Image Label Disambiguation for Rich Semantic Representations in Language-Prompted Segmentation Models

Justin Hall
Stanford University
jhall2025@stanford.edu

Walter Lopez Chavez
Stanford University
walterch@stanford.edu

Abstract

The field of video segmentation, specifically through manual frame-by-frame image masking to create ground truths for downstream applications, is tedious. For automatic ground-truth labeling, both disambiguation of uncommon terms and generalization of others are essential. With utilization of the Densely Annotated VIdeo Segmentation (DAVIS) dataset, we introduce What am I Looking At (WAILA), a chaining model combining classic NLP techniques (Word2Vec and ConceptNet Numberbatch [CNet-NB]) with the text prompt to image segmentation capabilities of Language Segment-Anything (LangSAM). We find that WAILA leads to improvements in mask generation both quantitatively at an average improvement of 39.7% using common accuracy metrics (IoU) and qualitatively for the DAVIS dataset, but leaves questions for generalization to unseen data, possibly due to a small dataset size and specific, misleading terms generated from Word2Vec and CNet-NB.

1. Introduction

Imagine you are a researcher creating a pixel-wise annotated video dataset featuring a piece of second-hand furniture as the primary object to be segmented and labeled. However, because the piece of furniture was second-hand, it's difficult to determine its exact name, so you go with the most closely related name you can think of. As a result of the ambiguous label provided, the language-prompted segmentation model you planned to use for automatic annotation struggles to properly segment this object throughout the videos you want to include your dataset. Thus, how can we achieve accurate object segmentation in videos when the object's name is unknown or ambiguous?

The creation of training datasets such as this one is a core problem to computer vision and semantic segmentation as the cost of manually annotating images with pixel-wise semantic labels is cost prohibitive due to the intensive labor involved from using human annotators [1]. Working with the data available, recent models such as Segment-Anything (SAM) and CLIP allow for language prompted segmentation in an image thus extending automation over this task. However, to use a segmentation model to accurately label an unknown object or use an ambiguous prompt, the model would require the same data needing to be labeled.

To address this, we propose an algorithm called What am I Looking At, or WAILA, a method of disambiguating the initial object label for an image such that a language prompted segmentation model can leverage over its previously learned representations of objects and combine segmentation masks to more accurately segment a semantically unknown object. Using the Densely Annotated Video Segmentation (DAVIS) dataset we start with samples of fully annotated pixel-wise segmented images, taken from a set of 50 HD video sequences each with potentially ambiguous labels, and attempt to disambiguate these initial labels using Word2Vec and CNet-NB which are pretrained language models to find nearest neighbor terms. Our outputs are a set of conceptually related terms to the initial prompt, and using SAM as our segmentation model we determine optimal combinations of these related words and their corresponding segmentation masks by finding the combination of output masks with the highest match to the ground truth annotation from DAVIS. Broadly, the aim is to disambiguate initial labels to fit within the learned semantic representations of pretrained segmentation models and align with the desired example mask, effectively extending a pretrained model's usability for unknown objects and ambiguous labels without further training and minimal ground truth labels.

2. Related Work

2.1. Image Segmentation

We looked at the general idea of image segmentation and language prompted segmentation models as described in Minaee et al. [4] and Li et al. [1] to inform our problem. From both papers, a lack of labeled training data severely

hampers the development of downstream models and applications. Li et al. describes their inspiration from CLIP, a contrastive learning model [8], to inform their method of combining text and image embeddings to allow for language prompted segmentation. CLIP was also considered as inspiration on how to encode natural language to reference visual objects for downstream tasks. The segmentation model used in our project is based on the Segment Anything Model (SAM 2) [9], with SAM 2 itself an encoder-decoder based model which is common for image-to-image tasks [4].

2.2. Data

We then searched for an appropriate dataset to match our segmentation model and our problem, and we found the Densely Annotated VIdeo Segmentation (DAVIS) dataset produced by Perazzi et al. [7] which provided high quality video data with pixel-wise annotations and raw images. From the DAVIS dataset we noticed the diverse range in ambiguity for its labels, with some labels being very accurate and easy to understand and others being uncommon or even previously unknown words.

