. (iii) We propose a solution for weakly supervised learning of a 4D language field. This approach uses a single language embedding, obtained from a vision-language video foundation model, which is shared across all times and pixels of virtual static viewpoint images.

For a detailed comparison between our method and prior work, please refer to the supplementary material. In particular, we explain how our approach differs from concurrent work by highlighting our second and third contributions.

[2] Although video foundation models extract multiple local features from the input video, these are not language embeddings that are fully temporally and spatially aligned. Therefore, it is necessary to aggregate the embeddings at the video level using methods such as feature pooling in the visual encoder.

2 Related Work

This section first focuses on 3D language field methods. We then provide an overview of the vision-language video foundation models that use 2D videos as inputs for our approach. For a comprehensive review of novel view synthesis in general, we refer the reader to the literature [4].

2.1 3D Language Fields

Following the developments in the novel view-rendering technologies, researchers have developed methods to distill scene image features into a 3D representation that can be projected onto arbitrary viewpoints [21]. These features are obtained from vision encoders trained on large datasets. Advancements in natural language processing have enabled the quantitative extraction of abstract meanings from user-input sentences via language embedding representations [3]. By leveraging these advancements, researchers have actively pursued methods that directly distill the distribution of language embeddings into 3D space, which can then be projected onto arbitrary viewpoints. Representative methods include LERF [8] and LangSplat [16], which represent 3D language fields using NeRF [11] and Gaussian splatting [7], respectively. These methods apply CLIP [17] to images captured from multiple viewpoints of a scene to obtain language embeddings at the image patch or pixel level. These embeddings are subsequently used to learn the language field. Inspired by LangSplat, our proposed method learns a 4D language field in which the meanings of the language embeddings change over time.

2.2 Vision-Language Video Foundation Models

With the advancements in natural language processing technologies discussed previously, language models [3, 12, 20, 23] have been introduced into the field of video understanding. This has led to significant progress in the development of vision-language video foundation models. In previous research [24, 28], sentences used for supervision were annotated to each video at various granularities, with sparse annotation of sentences at the video level used as a common approach. Alternatively, methods have been proposed for localizing query sentences temporally or spatiotemporally within videos by annotating sentences at specific times or providing image regions that correspond to sentences [10, 29] (*e.g.*, object bounding boxes or segmentation masks). However, because pretraining video foundation models requires large datasets, finer annotation granularity increases the human cost of dataset preparation.

In response, weakly supervised learning methods [22, 26] that perform the aforementioned localization using only sentences annotated at the video level have also been proposed. However, their localization accuracy is significantly inferior to that of supervised learning methods, making them unsuitable for learning language fields. Therefore, to reduce the annotation cost required for pretraining, we propose a weakly supervised learning method that directly uses video-level language embeddings from video foundation models trained with only video-level annotations.

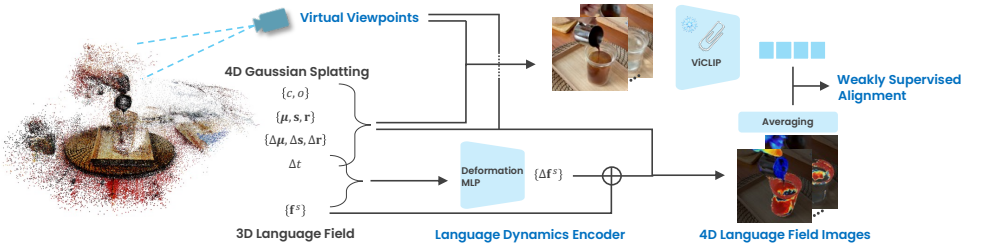


Figure 2: Overview of the proposed method. We introduce the novel LDE to estimate temporal changes in language embeddings. The language embeddings used for training are computed at the video level, rather than per pixel, using a vision-language foundation model.

3 Proposed Method

This section first introduces the 4D Gaussian splatting technique, which forms the basis of our method. Then, it describes our proposed 4D language field model and its weakly supervised learning approach.

