

# Efficient Pre-training for Localized Instruction Generation of Videos

Anil Batra<sup>1</sup>, Davide Moltisanti<sup>2</sup>, Laura Sevilla-Lara<sup>1</sup>,  
Marcus Rohrbach<sup>3</sup>, Frank Keller<sup>1</sup>

<sup>1</sup>University of Edinburgh, <sup>2</sup>University of Bath, <sup>3</sup>TU Darmstadt



THE UNIVERSITY *of* EDINBURGH  
UKRI Centre for Doctoral Training  
in Natural Language Processing



UNIVERSITY OF  
**BATH**

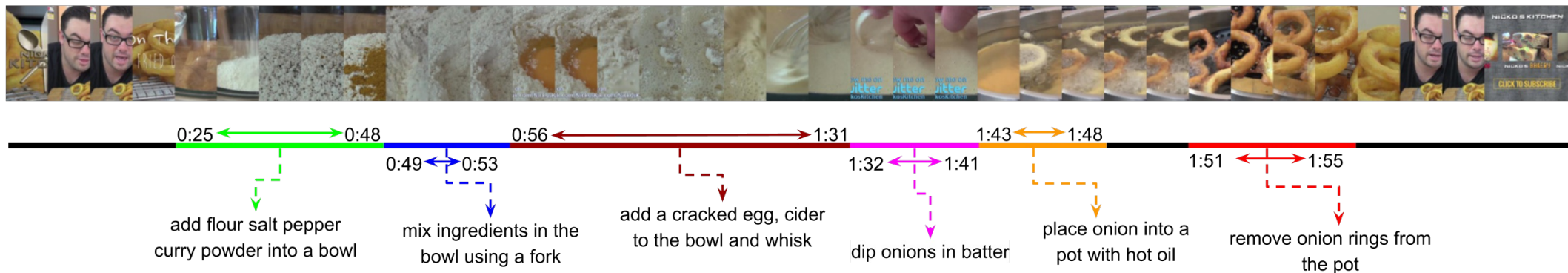


TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

# LOCALIZED INSTRUCTION GENERATION

## Goal

- ▶ Generate temporally localized instruction sequence in the procedural video.
  - ◆ Predicting temporal boundaries of each step.
  - ◆ Generating detailed textual descriptions of each step.



Overview

Approach


Results

Conclusion

# LOCALIZED INSTRUCTION GENERATION

## Challenges

- Prior pre-training approaches [1] utilize raw and noisy ASR text as pseudo labels, requiring large pre-training datasets.



[0 - 6.2] Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

[108.5 - 116.5] So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling the cookie crumbs.

[128.5 - 136.5] And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

[143.5 - 155.5] And what I did is that I placed the heavy cream into the microwave for a ... So this could melt the chocolate away.

[218.5 - 227.5] I send you lots of, . . . and see you next week, hopefully, with another of my vids

*Standard Approach: use raw transcripts*



[0 - 6.2] Hi everyone. Welcome back to another [..]. Today I'm very excited to share with you [..].

[128.5 - 136.5] [..] Just remember to store any leftover truffles in a tightly [..]



**Pre-Training**

1. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning.

Overview

Approach


Results

Conclusion

# LOCALIZED INSTRUCTION GENERATION

## Challenges

- ▶ Prior pre-training approaches [1] utilize raw and noisy ASR text as pseudo labels, requiring large pre-training datasets.
- ▶ Stylistic differences between ASR text and human written instructions introduce a domain gap.
- ▶ Lack of ASR text at test time e.g. Tasty



[0 - 6.2] Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

[108.5 - 116.5] So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling the cookie crumbs.

[128.5 - 136.5] And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

[143.5 - 155.5] And what I did is that I placed the heavy cream into the microwave for a ... So this could melt the chocolate away.

[218.5 - 227.5] I send you lots of, . . . and see you next week, hopefully, with another of my vids

CHURRO ICECREAM SANDWICH

Style Difference

Mix cookie crumbs with cheese


Shape mixture into balls and dip into chocolate

Melt the chocolate in microwave.