2.3. Word Disambiguation

In order to find methods of disambiguating these ambiguous labels, and making them easier for the underlying SAM 2 model to segment their corresponding objects, we looked at a survey on the topic produced by Ranjan et. al which proved word disambiguation is not a simple topic in natural language. In fact, it is an NP-complete equivalent problem due to the complexity of how natural language requires context and knowledge to describe inner semantic relationships [5]. Regardless, we found notable models for this application which were light-weight and deterministic to allow for repeatable experiments: Word2Vec and ConceptNet Numberbatch [12]. Word2Vec is a well-known language model trained on the Google News corpus to construct word vectors for the first one million most commonly used words [3]. For this project Word2Vec was implemented using gensim, an open-source Python library used to load many language models [10]. Additionally, we looked at ConceptNet Numberbatch which is an update to ConceptNet, with ConceptNet being a knowledge graph model of word vector embeddings notably focusing on the meaning of words rather than solely relational semantics [11]. Thus, Conceptnet Numberbatch, which we will refer to as CNet-NB, is an improvement on the base ConceptNet as it retrofits its knowledge graph with the learned vectors of Word2Vec and GLoVE [12], where GLoVE is another learned vectorized representation of word vectors [6]. Notably, we did not pursue LLMs due to their large size, risk of hallucinations, and lack of reproducibility from their non-deterministic nature during prompt evaluations.

3. Dataset and Features

Our main dataset throughout our project was DAVIS, specifically the original 2016 dataset. DAVIS has 50 high-quality videos of various categories, each with a starting label via the folder name for the category, in both 480p and 1080p, segmented into individual JPEG frames. For the sake of simplicity, we only focused on the 480p images for WAILA as they resulted in faster calculation for our segmentation model.

DAVIS also provides pixel-accurate, per-frame ground-truth, binary masks for each category. These masks were carefully and manually annotated by humans. Furthermore, DAVIS intentionally selected their video data to have one object of note at a time (or two spatially connected objects), as having a single object per sequence simplifies the detection performed by the segmentation model used, which in this case is SAM 2 [7].

Overall, DAVIS contains 3,455 JPEG images, 3,455 ground truth masks, and 50 categories to experiment with. For our purposes, we only used the categories that SAM segmented poorly in order to test if these same categories could be improved using WAILA at all. Thus, we used a subset of the DAVIS dataset, selecting the 20 categories out of 50 that SAM performed the worst on via an Intersection Over Union (IoU) baseline metric. This process will be discussed in more detail in the following Methods section.

An example of the original image for frame 6 of the video labeled ”bmx-bumps”, its ground-truth mask, and its baseline generated mask, can be found below:

Figure 1: JPEG Example Image for category ”bmx-bumps”

Figure 2: Ground Truth Annotation for ”bmx-bumps”

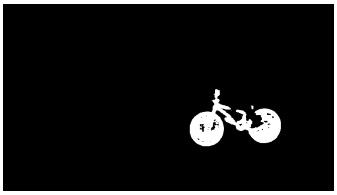
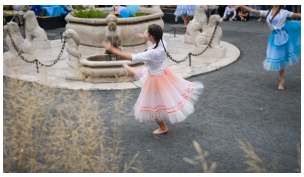


Figure 3: Baseline Mask Generation via SAM for ”bmx-bumps”

We additionally selected two categories out of the remaining 30 (”scooter-black” and ”dance-twirl”) as unseen test categories in order to test for generalization of WAILA, as the scooter-black video was comparable to scooter-gray, and dance-twirl was comparable to dance-jump, with scooter-black and dance-jump respectively being optimized through WAILA. As a visual metric to illustrate comparability, Figure 4 shows the similarity of the categories for the first frame of each respective video.



(a) ”dance-jump”

(b) ”dance-twirl”

Figure 4: Side-by-side comparison of ”dance-jump” and ”dance-twirl”.