3.1 4D Gaussian Splatting

We model a scene's appearance using 4D Gaussian splatting, which represents it as a set of time-varying anisotropic 3D Gaussians. Each 3D Gaussian is defined as

$$G(\mathbf{x},\Delta t) = \exp\left\{-\frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}})\right\}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ denotes the input 3D coordinates, and Δt represents the temporal displacement from the reference time. The mean $\hat{\boldsymbol{\mu}} \in \mathbb{R}^3$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ are computed as follows:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \Delta\boldsymbol{\mu}, \tag{2}$$

$$\hat{\boldsymbol{\Sigma}} = \hat{\mathbf{R}}\hat{\mathbf{S}}\hat{\mathbf{S}}^\top\hat{\mathbf{R}}^\top, \tag{3}$$

where $\Delta\boldsymbol{\mu}$ is the displacement of the mean due to temporal transitions. To enable differentiable optimization of the covariance matrix, $\hat{\boldsymbol{\Sigma}}$ is decomposed into a scaling matrix $\hat{\mathbf{S}}$ and rotation matrix $\hat{\mathbf{R}}$. Following prior work [27], $\hat{\mathbf{S}}$ and $\hat{\mathbf{R}}$ are calculated using scale factors $\mathbf{s} \in \mathbb{R}^3$ and rotation factors $\mathbf{r} \in \mathbb{R}^4$ at the reference time, along with their temporal displacements $\Delta\mathbf{s}$ and $\Delta\mathbf{r}$.

To optimize the parameters of these time-varying 3D Gaussians using the input videos, we project them onto images at each time step and efficiently render them using a tile-based rasterizer [16], as follows:

$$C(p,\Delta t) = \sum_{i\in\mathcal{N}} c_i \hat{\alpha}_i \prod_{j=1}^{i-1}(1-\hat{\alpha}_j), \tag{4}$$

where c_i is the color of the i-th Gaussian, $\mathcal{N}$ denotes the set of Gaussians within a tile, and $C(p,\Delta t)$ is the rendered color at pixel $p \in \{1,\ldots,H\}\times\{1,\ldots,W\}$ after temporal displacement Δt from the reference time. The term $\hat{\alpha}_i$ is computed as

$$\hat{\alpha}_i = o_i G_i^{2\mathrm{D}}(p,\Delta t), \tag{5}$$

where o_i is the opacity of the i-th Gaussian, and $G_i^{2D}(\cdot,\cdot)$ is the projected function on the image plane.

The parameters $\{\boldsymbol{\mu}, \mathbf{s}, \mathbf{r}, c, o\}$ are used to model the scene's appearance at the reference time and can be optimized using conventional 3D Gaussian splatting methods [7]. In addition, previous 4D Gaussian splatting research learned the parameters $\{\Delta\boldsymbol{\mu}, \Delta\mathbf{s}, \Delta\mathbf{r}\}$ to model temporal changes in a scene's appearance [25, 27].

3.2 4D Language Field

An overview of the proposed method is presented in Fig. 2. In our proposed 4D language field, we assign two types of language embeddings to each Gaussian: a static embedding $\mathbf{f}^s \in \mathbb{R}^D$ that captures time-invariant semantics such as object classes, and a dynamic embedding $\mathbf{f}^d \in \mathbb{R}^D$ that encapsulates semantics that change over time. The dynamic language embedding $\mathbf{f}^d$ is defined as

$$\mathbf{f}^d = \mathbf{f}^s + \Delta\mathbf{f}^s, \tag{6}$$

where we model $\Delta\mathbf{f}^s$ in a data-driven manner using the LDE, denoted as $\boldsymbol{\Phi}(\mathbf{f}^s, \Delta t)$. This model takes $\mathbf{f}^s$ and Δt as inputs, following the concept of residual blocks [6], to facilitate learning:

$$\mathbf{f}^d = \mathbf{f}^s + \boldsymbol{\Phi}(\mathbf{f}^s, \Delta t). \tag{7}$$

Using a tile-based rasterizer, we render $\mathbf{f}^s$ and $\mathbf{f}^d$ onto pixel p as follows:

$$\boldsymbol{F}^l(p, \Delta t) = \sum_{i \in \mathcal{N}} \mathbf{f}_i^l \hat{\alpha}_i \prod_{j=1}^{i-1} (1 - \hat{\alpha}_j), \quad l \in \{s, d\}, \tag{8}$$

where $\hat{\alpha}_i$ is calculated using Eq. (5).

We employ the method proposed by Yang *et al.* [27] to learn the 4D Gaussian splatting parameters $\{\boldsymbol{\mu}, \mathbf{s}, \mathbf{r}, c, o\}$ and $\{\Delta\boldsymbol{\mu}, \Delta\mathbf{s}, \Delta\mathbf{r}\}$, excluding the language embeddings. Subsequently, we sequentially optimize the language embeddings $\mathbf{f}^s$ and $\boldsymbol{\Phi}(\cdot,\cdot)$. By applying a vision-language foundation model (detailed in the next section) to the input images, we obtain D-dimensional per-pixel language embeddings $\left\{\boldsymbol{L}^l_{p,\Delta t} | l \in \{s, d\}\right\}$. Using these embeddings, we compute the loss $\mathcal{L}$ as follows:

$$\mathcal{L}^l = \frac{1}{THW} \sum_{p,\Delta t} M\left(\boldsymbol{L}^l_{p,\Delta t}, \boldsymbol{F}^l(p, \Delta t)\right), \quad l \in \{s, d\}, \tag{9}$$

where $M(\cdot,\cdot)$ denotes a distance function, defined in this study as the weighted sum of the L1 norm and cosine distance.

