GRIZAL: Generative Prior-guided Zero-Shot Temporal Action Localization

Onkar Susladkar[1], Gayatri Deshmukh[2], Vandan Gorade[2], Sparsh Mittal[3]†,
[1]Yellow.ai, India, [2]Northwestern University, USA,
[3]IIT Roorkee, India. †Corresponding Author
onkarsus13@gmail.com,{gayatri.deshmukh,vandan.gorade}@northwestern.edu,sparsh.mittal@ece.iitr.ac.in

Abstract

Zero-shot temporal action localization (TAL) aims to temporally localize actions in videos without prior training examples. To address the challenges of TAL, we offer GRIZAL, a model that uses multimodal embeddings and dynamic motion cues to localize actions effectively. GRIZAL achieves sample diversity by using large-scale generative models such as GPT-4 for generating textual augmentations and DALL-E for generating image augmentations. Our model integrates vision-language embeddings with optical flow insights, optimized through a blend of supervised and self-supervised loss functions. On ActivityNet, Thumos14 and Charades-STA datasets, GRIZAL vastly outperforms state-of-the-art zero-shot TAL models, demonstrating its robustness and adaptability across a wide range of video content. The code and models are available on `https://github.com/CandleLabAI/GRIZAL-EMNLP2024`.

1 Introduction

Temporal action localization (TAL) seeks to accurately identify specific actions occurring within extensive, unedited videos. Its applications include real-time surveillance for security, improving sports training with in-depth reviews of player actions, and optimizing video content organization by enabling effective search and indexing capabilities. The deep-learning techniques have achieved a significant milestone for TAL, however, these techniques necessitate training on comprehensive datasets for optimal performance. Given the challenges of gathering exhaustive, annotated videos, some TAL models tend to misidentify actions not encountered during training. To address this challenge, zero-shot learning seeks to identify actions without prior exposure to labeled instances of those actions during training. These techniques leverage semantic linkages and incorporate pre-trained Visual-Language (ViL) models to recognize actions by comparing the semantic similarity between action descriptions and video content.

Recent methods like STALE (Nag et al., 2022) propose a parallel localization and classification architecture. UnLoc (Yan et al., 2023) introduces an end-to-end trainable one-stage approach, starting directly from a CLIP two-tower model. Existing self-supervised learning (Purushwalkam and Gupta, 2020; Huang et al., 2021; Rebuffi et al., 2021; Wang and Qi, 2022) literature emphasizes the significance of augmentations for achieving generalized representation through diversity. Both STALE and UnLoc use learned text encoders only to get the representation whereas GRIZAL uses multi-modal models like GAFNet. STALE and UnLoc do not use any generative models to generate new data to use as external augmentations. Hence, their performance is inferior to the methods that utilize external augmentations (Ju et al., 2023).

Prevailing TAL methods predominantly rely on either retrieval-augmented techniques (Yasunaga et al., 2022) or stochastic-augmented training approaches (Wang et al., 2021b; Jing et al., 2018; Lin et al., 2020). For example, (Xu et al., 2021) incorporates traditional training augmentation with a non-parametric retrieval component, while (Lin et al., 2020) applies transformations without explicit dependence on pre-existing samples. This limited sample diversity leads to 1) Overcomplete representation stemming from semantic inconsistency. This occurs when varied visual representations of the same action impede the model's generalization. 2) Undercomplete representation results from a lack of contextual understanding. Here, the meaning of an action varies based on the context, leading to different interpretations. As shown in Fig. 1, both GRIZAL (without GPT-4

Support for this work was provided by Science and Engineering Research Board (SERB) of India, under the project CRG/2022/003821

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19046–19059
November 12-16, 2024 ©2024 Association for Computational Linguistics

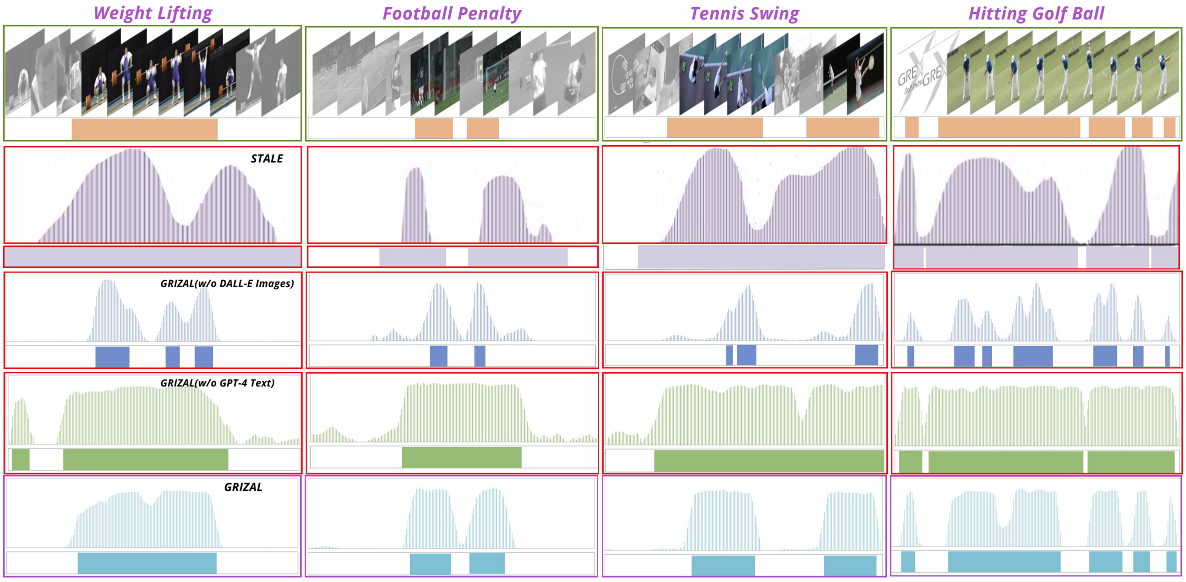


Figure 1: Row 1: Video frames and corresponding ground-truth action intervals. Row 2: STALE. Row 3 and Row : GRIZAL variants. Row5: GRIZAL. In rows 2 to 5, the y-axis shows action probability of each frame. While GRIZAL variants and STALE suffer from over/under-completeness, the full GRIZAL avoids these issues

text) and GRIZAL (without DALL-E image) suffer from these issues. Similarly, STALE suffers from over-complete representation. Some methods (Kalakonda et al., 2023; Ju et al., 2023) utilize generative models to generate text and images related to action, which act as external augmentations. These methods pass generated content through a pre-trained model to get rich feature representations and utilize them for specific tasks. However, since these extracted features are not passed through any additional learnable layers, the model does not get fine-tuned to the specific task at hand.