# LOCALIZED INSTRUCTION GENERATION

## Contributions

- ▶ A new framework *Sieve & Swap* to remove the irrelevant ASR segments and bridge the stylistic domain gap.



[0 - 6.2] Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

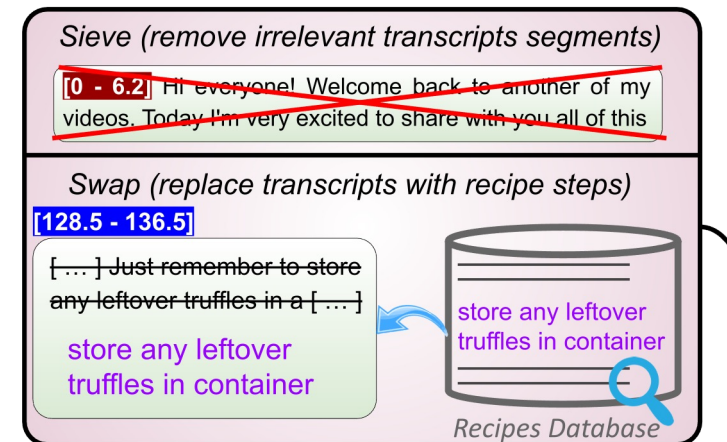
[108.5 - 116.5] So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling the cookie crumbs.

[128.5 - 136.5] And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

[143.5 - 155.5] And what I did is that I placed the heavy cream into the microwave for a ... So this could melt the chocolate away.

[218.5 - 227.5] I send you lots of, . . . and see you next week, hopefully, with another of my vids

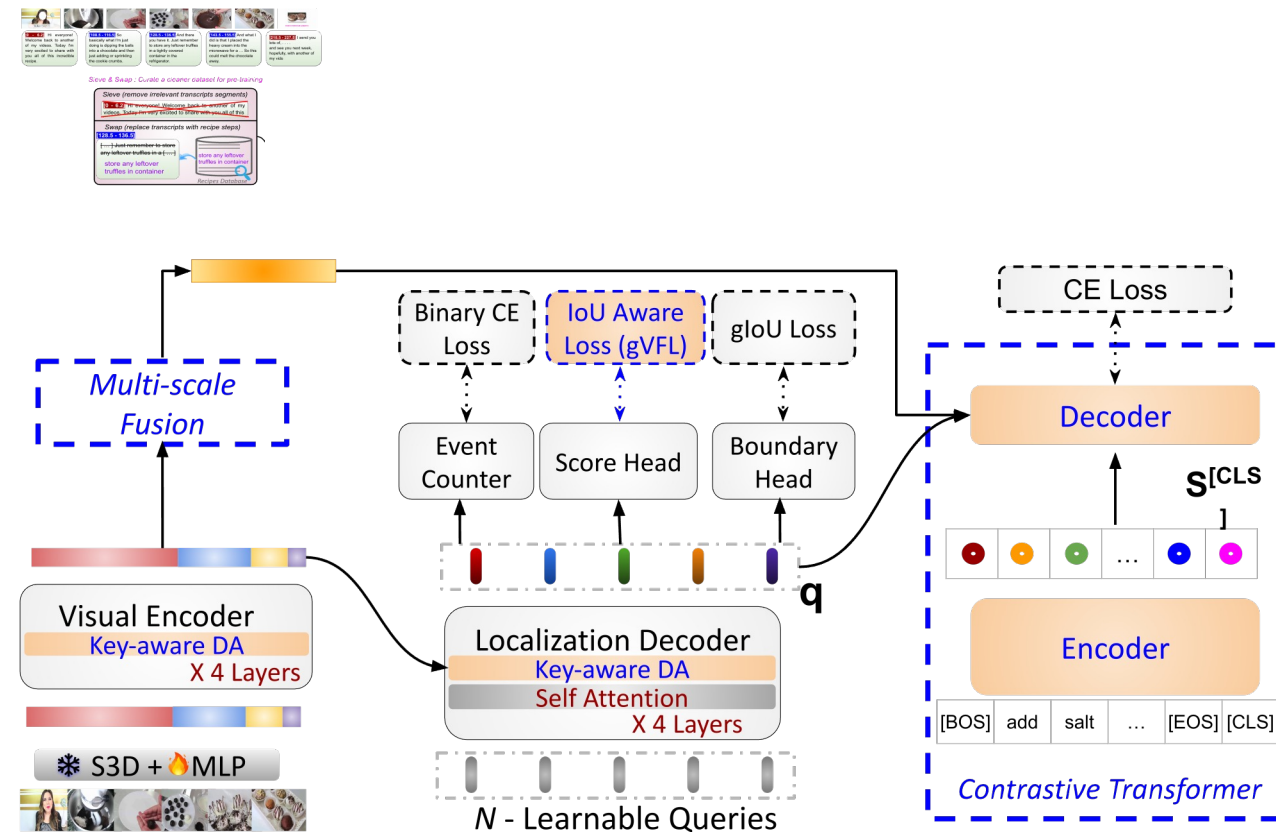
*Sieve & Swap : Curate a cleaner dataset for pre-training*