4. Methods (2 pages)

Our main methodology for WAILA contains three steps:

1. Utilize word vector representations in order to generate similar words to an ambiguous category label.

2. Generate masks with SAM for each similar word.

3. Generate the set of all subsets of mask combinations (power set), selecting the subset of terms that led to the highest Intersection Over Union (IoU) score with the ground truths.

4.1. Similar Word Generation

To generate similar words compared to a given starting category label, we explored two different natural language processing (NLP) models: Word2Vec and ConceptNet.

Word2Vec is an NLP model that transforms words into vector embeddings, allowing for quantitative comparison among words. With vector representations for words A and B, the cosine similarity equation below can be applied in order to see how similar A and B are to one another:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

A higher cosine similarity indicates that A and B are highly similar, whereas a smaller score indicates dissimilarity and a lack of correlation.

ConceptNet, as shown in 5, is another NLP model similar to Word2Vec, but instead of relying solely on vector representations, it employs a knowledge graph that connects words and phrases of natural language with labeled, weighted edges, allowing for a more intricate representation of meaning in words [11].

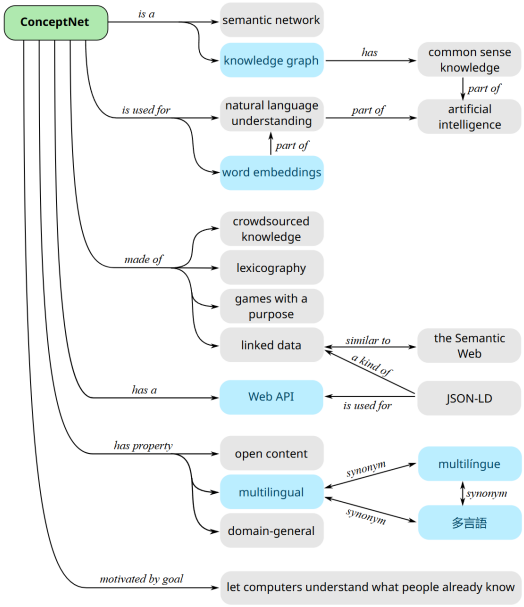


Figure 5: Visual Explanation of how ConceptNet works

For WAILA, we used one of its branches, ConceptNet Numberbatch (CNet-NB). CNet-NB is a combination of many other models like Word2Vec and GloVe. It contains a set of semantic vectors that can be used directly as in Word2Vec in order to generate similar words conceptually via cosine similarity [12].

Given that our project is largely prompt-tuning and word-disambiguation based, we hypothesized that CNet-NB would perform well on the DAVIS dataset in comparison to Word2Vec.

4.2. Mask Generation

For video frame segmentation, we rely on Language Segment-Anything (LangSAM), a zero-shot text prompting to segmentation model that uses SAM 2.1 for visual segmentation. [2] LangSAM takes a set of images and a set of text prompts as input and outputs a set of masks (some-

times multiple per text prompt for an image), along with confidence scores, as output.

For consistency, we use `sam2.1_hiera_small`, a masked image encoder and pretrained hiera image encoder, which is hierarchical, allowing for the use of multiscale features during decoding. [9]. This was due to its balance of performance and fast runtime when compared to the hiera-tiny, hiera-medium, and hiera-large models.

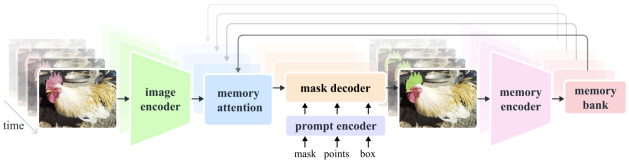


Figure 6: SAM 2.1 Architecture

Due to our selection of DAVIS as our dataset– and thus not needing to consider the detection of multiple objects within an image, we made a simplification for WAILA; we would always select just the highest confidence mask from LangSAM for each provided text prompt for a given image.

Figure 7 illustrates an example on Gradio, demonstrating how LangSAM and the underlying Sam 2.1 work with a JPEG image of a bear from the DAVIS dataset.