To address these challenges, we propose GRIZAL. We demonstrate that incorporating diverse and contextually rich augmentations in TAL results in more discriminative and controlled representations. As shown in Fig. 1, the full GRIZAL model precisely localizes temporal action boundaries. Clearly, diversity mitigates over-complete and under-complete representation issues. To achieve sample diversity, we leverage large-scale generative models such as GPT-4 for generating textual augmentations and DALL-E for generating image augmentations. These generated images and texts are passed through a pre-trained multimodal model to get a rich feature map. This feature map is passed through additional learnable layers to blend these features effectively. Our contributions are:

- We incorporate diverse and contextually rich augmentations in TAL to achieve more discriminative and controlled representations. We showcase this by using large vision-language models such as GPT-4 and DALL-E to generate textual and image augmentations, respectively.

- We introduce GRIZAL, a novel Generative Augmentation Guided Transformer-based architecture designed for zero-shot Temporal Action localization. This innovative approach incorporates generative augmentations to enhance the model's ability to handle diverse scenarios. By utilizing both textual and visual representations, GRIZAL leads to more controlled representations, avoiding under- or over-completeness.

- The experiments on ActivityNet-V1.3, THUMOS14, and Charades-STA datasets confirm that GRIZAL effectively localizes actions from both known and unknown classes and outperforms existing CLIP-based methods. For instance, compared to the SOTA method STALE, under open-set scenario (75-25%), GRIZAL improves mIOU by 5.2pp on the ActivityNet and 3.2pp on the THUMOS14 dataset (pp =percentage point).

2 Related Work

There have been numerous efforts at the intersection of computer vision and natural language processing. Radford et al. (Radford et al., 2021) introduced CLIP, a large-scale pretrained Vision-Language (ViL) model, trained using a contrastive

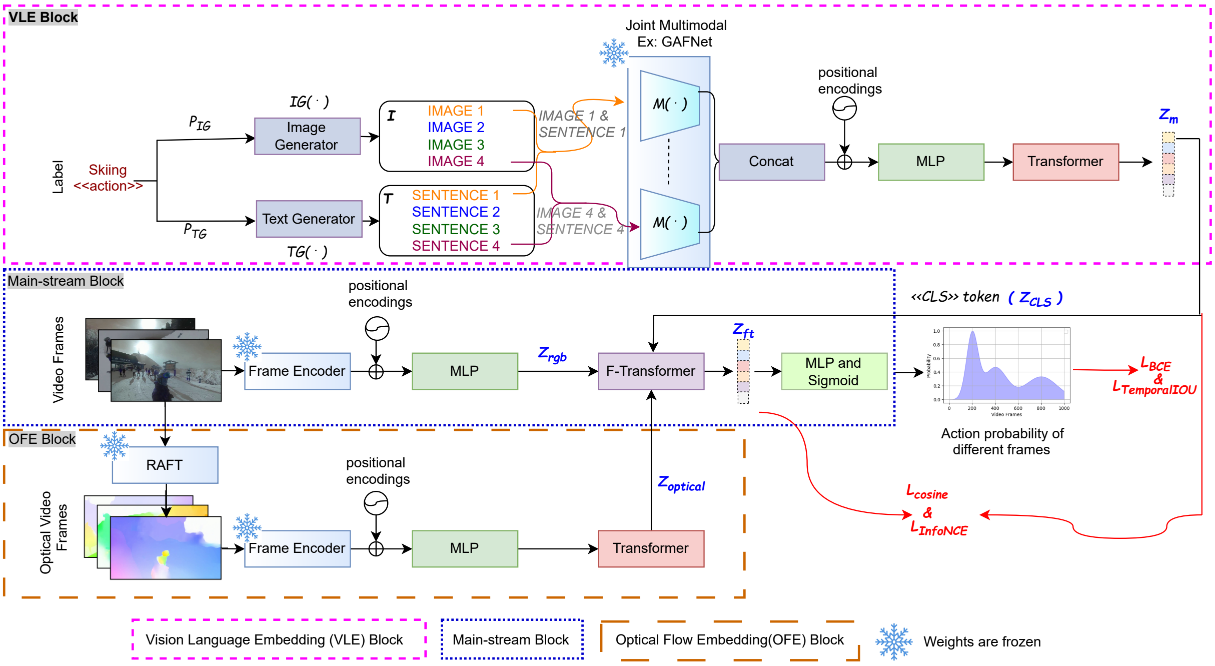


Figure 2: GRIZAL architecture

learning strategy on 400 million image-text pairs. CLIP demonstrated remarkable zero-shot transferability across 30 classification datasets. This motivated subsequent works to propose enhancements in training strategies, such as CoOp (Zhou et al., 2022) and CLIPAdapter (Gao et al., 2023). A similar approach has been explored for videos (Miech et al., 2020). ActionCLIP (Wang et al., 2021a) applies CLIP for action localization.

The existing supervised learning TAL networks are either two-stage (Lin et al., 2019; Shou et al., 2016) or single-stage networks (Zhang et al., 2022). EffPrompt (Ju et al., 2022) introduces a two-stage sequential architecture for zero-shot action localization. It involves generating an action proposal using a pre-trained detector like BMN (Lin et al., 2019), followed by the proposal classification using CLIP features. We aim to pioneer a proposal-free framework that leverages contextual augmentations and eliminates the reliance on a proposal generation stage.

Using generative models to augment training data has significantly enhanced model generalization. For example, (Bowles et al., 2018; Antoniou et al., 2017) have incorporated GAN-derived synthetic data into training sets. DALL-E (Ramesh et al., 2021) can create diverse images from textual prompts, while GPT-4 (OpenAI, 2023) excels in language understanding. Traditional CLIP-based models for action recognition often rely on manual augmentation or retrieval-based hard negatives, which can constrain representation quality. Our approach leverages generative models to create contextually relevant augmentations tailored to specific modalities.

3 GRIZAL: Our Proposed Method

Problem Formulation: Consider a dataset D, composed of two disjoint subsets: the training set D_{train} and the validation set D_{val}. Each subset has a collection of (V,AL,F), where: V is a video sequence, AL denotes the action label corresponding to V, and $F = \{f_1, f_2, \ldots, f_n\}$ represents binary annotations for each frame within V. The annotation for frame f_i is defined as $f_i = 1$ if the action specified by AL is present, and $f_i = 0$ otherwise.

3.1 GRIZAL Network Architecture

GRIZAL is a novel zero-shot TAL technique for understanding complex visual-textual relationships across diverse and novel contexts. GRIZAL offers a solution to bridge video content with textual and visual descriptions without the biases found in fully or weakly supervised methods. Figure 2 shows the architecture of GRIZAL. It consists of three blocks: VLE (Vision Language Embedding), OFE (Optical Flow Embedding), and mainstream. Given video frames (V) and action labels (AL) as input, GRIZAL pinpoints the frames where the specified action (AL) occurs. The VLE block enriches the mainstream block with contextual information, enhancing the model's understanding of the actions described in AL. The OFE block provides cross-attention to the mainstream, helping recognize action transitions within the video. GRIZAL also leverages a proposed F-Transformer block, which merges frequency and spatial domain features through Fourier Transform to enrich feature representation. The F-Transformer has been explained in the Appendix section.