Weakly Supervised Learning Using Vision-Language Video Foundation Models. As discussed in the previous section, the fidelity of the language embeddings $\mathbf{f}^s$ and $\mathbf{f}^d$ that represent the 4D language field depends on the accuracy of the per-pixel language embeddings $\boldsymbol{L}^s_{p,\Delta t}$ and $\boldsymbol{L}^d_{p,\Delta t}$ extracted from the input video. The embedding $\boldsymbol{L}^s_{p,\Delta t}$ captures time-invariant semantics such as object classes. If $\boldsymbol{L}^d_{p,\Delta t}$ does not encode spatiotemporal features, it can be computed using vision-language foundation models designed for static images, as in previous work [17]. However, if $\boldsymbol{L}^d_{p,\Delta t}$ is intended to represent meanings that consider spatiotemporal features, it is preferable to compute it using video foundation models that can generate high-precision per-pixel language embeddings. Training such models typically requires

dense annotations of sentences throughout space and time, which is resource intensive. In contrast, recent state-of-the-art video foundation models are trained using video-level sentences that can be sparsely annotated at a relatively low human cost. This enables them to generate sentences or language embeddings for input videos.

Therefore, we perform weakly supervised learning of dynamic language embedding $\mathbf{f}^d$ by leveraging video-level embeddings obtained from pretrained video foundation models with low annotation costs. Considering that vector operations between language embeddings can represent new meanings, we reformulate the dynamic language embedding loss $\mathcal{L}^d$ in Eq. (9) as follows:

$$\mathcal{L}^d = M\left(\mathbf{l}^d, \frac{1}{THW}\sum_{p,\Delta t}\boldsymbol{F}^d(p,\Delta t)\right), \tag{10}$$

where $\mathbf{l}^d$ denotes the video-level language embedding obtained by applying the video foundation model to the input video. This formulation implies that we train the LDE under the assumption that the average of the projected dynamic language embeddings from the 4D language field matches the video-level embedding provided by the video foundation model.

Learning the 4D Language Field Using Virtual Static Viewpoint Videos. To accurately learn 4D Gaussian splatting, it is beneficial to capture the target object from diverse camera positions with varying parallaxes, thereby enhancing the precision of the camera pose estimation. For example, the HyperNeRF dataset involves moving a camera in a circular path around an object. However, the distillation accuracy of a 4D language field depends on the language embeddings estimated using a video foundation model. Therefore, the input video for the foundation model should capture the target object from a viewpoint that allows for accurate language embedding estimation.

Temporal changes in the camera position and orientation, such as those in the HyperNeRF dataset, can hinder a video foundation model from accurately recognizing the dynamic semantics of a target object, such as actions. To mitigate this, we use a video rendered from the 4D Gaussian splatting model at a static virtual viewpoint as input to the video foundation model, ensuring that its recognition accuracy is not compromised. In our experiments, we randomly selected a static viewpoint from the camera perspectives of the input video.

4 Experimental Results

Datasets. We constructed a dataset to evaluate the 4D language grounding and segmentation task based on the HyperNeRF dataset. The HyperNeRF dataset comprises videos captured in real-world environments, featuring objects whose motions and shapes change over time. From this dataset, we selected six scenes in which the types and meanings of the object movements or shape changes switched at certain time points: `americano`, `espresso`, `split-cookie`, `chickchicken`, `cut-lemon`, and `torchocolate`. These scenes are suitable for verifying the temporal changes in language embeddings in the 4D language field. Each scene comprises a 15 FPS video of approximately 30–60 s, and the camera poses for every frame were calibrated using COLMAP [18].

We quantitatively evaluated the 4D language grounding and segmentation task by projecting the 4D language field onto the evaluation camera viewpoints and comparing the estimated 2D masks of the target objects with the ground truth (GT). We generated 2D segmentation masks for the target objects by applying object segmentation methods [2] to the

selected scene videos. Each mask was manually corrected and its frames were annotated with sentences describing the target object. These sentences were created by visually identifying the interval from the moment an object's state begins to change until it transitions to the next state. As described in Tab. 4, we investigated how different ways of phrasing these annotations affect the distillation accuracy of the proposed 4D language field.