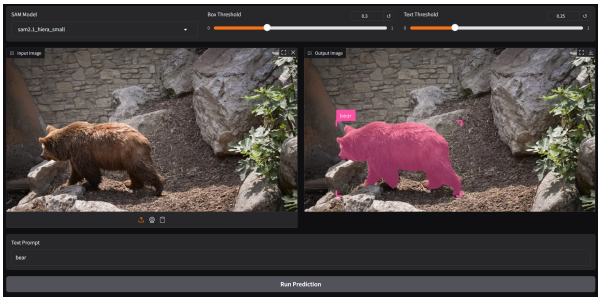


Figure 7: LangSAM demonstration on an image from category "bear"

4.2.1 Baselines

Since DAVIS provides a starting label for every category, for our baseline, for each category, we ran LangSAM with the starting label as the text prompt over all the frames of the category, retrieved the first (highest confidence) mask, and compared the mask of each frame to the ground truth annotation from DAVIS. For each frame i, we computed IoU with ground truth mask G_i and the LangSAM-generated mask P_i. IoU between G_i and P_i represents the quality of our generated mask P_i. A high IoU score for P_i signifies that we've identified most true parts of the ground truth mask (recall) and didn't falsely identify parts of the G_i that were not originally there (precision). We then average out the IoU scores for each frame to calculate an IoU score for each category. The equation we used can be seen below, where N represents the number of frames in a category:

$$\overline{\text{IoU}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|G_i \cap P_i|}{|G_i \cup P_i|}$$

Once achieving our baselines, we selected the 20 categories (indicated in bold in Table 1 with the lowest average IoU scores to experiment with WAILA on.

Table 1: Baselines Table

Category	Avg IoU	Category	Avg IoU
car-roundabout	0.9847	boat	0.7783
car-shadow	0.9755	motocross-jump	0.7121
car-turn	0.9751	horsejump-low	0.7102
rhino	0.9749	scooter-black	0.7070
bear	0.9654	horsejump-high	0.6828
drift-turn	0.9624	**scooter-gray**	**0.6624**
elephant	0.9614	**drift-chicane**	**0.6415**
cows	0.9560	**parkour**	**0.6312**
breakdance-flare	0.9530	**paragliding-launch**	**0.6080**
soccerball	0.9382	**stroller**	**0.5634**
dog-agility	0.9268	**motorbike**	**0.5607**
bus	0.9238	**kite-walk**	**0.5544**
goat	0.9217	**motocross-bumps**	**0.5396**
breakdance	0.9142	**drift-straight**	**0.4657**
train	0.9050	**rollerblade**	**0.3313**
libby	0.9036	**bmx-bumps**	**0.2823**
lucia	0.9014	**kite-surf**	**0.2805**
blackswan	0.8898	**hockey**	**0.2600**
flamingo	0.8888	**bmx-trees**	**0.1746**
mallard-fly	0.8829	**swing**	**0.1230**
paragliding	0.8820	**surf**	**0.1036**
dance-twirl	0.8792	**dance-jump**	**0.0710**
camel	0.8665	**hike**	**0.0001**
dog	0.8640	**mallard-water**	**0.0000**
tennis	0.8007	**soapbox**	**0.0000**

4.3. WAILA

4.3.1 Finding the best term

For every category, we generated n_{terms} amount of similar terms using an input $word\_model$; this word model was either the aforementioned Word2Vec or CNet-NB. In contrast to the baseline model, we sampled n_{samples} images from each category to experiment with and optimize. We then ran a prediction with LangSAM for every term in the sampled image, generating a set of baseline masks $M = [M_1, M_2, ..., M_t]$ for each of the t terms.

With the baseline masks, we then formed the powerset $P(M)$ of all baseline masks, combining masks via addition. Subsequently, we performed IoU with each element of $P(M)$ and the ground truth of the sampled frame. We save the mask and best combination of terms for sampled frame z (BC_z) that achieve the highest IoU score. In the example case, this subset of masks was solely the term "bump".