VLE block: Given an action label AL (e.g., "skiing"), the VLE block uses it in prompts designed for image-generator IG (DALL-E in our case) and text-generator TG (GPT-4 in our case), denoted as P_{IG} and P_{TG}, respectively. The IG and TG models take these prompts as inputs and produce sets of images and sentences where $I = IG(P_{IG})$ yields a set of images $I = \{Img_1, \ldots, Img_k\}$, where $Img_i \in \mathbb{R}^{c \times h \times w}$ and $k = 4$ in our case. Similarly, $T = TG(P_{TG})$ produces a set of sentences $T = \{Sent_1, \ldots, Sent_k\}$, where each $Sent_i$ is a string of text of variable length. Here, AL acts as a clue that helps to create images and sentences about the action to be localized. Each image and sentence shows a different view of the action to give a complete understanding.

Next, the VLE block uses a Joint Multimodal unit M, for feature extraction and semantic coherence. The parameters of M are kept frozen. For each image-sentence pair (I_i, T_i), where i ranges from 1 to k, M extracts semantically coherent embeddings. It enhances action understanding by combining the vivid, instant representation of actions in images with the detailed, context-rich explanations provided by sentences.

Formally, let $E_i = M(I_i, T_i)$, the resulting embedding for pair i has a dimensionality of $\mathbb{R}^{B \times S \times E}$, with B representing the batch size, S the sequence length, and E the embedding dimension. The embeddings E_i for all pairs are concatenated along the sequence dimension to form a single tensor T, such that $T = Concat(E_1, E_2, E_3, E_4)$. The resulting tensor T has dimensions $\mathbb{R}^{B \times 4S \times E}$. Before the concatenation of embeddings, each embedding is padded to match the maximum sequence length. Positional encodings are then added to T followed by an MLP (multilayer perceptron) that aligns multimodal features and reduces the dimensionality of a tensor T to shape $\mathbb{R}^{B \times 4S \times 512}$. The tensor T is then fed into a transformer comprising N blocks where $N = 7$, to yield the final tensor Z_m of dimension $\mathbb{R}^{B \times 4S \times 512}$.

The vector resulting from feature concatenation has a higher dimensionality that preserves both modalities' dimensionality, context, and unique characteristics. Addition may lead to feature cancellation, especially if the vectors contain both positive and negative values. This can result in the loss of crucial information necessary for tasks such as action localization or multimodal understanding.

Through this process, the VLE analyzes the context of actions depicted in both images and sentences, generating a set of context-aware embeddings (Z_m) that encapsulate the AL. At last, the embedding corresponding to the «CLS» token is extracted, yielding a $\mathbb{R}^{B \times 512}$ shaped tensor (denoted as Z_{CLS}). This token embedding is forwarded to the F-transformer in the mainstream block.

OFE block: Given a set of RGB video frames, RAFT algorithm (Teed and Deng, 2020) is applied to compute optical flow frames. Let $V_{rgb} = \{v_{rgb1}, \ldots, v_{rgbm}\}$ denote the set of RGB frames, where $v_{rgbi} \in \mathbb{R}^{3 \times H \times W}$ and H, W are the height and width of the frames, respectively. RAFT transforms V into a set of optical flow frames $O = \{o_1, o_2, ..., o_m\}$, with $o_i \in \mathbb{R}^{3 \times H \times W}$. This produces a detailed pixel-by-pixel motion depiction across frames. O is then passed through a frame encoder (FE), which extracts feature vectors F_o. While we employ a CLIP image encoder (Radford et al., 2021) to obtain video features, any other frame encoder can also be used. These features F_o are then passed through an MLP. RAFT and the frame encoders are pre-trained, and their weights are frozen during training. After an MLP, the features F_o pass through a transformer with N blocks, capturing dynamic motion in optical flow to generate feature representations $Z_{optical}$, used in F-Transformer's cross attention (refer Section S.1).

Mainstream Block: It performs action localization by integrating the feature maps $Z_{optical}$ from the OFE block and Z_{CLS} from the VLE block. As outlined in the OFE block, RGB video frames V_{rgb} undergo a similar initial process, where V_{rgb} is fed into a frame encoder that produces features. These features are then aligned dimensionally through an MLP, yielding a transformed feature set Z_{rgb}. The F-Transformer block fuses Z_{rgb}, $Z_{optical}$, and Z_{CLS} to generate the feature map Z_{ft}, which contains enriched information for action localization. At last, Z_{ft} goes through an MLP and then a sigmoid function to obtain a probability distribution across video frames, indicating the likelihood of each frame containing the target action.

3.2 Learning Objective

Biases can form when models learn only from specific examples (i.e., fully supervised learning) or when they lack detailed temporal annotations (i.e., weakly supervised learning). To address this, GRIZAL employs a combination of supervised and self-supervised loss functions during training. **Supervised loss functions:** We use L_{BCE} (i.e., Binary Cross Entropy) and $L_{TemporalIOU}$ (Temporal Intersection Over Union) losses to train the model in a supervised fashion. L_{BCE} focuses on frame-level classification. It helps the model to discriminate between frames with and without the given action (AL), maintaining frame-wise accuracy. However, it treats each frame independently and does not enforce the continuity or duration of the action within the video sequence. $L_{TemporalIOU}$ complements BCE by considering the temporal structure of the action segments. $L_{TemporalIOU}$ evaluates the overlap between the predicted action segment and the ground-truth segment. It encourages the model to predict action segments that are temporally contiguous and have accurate start and end boundaries.

Self-supervised loss functions: For self-supervised training of the model, we use cosine similarity and InfoNCE losses. These losses operate on the embeddings produced by the VLE block (Z_{CLS}) and the main-stream block (Z_{ft}). While BCE and Temporal IOU optimize for accuracy with respect to known ground truths, Cosine Similarity and InfoNCE encourage the model to explore and exploit the inherent structure within the data. This addresses the challenge of over-reliance on labeled data and empowers the model to learn a more generalized and robust action representation. Cosine Similarity Loss ensures that the semantic information captured by the VLE block (which processes textual descriptions and related images) is aligned with the semantic content of the video frames processed by the main-stream block. The InfoNCE loss acts as a contrastive learning mechanism within the same embeddings. It pushes the model to increase the mutual information between corresponding video frames and action labels.

4 Experimental Setup

We perform experiments on ActivityNet-v1.3, THUMOS14 and Charades-STA datasets and use the dataset splits proposed by (Nag et al., 2022). We evaluate two scenarios. In *open-set scenario*, we have $D_{\text{train}} \cap D_{\text{val}} = \emptyset$, i.e., action categories for training and validation are disjoint. Here, we evaluate two splits, viz., 75%:25% and 50%:50%, of action categories in training and testing. In *closed-set scenario*, $D_{\text{train}} = D_{\text{val}}$. More details are provided in the Appendix.