Evaluation Protocol. As an evaluation protocol for the new 4D language grounding and segmentation task, we projected the 4D language field onto images from evaluation camera viewpoints. By thresholding the cosine similarity between the projected embeddings and language embedding of the query, we estimated the 2D segmentation masks corresponding to the query. Following previous studies [16], we compared the distillation accuracy of the 4D language field of each method using the mean intersection over union (mIoU) between the estimated masks and GT. For each scene, frames were alternately selected in a 1:1 ratio for the training and evaluation of the 4D language field.

As the implementation of a weakly supervised learning method for the 4D language field is not publicly available, we implemented three variants. In selecting the baselines, we aimed to train the 4D language field by combining all plausible techniques from prior work.

- **CLIP-based method**: For each frame, we estimate per-pixel language embeddings $\boldsymbol{L}^d_{p,\Delta t}$ (Sec. 3.2) using CLIP and distill them directly in 4D Gaussian splatting. Specifically, we predict a CLIP embedding for the object mask in each frame, assign this embedding to every pixel inside the mask to obtain $\boldsymbol{L}^d_{p,\Delta t}$, and then minimize its distance to the dynamic feature $\mathbf{f}^d$.

- **LangSplat-based method**: Analogous to LangSplat, this variant distills a constructed GT embedding, $\bar{\boldsymbol{L}}^d_{p,\Delta t}$ (described below), with the static language embeddings $\mathbf{f}^s$. We generate $\bar{\boldsymbol{L}}^d_{p,\Delta t}$ from the ground-truth masks and captions and minimize its distance to $\mathbf{f}^s$.

- **GT-based method**: We directly model the GT embeddings $\bar{\boldsymbol{L}}^d_{p,\Delta t}$, obtained from the masks and captions, within 4D Gaussian splatting and optimize the loss in Eq. (9) using $\bar{\boldsymbol{L}}^d_{p,\Delta t}$.

For both the LangSplat-based and GT-based methods, we convert the annotated sentences into language embeddings using the ViCLIP text encoder. Subsequently, we assign these embeddings to all pixels inside the corresponding masks to obtain the per-pixel GT language embeddings $\bar{\boldsymbol{L}}^d_{p,\Delta t}$.

Implementation. The training of the 4D Gaussian splatting in our proposed method followed the approach of previous work. Training was conducted for approximately 100 epochs, resulting in each scene being composed of more than 1.0×10^5 Gaussians. We first trained the 4D Gaussian splatting, and then sequentially learned the static language embeddings $\mathbf{f}^s$ and dynamic language embeddings $\mathbf{f}^d$ for the 4D language field. At each stage, we fixed the previously learned parameters and did not update them in subsequent steps. We obtained the video-level embeddings $\mathbf{l}^d$ described in Eq. (10) using the video foundation model ViCLIP (`ViCLIP-L-14`) [24]. For the architecture of the LDE, we adopted the deformation MLP from previous studies [27] and set both $\mathbf{f}^s$ and $\mathbf{f}^d$ as 768-dimensional vectors. The other hyperparameters related to the LDE are listed in the supplementary material.

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`).

To minimize experimental uncertainties, we selected virtual static camera viewpoints using a simple method. At each iteration of video training, we randomly sampled a viewpoint from the camera viewpoints of the training videos. We performed data augmentation on the virtual viewpoint videos by creating sub-videos and randomly altering their scale and start/end times.

Comparative Experiments Against Baselines. Tab. 1 summarizes the mean accuracy of our method and the baseline approaches for the 4D language grounding and segmentation task. Our method attains substantially higher accuracy than baselines that do not rely on GT masks and sentences, and it closes the gap to the GT-based method to within a few percentage points. These findings indicate that the weakly supervised learning of the dynamic language embedding $\mathbf{f}^d$ successfully captures the temporal evolution of scene-level language embeddings.

Effect of Virtual Viewpoints in Input Videos. We investigated the impact of different camera viewpoint selection methods on the learning of the 4D language field. Tab. 2 presents the distillation accuracy when camera viewpoints were sampled using different patterns, evaluated on the same 4D language grounding and segmentation task as in Tab. 1. The compared methods included using the original video viewpoints (*Original*) and sampling all camera viewpoints from virtual static viewpoints (*Virtual*). The accuracy of the 4D language field learned using the original viewpoints is significantly lower than that of our proposed method, which uses virtual viewpoints. This demonstrates that incorporating virtual static viewpoints is effective for obtaining high-precision language embeddings from a video foundation model and learning an accurate 4D language field.