(a) 'bmx' (b) 'bump' (c) 'camber-corners'

Figure 8: Masks for Similar Terms of Sampled Image 1 of category "bmx-bumps". Similar terms were "bmx", "bump", and "camber-corners"

Figure 9: Best combination: ["bump"]

4.3.2 Evaluating best combination's performance

Once we find BC_z for a category and sample frame z, we run LangSAM on all other frames of the category, using each term of BC_z as a text prompt. We combine all the masks in an analogous way to BC_z, and run average IoU with the resulting masks across all ground truth frames for the category. This produces IOU_z.

We repeat this entire process for all $n\_samples - 1$ other sample frames. Once all sample frames have concluded, we save the subset of terms BC^*

$$z^* = \arg \max_{z \in \{1,\ldots,n_{\text{samples}}-1\}} \text{IoU}_z$$

$$BC^* = BC_{z^*}$$

which led to the highest IoU score: this is our final prediction of terms for the category. We then repeat this for every the remaining 19 categories to disambiguate a subset of terms for every category.

4.4. Codebase

In our code, we utilized both the DAVIS dataset [7] and added to the LangSAM GitHub repository [2]. The Word2Vec model used was loaded using `gensim`, a python library [10]. Everything in the "WAILA (Our Code for CS 231N)" folder in lang-segment-anything and in the "similar_words_generation" folder in the base CS 231N project folder is our code, everything else comes from either DAVIS or LangSAM.

5. Experiments/Results/Discussion

5.1. Description and Purpose of Experiments

5.1.1 WAILA Hyperparameter Experiments

With WAILA implemented, we ran three experiments with the following hyperparameters:

- **SmallW2V:** $n_{\text{terms}} = 5$, $n_{\text{samples}} = 3$, `word_model` = W2V.
- **LargeW2V:** $n_{\text{terms}} = 10$, $n_{\text{samples}} = 5$, `word_model` = W2V.
- **CNet-NB:** $n_{\text{terms}} = 10$, $n_{\text{samples}} = 5$, `word_model` = CNet-NB.

The purpose of these experiments was to determine whether WAILA would lead to higher IoU scores with the ground-truth masks—i.e., whether the combined masks would more closely align with the annotations.

5.1.2 Generalization Evaluation

Each experiment outputs a set BC_i^* of terms for experiment i. We then selected two categories that were not used in our WAILA pipeline—namely, `scooter-black` and `dance-twirl`—and performed the following evaluations on experiments 2 and 3 from above ($i = [2, 3]$):

(a) Ran LangSAM on category `dance-twirl` using each different BC_i^* for `dance-jump`, and compared these results to using the tag `dance-jump`.

(b) Ran LangSAM on category `scooter-black` using each different BC_i^* for `scooter-gray`, and compared these results to using the tag `scooter-gray`.

The goal of these experiments was to evaluate WAILA's generalization. In a real-world setting—where annotated ground-truth masks for unseen data are unavailable; we wished to determine whether the selected term subsets would generalize to similarly unseen scenarios.

5.2. Quantitative Experiments / Evaluation

5.2.1 CNet-NB vs Word2Vec vs Smaller Word2Vec vs Baselines

For the table below 2, we found the best IoU scores for each model and we calculate the minimum improvement in the column `min_improv`.

The minimum increase in IoU accuracy was only 1.2 percent whereas the largest increase was 94.9 percent. The median IoU improvement was 39.7 percent, with a std dev. of 28 percent. We find that in general, there is a noticeable improvement in the IoU score following our methodology. However, it may also be important to note we focused on

Table 2: Baseline to Model Comparison

Category	baseline	small_w2v	large_w2v	conceptnet	min_improv
scooter-gray	0.662	0.865	0.846	**0.867**	0.205
drift-chicane	0.641	0.907	0.919	**0.92**	0.279
parkour	0.631	**0.953**	**0.953**	**0.953**	0.322
paragliding-launch	0.608	0.616	0.618	**0.62**	0.012
stroller	0.563	0.563	**0.924**	0.598	0.361
motorbike	0.561	0.568	0.569	**0.578**	0.017
kite-walk	0.554	0.728	0.749	**0.75**	0.196
motocross-bumps	0.54	0.595	0.916	**0.92**	0.38
drift-straight	0.466	**0.96**	**0.96**	**0.96**	0.494
rollerblade	0.331	**0.938**	**0.938**	**0.938**	0.607
bmx-bumps	0.282	0.417	0.417	**0.504**	0.222
kite-surf	0.281	0.688	**0.696**	0.685	0.415
hockey	0.26	0.905	0.906	**0.907**	0.647
bmx-trees	0.175	0.001	**0.471**	0.384	0.296
swing	0.123	0.858	0.864	**0.868**	0.745
surf	0.104	0.834	**0.835**	0.832	0.731
dance-jump	0.071	0.81	0.811	**0.814**	0.743
hike	0	0.938	**0.946**	0.945	0.946
mallard-water	0	**0.949**	**0.949**	**0.949**	0.949
soapbox	0	0.293	0.543	**0.658**	0.658