4.1 Quantitative Results:

Open-set Scenario. As shown in Table 1, GRIZAL performs best on all metrics, including stringent criteria such as IOU@0.95. This underscores GRIZAL's ability to localize actions precisely. The higher performance across various IoU thresholds underscores GRIZAL's robustness in handling different levels of object overlap.

GRIZAL shows a notable improvement over other methods, such as EffPrompt. In contrast to GRIZAL, which extensively uses both text and generative images, EffPrompt only uses efficient prompting strategies with text, which restricts its ability to grasp the context. Furthermore, STALE fails, especially on the ActivityNet and THUMOS datasets, where descriptions are minimal, such as "girl in pink dress doing archery". STALE relies exclusively on these brief sentences, lacking the enriched contextual backdrop that GRIZAL employs. Moreover, GRIZAL's sophisticated technique handles complicated scenes more skillfully than the ICCV19 method (Nam et al., 2021), which simplifies phrases before feeding them to the encoder. This makes the ICCV19 method unsuitable for processing long-context videos.

While VideoCLIP and VAC are retrieval-based methods, other methods focus more on architectural aspects. GRIZAL provides substantially superior performance over retrieval-based methods such as VideoCLIP by incorporating diverse and contextually rich augmentations. Substantiating this hypothesis, training GRIZAL without DALL-E generated images and GPT-4 generated text separately results in performance deterioration. For example, in the first setting (75-25%), not using DALL-E generated images degrades mIoU by 8pp on ActivityNet

Table 1: Comparison with state-of-art under open-set scenario on ActivityNet and THUMOS14 dataset.

Method	**ActivityNet**				**Thumos14**			
	IOU@0.50	**IOU@0.75**	**IOU@0.95**	**mIOU**	**IOU@0.30**	**IOU@0.50**	**IOU@0.70**	**mIOU**
Open-set Scenario(75-25%)								
LGI(Mun et al., 2020)	32.4	17.0	3.1	17.9	37.9	20.0	3.2	19.1
VideoCLIP(Xu et al., 2021)	33.4	18.0	4.4	18.9	38.1	19.9	2.9	21.0
VAC(Wang et al., 2021c)	35.7	20.1	5.0	20.0	41.1	21.3	7.7	24.7
iCCV 19(Nam et al., 2021)	29.0	20.0	2.0	16.2	40.2	21.1	4.5	23.0
EffPrompt(Ju et al., 2022)	37.6	22.9	3.8	23.1	39.7	23.0	7.5	23.3
STALE(Nag et al., 2022)	38.2	25.2	6.0	24.9	40.5	23.5	7.6	23.8
GRIZAL(w/o DALL-E Images)	40.0	25.6	2.0	22.1	36.5	20.0	6.0	23.7
GRIZAL(w/o GPT-4 Text)	44.2	30.2	4.1	28.9	41.3	24.0	8.2	22.1
GRIZAL(w OpenWorld images)	46.3	31.8	5.9	29.9	42.3	25.0	9.5	26.6
GRIZAL	**46.4**	**32.5**	**6.8**	**30.1**	**43.2**	**25.7**	**9.8**	**27.0**
Open-set Scenario(50-50%)								
LGI(Mun et al., 2020)	28.9	15.2	2.1	19.8	35.0	18.7	4.2	19.1
VideoCLIP(Xu et al., 2021)	29.9	15.9	1.9	18.9	33.1	19.0	5.2	19.9
VAC(Wang et al., 2021c)	30.0	18.2	3.1	20.2	38.9	22.2	7.2	21.0
iCCV 19(Nam et al., 2021)	26.7	14.5	2.0	20.0	35.8	20.0	6.7	20.0
EffPrompt(Ju et al., 2022)	32.0	19.3	2.9	19.6	37.2	21.6	7.2	21.9
STALE(Nag et al., 2022)	32.1	20.7	5.9	20.5	38.3	21.2	7.0	22.2
GRIZAL(w/o DALL-E Images)	33.0	18.9	3.0	20.4	34.6	22.0	6.7	21.5
GRIZAL(w/o GPT-4 Text)	37.8	22.4	5.9	23.5	38.7	23.3	8.1	23.6
GRIZAL(w OpenWorld images)	39.3	24.8	6.3	24.9	38.9	23.9	8.8	24.6
GRIZAL	**39.9**	**25.7**	**6.6**	**25.7**	**40.0**	**25.0**	**9.1**	**25.2**

and 3.3pp on THUMOS14. Not using GPT-4 generated text degrades mIoU by 1.2pp on ActivityNet and 4.9pp on THUMOS14. These results quantitatively affirm that GRIZAL effectively simulates real-world scenarios by handling instances from both known and unknown classes more adeptly than existing CLIP-based methods. We further introduced a variant for our proposed method, viz., "GRIZAL (with OpenWorld images)", which uses open-world images from Wikipedia instead of DALL-E generated images. This variant achieves comparable performance on the ActivityNet and THUMOS14 datasets in both open and closed settings. This variant can save the cost of using the DALL-E model.

GRIZAL's improvement over STALE is more pronounced on the ActivityNet dataset than on THUMOS, likely due to the characteristics of each dataset. THUMOS, with its shorter actions in longer videos, demands higher localization precision. Additionally, untrimmed videos in THUMOS add background clutter and irrelevant scenes, affecting the sensitivity of IOU@0.5. Despite these challenges, GRIZAL consistently outperforms previous methods. The introduction of diverse augmentations reduces sensitivity to background clutter and enhances the model's ability to learn from shorter actions in longer videos.

Results Under Closed-set Scenario. Table 2 showcases GRIZAL's performance, compared with seven TAL methods featuring I3D encoder backbones and five CLIP-based methods. On both datasets, GRIZAL consistently surpasses existing TAL methods by a wide margin as the volume of labeled data grows.

Architecture-based methods like Context-Loc and VSGN robustly compete with CLIP-based state-of-the-art techniques like STALE. Conversely, VideoCLIP underperforms in closed-set settings, underscoring the importance of diverse samples for learning discriminative representations. This is further evidenced by our GRIZAL models (without DALL-E images and GPT-4 text), which show significant performance drops without generative augmentation.

Furthermore, the results on Charades-STA, as presented in Table 3, showcase GRIZAL's ability to learn much more complex scenes involving multiple actors, overlapping objects, and various interacting objects. GRIZAL outperforms the more recent state-of-the-art (SOTA) architectural-based

Table 2: Comparison with state-of-art under closed-set scenario on ActivityNet and Thumos14 datasets.