# LOCALIZED INSTRUCTION GENERATION

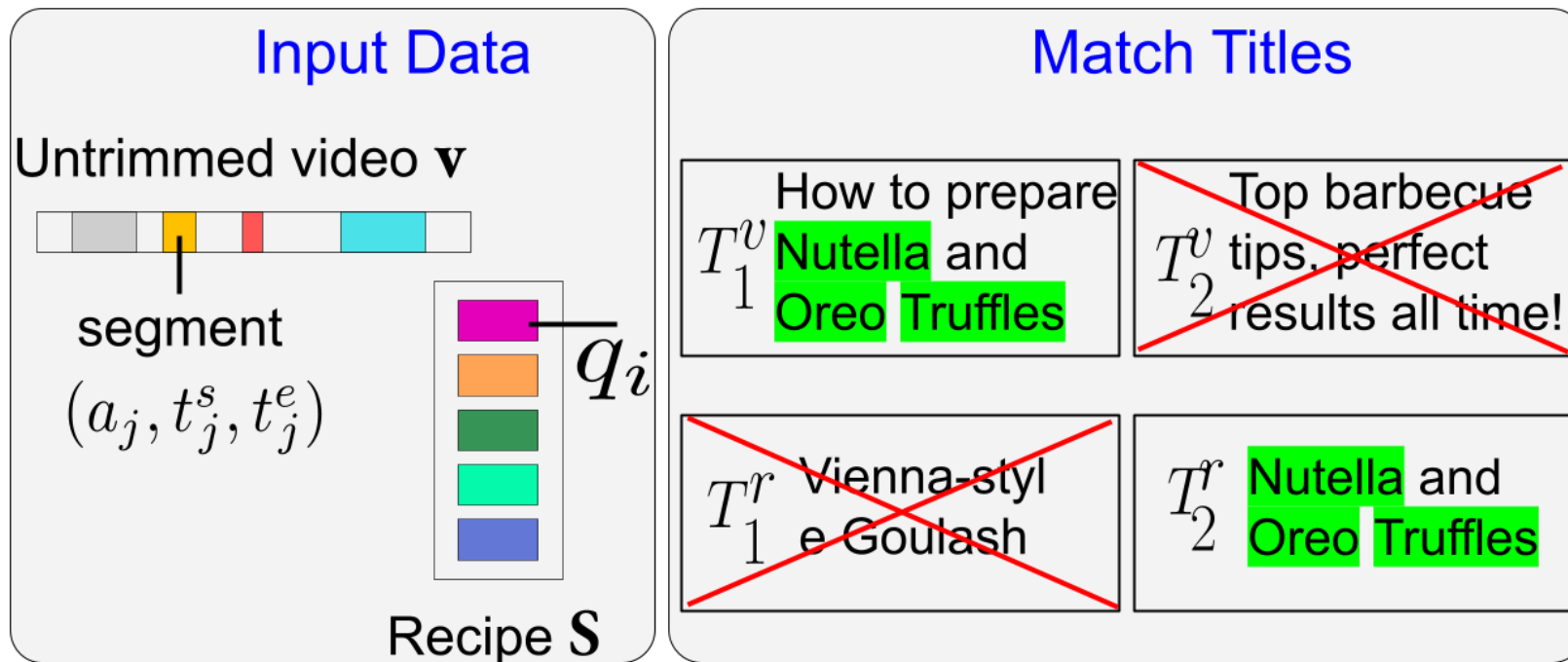
## Contributions

- ▶ A new framework *Sieve & Swap* to remove the irrelevant ASR segments and bridge the stylistic domain gap.