the lowest performing baseline IoU scores from the initial category label. We see that for the smaller word2vec sample size, its performance was only comparable to its larger word sample generations when the resulting score across all experiments was the same. In general, for the same model word2vec performed either the same as its smaller counterpart, or with a marginal improvement. Additionally, we also see that CNet-NB receives the highest score 11/20 times, larger Word2Vec 5/20 times, with all three models tying 4/20 times. Thus we conclude that CNet-NB is twice as likely to result in the highest improvement compared to Word2Vec, ignoring ties. The figure below 10 includes a clearer visualization for the table.

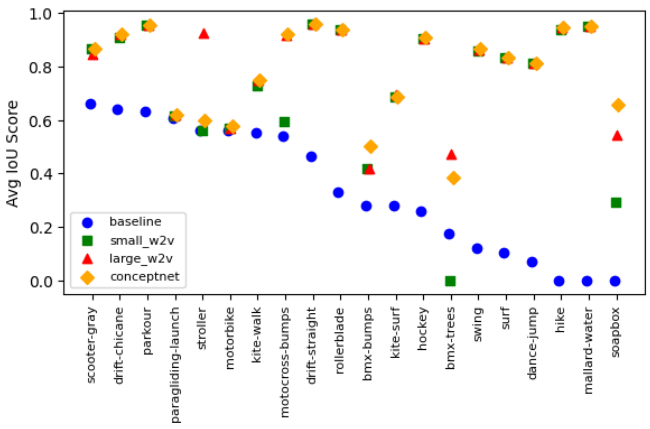


Figure 10: Avg IoU Score Per Category Per Model

5.3. Qualitative Evaluation / Analysis

5.3.1 Mask Generation Analysis

Our qualitative results from WAILA on DAVIS– our output mask visualizations– saw notable improvements as well.

As an example, our largest categorical improvement was the `mallard-water` example. SmallW2V suggested [”mallard”], LargeW2V suggested [”bluebills”], and CNet-NB suggested [”aquatile”], all yielding the same score of 0.949. These sets of terms all relate to mallard in terms of cosine similarity, but notably omit ”water”. In general, this highlights that finding the ”optimal” prompt will require precision.

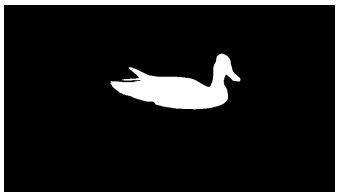
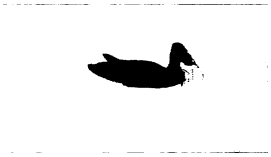


Figure 11: DAVIS Ground Truth for ”mallard-water”



(a) Baseline Mask for category: ”mallard-water”

(b) WAILA Mask (CNet-NB) for category: ”mallard-water”

Another example, `soapbox` had a different issue at baseline; LangSAM was not generating any mask at all as it had likely never seen the term ”soapbox” before.

In the context of DAVIS, a soapbox is synonymous with a gravity racer, which is a motorless vehicle raced on a downhill course, propelled purely by gravity. While the word2vec models did improve on the baseline, WAILA recommended largely unhelpful term combinations: [”whitty-retort”, ”whoop-dee”] and [”whitty-retort”, ”whoop-dee”, ”meaningless-drivel”, ”bludge”] for SmallW2V and LargeW2V respectively.