Method	Encoder	ActivityNet				Thumos14			
		IoU@0.5	IoU@0.75	IoU@0.95	mIoU	IoU@0.30	IoU@0.50	IoU@0.70	mIoU
TALNet(Chao et al., 2018)	I3D	38.2	18.3	1.3	20.2	53.2	42.8	20.8	39.8
GTAN(Long et al., 2019)	P3D	52.6	34.1	8.9	34.3	57.8	38.8	-	-
MUSES(Liu et al., 2021)	I3D	50.0	34.9	6.5	34.0	68.9	56.9	31.0	53.4
VSGN(Zhao et al., 2021)	I3D	52.3	36.0	8.3	35.0	66.7	52.4	30.4	50.1
Context-Loc(Zhu et al., 2021)	I3D	56.0	35.2	3.5	34.2	68.3	54.3	26.2	-
BU-TAL(Lin et al., 2021)	I3D	43.5	33.9	9.2	30.1	53.9	45.4	28.5	43.3
LGI(Mun et al., 2020)	I3D	43.2	29.1	6.0	31.0	66.3	54.3	30.0	49.8
VideoCLIP(Xu et al., 2021)	CLIP	42.1	23.4	4.1	29.8	65.5	52.4	26.8	47.6
iCCV 19(Nam et al., 2021)	CLIP	43.0	30.0	5.1	33.4	65.0	50.1	25.6	44.9
VAC(Wang et al., 2021c)	CLIP	44.0	31.1	6.1	34.0	67.9	56.7	32.0	51.1
EffPrompt(Ju et al., 2022)	CLIP	44.0	27.0	5.1	27.3	50.8	35.8	15.7	34.5
STALE(Nag et al., 2022)	CLIP	56.5	36.7	9.5	36.4	68.9	57.1	31.2	52.9
GRIZAL(w/o DALL-E Images)	CLIP	58.0	40.9	13.1	41.3	65.1	53.2	27.1	49.0
GRIZAL(w/o GPT4 text)	CLIP	62.2	48.0	16.0	43.4	68.9	57.9	31.8	53.3
GRIZAL(w OpenWorld Images)	CLIP	63.2	45.0	17.5	45.4	71.4	60.9	34.8	56.6
GRIZAL	CLIP	**64.0**	**53.1**	**18.7**	**46.3**	**72.4**	**62.7**	**36.7**	**57.8**

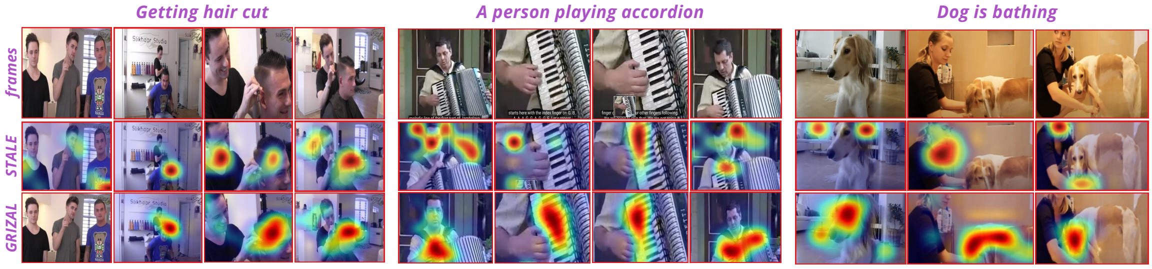


Figure 3: Gradient-activation maps for the ActivityNet dataset

Table 3: Closed-set scenario results on Charades-STA

Method	**Encoder**	**R@1**		**R@5**	
		IoU=0.5	**IoU=0.7**	**IoU=0.5**	**IoU=0.7**
CTRL (Gao et al., 2017)	C3D	23.6	8.9	58.9	29.5
2D TAN (Zhang et al., 2020b)	VGG	39.7	23.3	80.3	51.3
VSLNet (Zhang et al., 2020a)	I3D	47.3	30.2	-	-
UMT (Liu et al., 2022)	VGG	49.4	26.2	89.4	55.0
IVG-DCL (Nan et al., 2021)	C3D	50.2	32.9	-	-
M-DETR (Lei et al., 2021)	CLIP	55.7	34.2	-	-
LGI (Mun et al., 2020)	I3D	59.5	35.5	-	-
UnLoc-B (Yan et al., 2023)	CLIP	58.1	35.4	87.4	59.1
UnLoc-L (Yan et al., 2023)	CLIP	60.8	38.4	88.2	61.1
GRIZAL	CLIP	**62.1**	**41.0**	**91.2**	**64.0**

method UnLoC, which utilizes pre-trained image and text towers and feeds tokens to a video-text fusion model. This substantiates the importance of leveraging diverse and contextually rich augmentations, positioning GRIZAL as a superior alternative to CLIP-based approaches. The generative prior captures patterns of interactions involving object classes such as *football*, *microwave*, and *TV or LCD*. These interactions are more predictable, which benefits our approach more than previous baselines. For an object like *TV*, the spatial prior pattern of the interaction *(e.g.watch TV)* is more diverse and thus harder to model, resulting in only a tiny boost in the R@1 metric.

The Grad-CAM maps in Fig. 3 reveal GRIZAL's remarkable ability to model relationships between regions within images and the words present in the textual prompt. When prompted with *dog is bathing*, STALE primarily focuses on the term 'person' rather than *dog*. In contrast, GRIZAL accurately directs its attention to the 'dog' across the frames. Notably, in the third frame corresponding to the prompt "Getting a hair cut", GRIZAL focuses not only on the *person getting a haircut* but also on the *person performing the haircut*, effectively filtering out irrelevant information, e.g., background.

The t-SNE plots (Fig. 4) reveal that methods with limited diversity tend to produce less discriminative representations. The retrieval-based methods like VideoCLIP and VAC are prime examples of this trend. They lead to less discriminative representations. The embedding space of GRIZAL can separate the boundaries between the classes. GRIZAL exhibits higher discriminative capabilities, whereas the GRIZAL variants without GPT-4 and DALL-E generated augmentations have inferior capabilities.

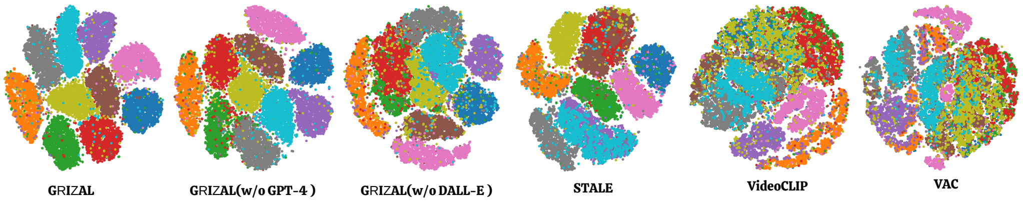


Figure 4: t-SNE plots on the ActivityNet dataset.

Clearly, diversity is essential to generalize to unseen data. This confirms our hypothesis that diversity serves as a pivotal contributor to the model's capacity to learn more discriminative representations, enhancing its generalization to unseen data. The diverse and contextually rich augmentations employed by GRIZAL help it generalize to the unseen data. This also explains GRIZAL's slightly higher performance in the open-set TAL setting.