Design of Dynamic Language Embeddings. We examined how different designs of the dynamic language embeddings $\mathbf{f}^d$ affect the distillation accuracy of the 4D language field. As shown in Tab. 3 and Fig. 3, modeling $\mathbf{f}^d$ as variations relative to the static embeddings $\mathbf{f}^s$ (*Ours w/* $\mathbf{f}^s$) yields higher accuracy than modeling $\mathbf{f}^d$ directly without $\mathbf{f}^s$ (*Ours w/o* $\mathbf{f}^s$). This suggests that utilizing static embeddings to model dynamic embeddings is effective, likely because it better captures the inherent relationship between static and dynamic features.

Analysis of the 4D Language Field Representation. In the 4D language grounding and segmentation task, we analyzed the 4D language field representations learned by the proposed method by evaluating its accuracy while varying the query formulation. Tab. 4 displays the distillation accuracy of the 4D language field for different query descriptions obtained us-

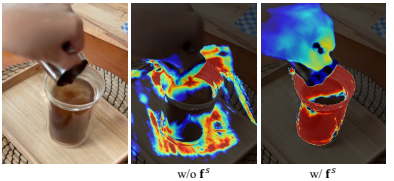


Figure 3: Visual results illustrating the impact of our dynamic embedding design. The proposed dynamic language embedding demonstrates improved distillation accuracy for the 4D language field.

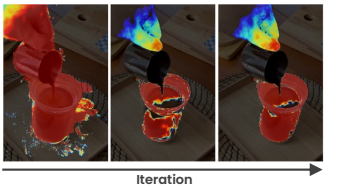


Figure 4: Projection results for novel viewpoints during the learning process of the 4D language field by the proposed method. The temporal accuracy of the projected language embeddings improved as training progressed.

ing our method. As indicated, the learned representation is robust, maintaining its distillation accuracy with only a few percentage points of fluctuation across different queries.

Qualitative Results. Fig. 1 shows the visualization results of the 4D language field learned by the proposed method. Our method successfully segmented the target object region, both temporally and spatially. Fig. 4 illustrates the learning progression of the 4D language field using the proposed method.

5 Conclusion

In this study, we aimed to construct a 4D language field in which the meanings of language embeddings change over time at arbitrary positions in 3D space. We focused on two challenges that arise when applying conventional 3D language field methods: (1) extending them to model language embeddings in 4D space and (2) leveraging video-level language embeddings obtained from vision-language video foundation models. To address these challenges simultaneously, we proposed Lang4D, which is a framework that learns a 4D language field based on 4D Gaussian splatting using video foundation models. Specifically, we introduced the LDE for the data-driven modeling of temporal changes in language embeddings in 3D space. We also proposed a weakly supervised learning method that uses input videos from virtual viewpoints, where the recognition accuracy is stable, to obtain video-level language embeddings common to all times and pixels. This approach enables us to assign language embeddings to numerous Gaussians and learn their variations. Through quantitative evaluations and ablation studies, we verified the effectiveness of our proposed method in addressing these two challenges.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3d question answering for spatial scene understanding. In *CVPR*, 2022.

[2] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023.

[3] Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.

[4] Anurag Dalal, Daniel Hagen, Kjell G. Robbersmyr, and Kristian Muri Knausgård. Gaussian splatting: 3d reconstruction and novel view synthesis: A review. *IEEE Access*, 12:96797–96820, 2024. doi: 10.1109/ACCESS.2024.3408318.

[5] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: Visual question answering in interactive environments. In *CVPR*, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[8] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language embedded radiance fields. In *ICCV*, 2023.

[9] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4D with dynamic 3d gaussians and composed diffusion models. In *CVPR*, 2024.

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024.

[11] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[12] OpenAI, Josh Achiam, Steven Adler, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[13] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021.

[14] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. volume 40. ACM, dec 2021.

[15] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020.

[16] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. LangSplat: 3d language gaussian splatting. In *CVPR*, 2024.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *CVPR*, 2021.

[18] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[19] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *ECCV*, 2022.

[20] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural Feature Fusion Fields: 3D distillation of self-supervised 2D image representations. In *3DV*, 2022.

[22] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Weakly-supervised spatio-temporal video grounding with variational cross-modal alignment. In *ECCV*, 2024.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[24] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2023.

[25] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024.

[26] Xiaobo Yang and Xiaojin Gong. Foundation model assisted weakly supervised semantic segmentation. In *WACV*, 2023.

[27] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024.

[28] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023.

[29] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020.