However, WAILA with CNet-NB was able to extract a different set [”soapbox-car”, ”stump-orator”], which notably contains ”soapbox-car”, giving LangSAM the necessary context in order to produce more meaningfully appropriate masks.

5.3.2 Word Generation Analysis

The table below is an example of the generated output from the Word2Vec model versus the CNet-NB model for

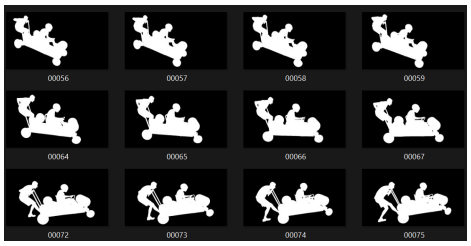


Figure 13: 12 DAVIS Ground Truths for category: ”soapbox”

(a) 12 Baseline Masks for the category: ”soapbox”

(b) 12 WAILA Masks (CNet-NB) for category: ”soapbox”

the top 10 related words for a given input ”rollerblade” 3.

Table 3: Word2Vec vs ConceptNet Numberbatch

”rollerblade”	Word2Vec	CNet-NB
1	”roller_blade”	”in_line_skate”
2	”biking”	”rollerblades”
3	”rollerblading”	”in_line_skater”
4	”rollerblades”	”in_line_skates”
5	”roller_blading”	”roller_blade”
6	”roller_bladers”	”rollerblading”
7	”bicyclers”	”rollerblader”
8	”Rollerblading”	”roller_skate”
9	”jogging_biking”	”rollerskater”
10	”rollerskate”	”roller_boot”

Although this is a small example, we can already see that Word2Vec and CNet-NB find similar words a bit differently. Word2Vec emphasizes similar words almost to the same spelling, with most of its suggestions being some form of ”rollerblade” whereas CNet-NB keeps the ”roller”-prefix and looks at iterations of the post-fix ”-skate.” Interestingly Word2Vec outside of its rollerblade-like suggestions also includes almost completely different terms in ”biking” and ”bicyclers.” In practice, Word2Vec is more likely to go off-topic from the original prompt whereas CNet-NB is generally better at keeping on-topic for its suggestions. However, from our results 10 we see that performance is comparable for both, with CNet-NB being slightly better.

5.4. Experiment on Unseen Data

5.4.1 Scooter-black

As described in 5.1.2, we tested the BC_i^* terms for `scooter-gray` on the set of images from `scooter-black`. For the sake of having a quantitative metric, we used the IoU process as described in section 4.3.2, though it should be noted that a quantitative metric would not be plausible for these experiments beyond the scope of this paper.

Table 4: Scooter-Gray $->$ Scooter-Black

	Baseline	LargeW2V	CNet-NB
Best Term(s)	[”scooter-gray”]	[”ungray”, ”gray”, ”grayly”, ”grayen”, ”maxis-cooter”]	[”gray”, ”Mongoose_mountain”]
Scooter-Black Avg IoU Score	0.707	0.612	0.589

For the scooter test, both LargeW2V and CNet-NB did significantly worse than simply using the baseline label of `scooter-gray`, receiving Average IoU scores of -0.095 and -0.118 relative to the baseline. We hypothesize that our word models may have over-emphasized the ”gray” part of ”scooter-gray” (Word2Vec's optimal terms were found to include ”ungray”, ”gray”, ”grayly”, and ”grayen”, for instance). In this specific scenario focusing on the ”scooter” semantically would have been more beneficial for generalization. Visual results for the ”scooter-black” category can be seen in Figures 15 and 16

Figure 15: Baseline Scooter Results for 12 Frames of ”Scooter-Black” (Avg IoU = 0.707

5.4.2 Dance-twirl

For the `dance-twirl` test, WAILA also did not perform as well as expected, as LargeW2V and CNet-NB performed with scores of -0.004 and $+0.004$ relative to the baseline respectively via Table 5, indicating an essentially negligible difference.