4.2 Ablation Studies

Architectural Components. From Table 4(a), the performance metrics degrade on excluding either one or both of Fourier transform and optical flow. Optical flow captures the motion information to ensure temporal consistency in videos by aligning frames over time. Fourier transform captures frequency-based features. Combining their complementary strengths helps achieve temporal stability.

Effect of Multimodal Architectures in GRIZAL. As depicted in Table 4(b), GAFNet (Susladkar et al., 2023a) outperformed others with an IOU@0.5 of 64.0 and mIoU of 46.3, demonstrating superior fusion of visual and language cues. ViLBert and TCL, though competitive, had slightly lower scores, indicating potential limitations in capturing nuances for zero-shot scenarios. BART-Encoder closely followed GAFNet, showing promise in preserving critical information.

Effect of Augmentation Pairs (k) Increasing the number of augmentation pairs improves the model's robustness and generalization (Table. 4(c)) since the augmented data provides a more diverse set of examples for the model to learn from. The average inference latency (in ms) for a batch size of 16 for various k values is as follows: 400 ($k = 1$), 578 ($k = 2$), 654 ($k = 3$), 702 ($k = 4$), and 1200 ($k = 8$). Thus, the performance saturates for $k \geq 5$, whereas the computational overhead and inference latency rise rapidly. To balance these factors, we chose $k = 4$.

Effect of Loss Function. As per Table 4(d), the supervised loss contributes more to the model's detection capability than the self-supervised loss. Nonetheless, a model trained with only self-supervised loss may still capture important features, especially in scenarios where labeled data is limited or unavailable. This finding aligns with the favorable results observed for GRIZAL under open-set settings. Thus, GRIZAL can adapt to varying degrees of labeled data availability.

Effect of different frame encoder As shown in Table. 4(e), R(2+1)D encoder has the best temporal feature extraction performance. I3D and C3D, while effective, demonstrate marginally lower performance. ViT-B16/L does not quite reach the temporal performance levels achieved by R(2+1)D. Notably, the original GRIZAL model incorporates CLIP as a frame encoder, which works better than the abovementioned encoders. This underscores the importance of advanced temporal encoding for accurate action localization.

Effect of Token Size The best results were achieved with a token size of 30 (Table. 4(f)). A token size that is too long for augmented summarized text may lead to the loss of contextual information.

Results with open-source models: Table 5 illustrates GRIZAL's ability to integrate seamlessly with various open-source text and image generation models. Demonstrating its flexibility and modularity, GRIZAL works effectively with different combinations like LLaMa2 (Touvron et al., 2023) paired with Stable Diffusion (Rombach et al., 2022), as well as Mixtral (Jiang et al., 2024) with Pixart- α (Chen et al., 2023). This compatibility highlights GRIZAL's potential for broad applicability across diverse pre-trained generators.

Effect of the number of VLE layers: Table 6 presents an ablation study analyzing the effect of varying VLE layer counts (N) on model performance within the GRIZAL architecture (depicted in Figure 2). The study explores how different

Table 4: Ablation studies on ActivityNet dataset

	IoU@0.5	IoU@0.75	IoU@0.95	mIoU		IoU@0.5	IoU@0.75	IoU@0.95	mIoU
Full network	64.0	53.1	18.7	46.3					
(a) Architectural components(OF=optical flow)					(b) Multimodal architecture (Full (i.e. proposed) network uses GAFNet)				
w/o optical flow	61.3	49.0	16.8	45.0	VilBert	62.2	51.2	17.1	44.1
w/o Fourier	62.3	51.2	17.0	45.1	TCL	61.8	50.0	17.2	45.8
w/o Fourier and OF	59.9	49.9	15.0	44.9	BART-Encoder	64.1	53.1	18.0	46.6
(c) Augmentation pairs (full network uses k=4)					(d) Loss function (full network uses both loss functions)				
k=1	56.0	42.2	14.1	41.1	Only supervised loss	62.1	50.9	17.1	44.4
k=2	58.8	46.0	15.7	42.8	Only self-supervised loss	60.0	49.0	15.4	43.1
k=3	61.0	50.0	17.0	44.6					
k=8	64.17	53.2	18.8	46.5					
(e) Frame encoder used (Full Network uses CLIP)					(f) Number of generated text tokens (full network uses 30 tokens)				
I3D	61.0	52.1	17.0	45.9	10	63.0	51.0	16.0	43.0
C3D	62.1	52.0	16.0	45.4	20	62.0	52.8	17.0	45.9
ViT-B16/L	62.6	52.8	17.0	46.0	40	63.9	51.0	16.0	45.1
R(2+1)D	63.3	53.9	19.0	47.0	50	62.1	52.2	16.0	43.4

Table 5: Open-source model results on ActivityNet

TG/IG	IoU@0.5	IoU@0.75	IoU@0.95	mIoU
LLaMa2-7b / SD	0.6102	0.5101	0.16	0.4512
LLaMa2-34b / SD XL	0.6256	0.5205	0.17	0.4566
Mixtral - 8B / Pixart- α	0.6298	0.5279	0.17	0.4601

configurations of the Vision Language Embedding (VLE) block impact the IoU scores across various thresholds and the mIoU. As shown in Figure 2, the VLE block is responsible for generating multimodal representations by fusing image and text embeddings produced by pre-trained generators. As the number of layers (N) in the VLE block increases, the model's hidden size and the number of parameters are adjusted accordingly. These changes directly influence the joint multimodal embeddings, affecting the model's ability to handle complex interactions between visual and textual data. The ablation results suggest how adding more layers in the VLE block impacts the depth of the joint embedding, with improvements seen in the IoU metrics. This indicates that a deeper VLE block provides more capacity to encode multimodal information.

Table 6: Ablation of VLE layer-count (N)

N	Params	Hidden Size	IoU@0.5	IoU@0.75	IoU@0.95	mIoU
3	44M	512	0.5881	0.4602	0.1571	0.4278
7	79M	512	0.6411	0.5312	0.1808	0.4633
13	127M	768	0.6519	0.5392	0.1874	0.4678
17	189M	1024	0.6588	0.5418	0.1893	0.4702
23	227M	1024	0.6592	0.5447	0.1901	0.4709

5 Conclusion

We introduce GRIZAL, a novel framework designed for zero-shot action localization. GRIZAL can synthesize information from vision-language embeddings and optical flow. This novel approach is designed to recognize and interpret actions in videos by using multimodal clues without being exposed to the action labels during training. The strategic integration of multimodal embeddings and the tailored use of loss functions contribute to the model's exceptional performance.

6 Limitation and Future Work

Augmentation Bias and Mitigation. Due to GRIZAL's dependency on pre-trained generative models, its performance and fairness may be challenging to quantify. This issue can be addressed by creating more task-specific prompts to allow highly controlled generation of augmentations. Another approach is to fine-tune pre-trained models using low-rank adaptors to generate task-specific augmentations. In the future, we will explore these approaches to create a bias-free GRIZAL.

Additional Modalities. Currently, GRIZAL is limited to the vision-language modality. In future versions, we will adapt our model to other modalities, such as audio, for sound localization.

References

Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. 2018. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. 2023. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*.

Gayatri Deshmukh, Onkar Susladkar, Dhruv Makwana, Sparsh Mittal, et al. 2024. Textual alchemy: Coformer for scene text understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2931–2941.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2023. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15.

Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. 2021. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*.

Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. 2018. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Chen Ju, Zeqian Li, Peisen Zhao, Ya Zhang, Xiaopeng Zhang, Qi Tian, Yanfeng Wang, and Weidi Xie. 2023. Multi-modal prompting for low-shot temporal action localization. *arXiv preprint arXiv:2303.11732*.

Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. 2023. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 31–36. IEEE.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858.

Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. 2021. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3320–3329.

Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498.

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898.

Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. 2021. Multi-shot temporal event localization: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3042–3051.

Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. 2019. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 344–353.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819.

Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. 2022. Zero-shot temporal action detection via vision-language prompting. In *European Conference on Computer Vision*, pages 681–697. Springer.

Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. 2021. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1470–1479.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Senthil Purushwalkam and Abhinav Gupta. 2020. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058.

Onkar Susladkar, Gayatri Deshmukh, Dhruv Makwana, Sparsh Mittal, R Teja, and Rekha Singhal. 2023a. Gafnet: A global fourier self attention based novel network for multi-modal downstream tasks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5242–5251.

Onkar Susladkar, Prajwal Gatti, and Santosh Kumar Yadav. 2023b. Slbert: A novel pre-training framework for joint speech and language modeling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021a. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. 2021b. Self-supervised learning for semi-supervised temporal action proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1905–1914.

Xiao Wang and Guo-Jun Qi. 2022. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560.

Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. 2021c. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1459–1468.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. 2023. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Chen Zhao, Ali K Thabet, and Bernard Ghanem. 2021. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. 2021. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525.

Appendix Section

S.1 F-Transformer

Recent research on learnable Fourier transform has been applied in domains such as speech-language modeling (Susladkar et al., 2023b), natural language processing (NLP) (Lee-Thorp et al., 2021), and scene-text understanding (Deshmukh et al., 2024). Inspired by these advancements, we propose the F-Transformer block.

F-Transformer: The F-Transformer plays a crucial role in fusing modalities. It synthesizes information from distinct modalities—RGB video frames from the mainstream block, optical flow from the OFE block, and the visual-textual embeddings from the VLE block to construct an integrated and meaningful representation. It manipulates the input feature vectors Z_{rgb}, Z_{CLS}, $Z_{optical}$ to enhance action localization performance. This architecture facilitates cross-modality learning and underscores the relevance of the action label within the learned representations.

As shown in Fig. S.5, the feature vector Z_{rgb} from the mainstream block is first subjected to a Multi-headed Self-Attention (MSA) mechanism. This self-attention mechanism allows the model to capture relationships between different positions in the sequence, enhancing its ability to understand the temporal dynamics of the video. After residual connection and layer normalization, the output is combined with the Z_{CLS} vector from the VLE Block through pointwise addition. This helps the model by providing more context relevant to the action label.

Following the integration of these vectors, the Fast Fourier Transform (FFT) is applied to transform the features into the frequency domain. This elucidates the distribution of features across various frequencies. Following FFT, an MLP layer captures distribution patterns within the frequency domain. This allows the model to emphasize important properties while downplaying less important ones. Subsequently, an inverse FFT (iFFT) operation is performed to map the features back to the spatial domain.

The concluding MSA module in our novel F-Transformer utilizes the optical flow feature ($Z_{optical}$) as both key and value vectors, with the preceding feature map serving as the query vector. This configuration enables the MSA to synthesize attention-pooled characteristics conditioned reciprocally across modalities. In particular, the optical flow features outline the action dynamics in the video frames, highlighting areas of motion and stillness. At the same time, the visual-textual embeddings provide a visual and contextual comprehension of the action label, assisting in accurately identifying frames where the action of interest takes place. The features are then passed through additional MSA and MLP layers, with layer normalization (LNorm) applied after each operation to stabilize the learning process. The final output of the F-Transformer block is denoted as Z_{ft}, which incorporates both the detailed spatial features and the refined frequency domain features. Doing so captures a comprehensive understanding of the video's content, both in terms of the global context provided by the self-attention and the local, detailed motion patterns highlighted by the optical flow. This multifaceted approach allows for a more accurate and robust action localization within videos.

S.2 Details of Experimental Platform

S.2.1 Datasets

1) The ActivityNet-v1.3 dataset (Caba Heilbron et al., 2015), comprises 19,994 videos spanning 200 action classes. To adhere to the standard evaluation protocol, we partitioned the videos into training, validation, and testing subsets, maintaining a ratio of 2:1:1. 2) The THUMOS14 dataset (Idrees et al., 2017) includes 200 validation videos and

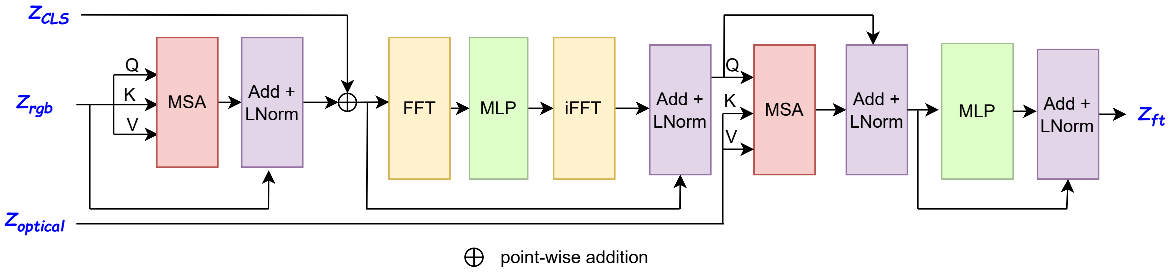


Figure S.5: F-Transformer

213 testing videos distributed across 20 action categories. Notably, THUMOS14 provides labeled temporal boundaries and action classes for each video. 3) The Charades-STA dataset (Gao et al., 2017), consists of 6,672 videos and 16,128 segment/caption pairs. We utilized 12,408 pairs for training purposes and 3,720 pairs for testing. Each video in Charades-STA is annotated with an average of 2.4 segments, with an average segment duration of 8.2 seconds.

S.2.2 Implementation Details

We leverage a multi-GPU distributed training setup with Nvidia A100-40GB GPUs, employing a total of 16 devices. The chosen hyper-parameters are tuned for optimal performance. The learning rate is set to 5e-07. The training is conducted in batches of 24 samples, while the evaluation utilizes larger batches of 56 samples. The total batch size for training is 2400, and for evaluation, it is 900. The Adam optimizer, configured with betas=(0.9, 0.999) and epsilon=1e-08 is used. The learning rate scheduler is linear, with a warm-up ratio of 0.1, providing a gradual increase in learning rates during the initial training epochs. The entire training process spans three epochs, ensuring sufficient exposure to the dataset while avoiding overfitting. For the CLIP encoders, video frames underwent preprocessing to achieve a spatial resolution of 224×224. The maximum number of textual tokens was constrained to 77, aligning with the original CLIP design. Each video's feature sequence, denoted as F, was rescaled to $T = 100/256$ snippets for ActivityNet/THUMOS/Charades, leveraging linear interpolation techniques to ensure consistency and accuracy in the temporal domain.