Because the categories we originally selected for WAILA were the 20 worst-performing categories (the best of those 20 being scooter-gray with a baseline of 0.662),

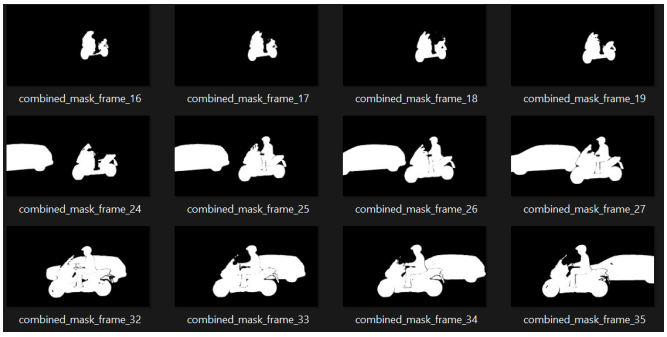


Figure 16: W2V Scooter Results for 12 Frames of "Scooter-Black" (Avg IoU = 0.589)

Table 5: Dance-Jump $->$ Dance-Twirl

	Baseline	**LargeW2V**	**CNet-NB**
Best Term(s)	["dance-jump"]	["jump", "jumps"]	["tripudiation"]
Dance-Twirl Avg IoU Score	0.881	0.877	0.885

improving upon the test categories proved quite challenging, as base LangSAM proved to already provide higher accuracy masks, especially in the case of `dance-twirl`, which already had an IoU baseline of 0.879 via table 1.

Additionally, the videos we compared may not have been similar enough for a fair comparison, as we solely used videos from the DAVIS dataset for both the generation of each best combination of terms and for testing of these particular combinations. If we had more time for the project, finding a different dataset with more similar, but not identical videos (i.e. a video at two different angles), could prove to be more effective.

6. Conclusion & Future Work

Through our results, we found that our method, WAILA, does improve segmentation performance on videos which LangSAM did poorly on– by a substantial margin on average with a 39.7% increase. Additionally, increasing the number of terms used for optimization and the sample size of frames had a marginal increase on performance. In general, we found that CNet-NB is more than twice as likely to result in the highest IoU score compared to its Word2Vec counterpart, which we believe is due to its methodology of keeping a more complete semantic representation of its pretrained word vectors. However, we noticed shortcomings in generalizing for unseen videos by way of our scooter-black and dance-twirl examples. Thus, this same specificity which improves performance can also have a marginal effect if not a detrimental one.

In future work, we would like to see this same process applied with a higher number of ambiguous label categories, and more image samples over the dataset, if not more data in general. Additionally, we would like to see a branching algorithm for the similar word generation so rather than focusing on only the initial prompt and its related words, each related word could generate its own related terms in a recursive fashion. Furthermore, we would like to see if this same method improves performance for other segmentation models with different architectures.

Overall, WAILA is promising and uses the state-of-the-art language models of a decade ago to improve the performance of a segmentation model representing the state-of-the-art of today which is a surprising and novel result. Although generalized usage is limited, there are many ways to explore and improve upon this concept for future research.

7. Contributions & Acknowledgments

Justin worked on roughly half of the milestone (even though the milestone isn't at all related to this project). He also worked on all of the code scripts in the WAILA folder, most notably `waila.py` and `waila_test.py`. Justin ran the experiments for WAILA, for scooter-black and for dance-twirl. Justin also worked on roughly half of the final report, specifically the methods, mask screenshot examples, and qualitative results, though Walter and Justin worked on all of the sections together.

Walter worked on the other half of the milestone specifically on the literature review and method portions. He implemented all the word generation code scripts, done in .pynb files, for Word2Vec and the CNet-NB generating JSONS for all of the similar words for each category in DAVIS. Additionally, he wrote the introduction, related work, quantitative experiments section 5.2, quantitative word generation analysis 5.3.2, and the conclusion and future work.

The LangSAM GitHub repository we used is reference #2 on the reference list.

References

[1] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation, 2022.

[2] L. Medeiros. Language segment-anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2024.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[5] R. Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2), Feb. 2009.

[6] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in*

Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[7] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[9] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[10] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[11] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018.

[12] R. Speer and J. Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, page 85–89. Association for Computational Linguistics, 2017.