Evaluation Metrics: Following (Nag et al., 2022), we compare the results in two types of metrics: (1) Recall at various intersection over union thresholds (R@tIoU). It measures the percentage of predictions that have larger IoU than the thresholds (we use threshold values of $\{0.5, 0.75, 0.95\}$ for Activitynet dataset, $\{0.3, 0.5, 0.7\}$ for Thumos14 dataset and $\{0.5,0.7\}$ for Charades-STA dataset.). (2) Mean intersection over union (mIoU) is an averaged temporal IoU between the predicted and the ground-truth region.

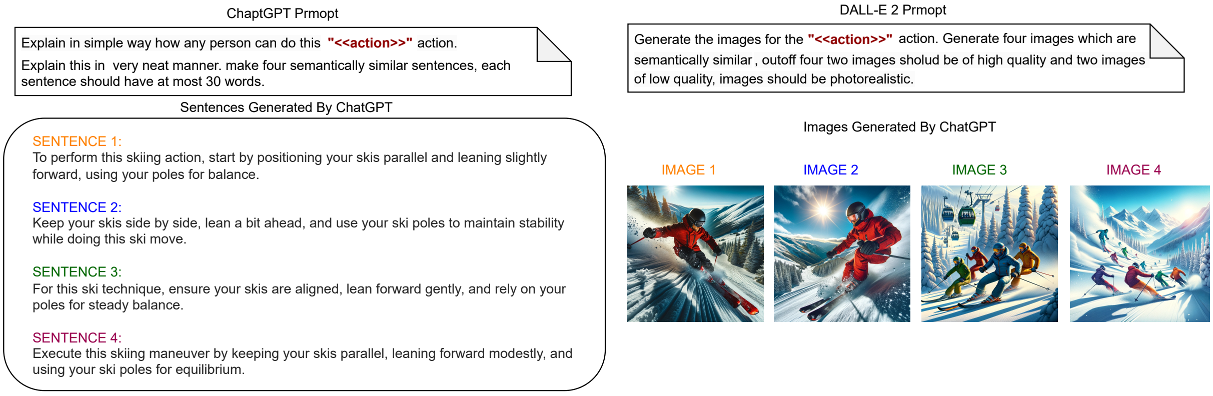
Consistent hyperparameters were maintained across all the evaluated networks for a fair comparison.

S.2.3 Visual Example of Generative Priors

As illustrated in Figure S.6, when a prompt is processed through the GPT-4 and DALL-E APIs, it generates multiple augmented, summarized textual descriptions and corresponding images. By creating cross-modal representations that integrate these summarized descriptions with generated images, we ensure stronger semantic consistency and a deeper contextual understanding of the original textual prompt. For instance, in Figure 2, without a nuanced understanding of both the generated descriptions and images, a system might struggle to differentiate between deliberate speed control in skiing and unintentional sliding in unrelated scenarios. However, with this contextual integration, the system can accurately interpret detailed skiing techniques, such as snow ploughing and parallel turns, enabling it to recognize intentional speed adjustments specific to skiing. This cross-modal approach significantly enhances the model's grasp of complex, domain-specific concepts.

S.3 Additional Qualitative Results

Qualitative analysis, as shown in Fig. S.7, further demonstrates the critical role of diversity in mitigating over-complete and under-complete representation issues in TAL. The results emphasize the significance of leveraging diverse samples, as demonstrated by GRIZAL, to enhance the model's



ChaptGPT Prmopt

Explain in simple way how any person can do this "<<action>>" action.

Explain this in very neat manner. make four semantically similar sentences, each sentence should have at most 30 words.

Sentences Generated By ChatGPT

SENTENCE 1:
To perform this skiing action, start by positioning your skis parallel and leaning slightly forward, using your poles for balance.

SENTENCE 2:
Keep your skis side by side, lean a bit ahead, and use your ski poles to maintain stability while doing this ski move.

SENTENCE 3:
For this ski technique, ensure your skis are aligned, lean forward gently, and rely on your poles for steady balance.

SENTENCE 4:
Execute this skiing maneuver by keeping your skis parallel, leaning forward modestly, and using your ski poles for equilibrium.

DALL-E 2 Prmopt

Generate the images for the "<<action>>" action. Generate four images which are semantically similar , outoff four two images sholud be of high quality and two images of low quality, images should be photorealistic.

Images Generated By ChatGPT

Figure S.6: Vision-Language Generative Priors

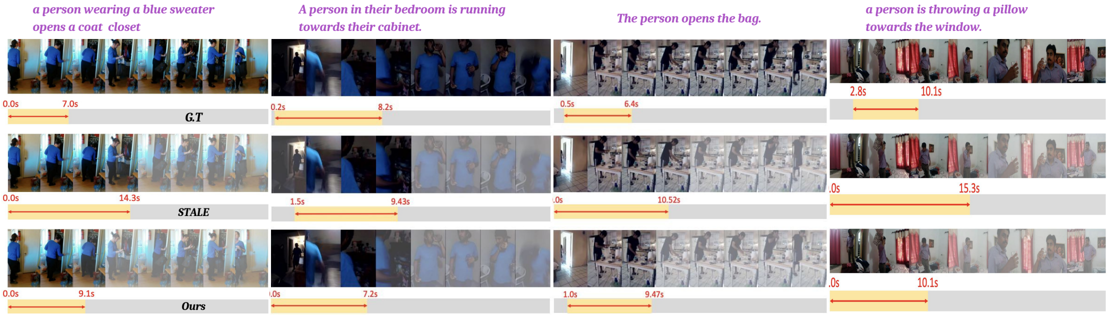

a person wearing a blue sweater opens a coat closet
A person in their bedroom is running towards their cabinet.
The person opens the bag.
a person is throwing a pillow towards the window.
0.0s
7.0s
G.T
0.2s
8.2s
0.5s
6.4s
2.8s
10.1s
0.0s
14.3s
STALE
1.5s
9.43s
.0s
10.52s
.0s
15.3s
0.0s
9.1s
Ours
.0s
7.2s
1.0s
9.47s
.0s
10.1s

Figure S.7: Qualitative Maps on ActivityNet Dataset showing *Boundary Localization* ability of GRIZAL and STALE with respect to Ground Truth.

precision in localizing temporal action boundaries. GRIZAL consistently exhibits more controlled localization of boundaries, effectively addressing over-complete and under-complete representation issues. In contrast, STALE suffers from over-complete representation in all cases.