- ▶ Improved Procedure Transformer (ProcX) to focus on video modality, achieving SoTA on YouCook2.

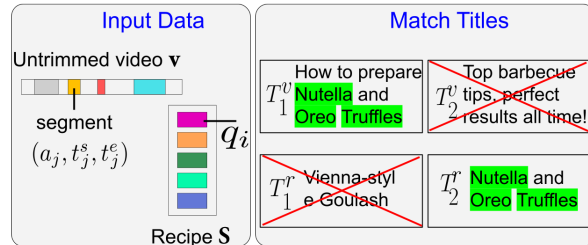




# Sieve Raw Datasets

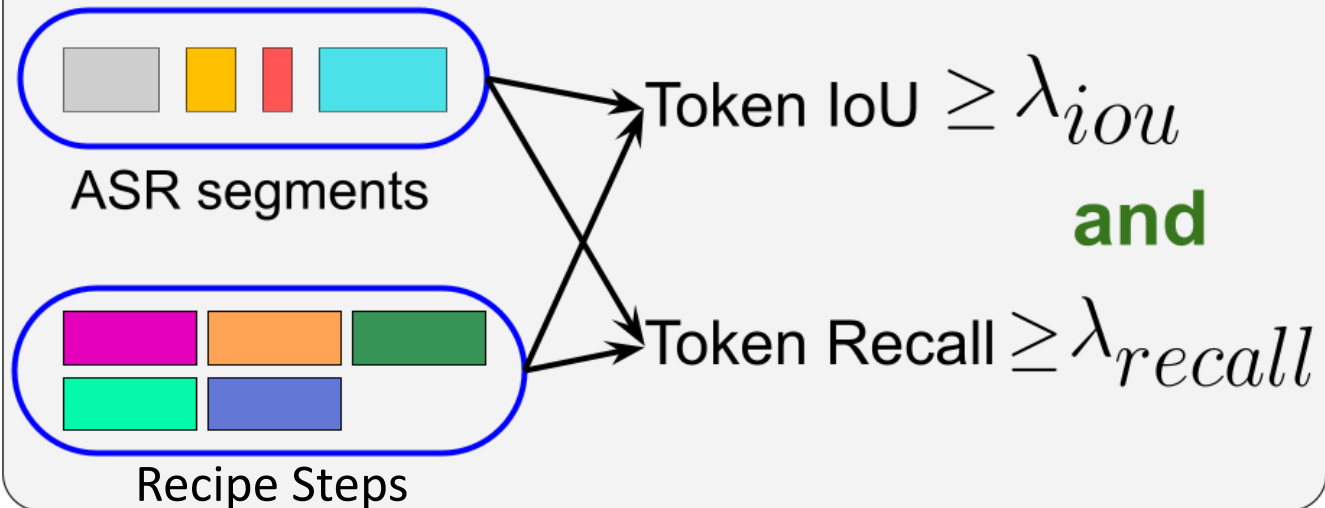


# Sieve Raw Datasets



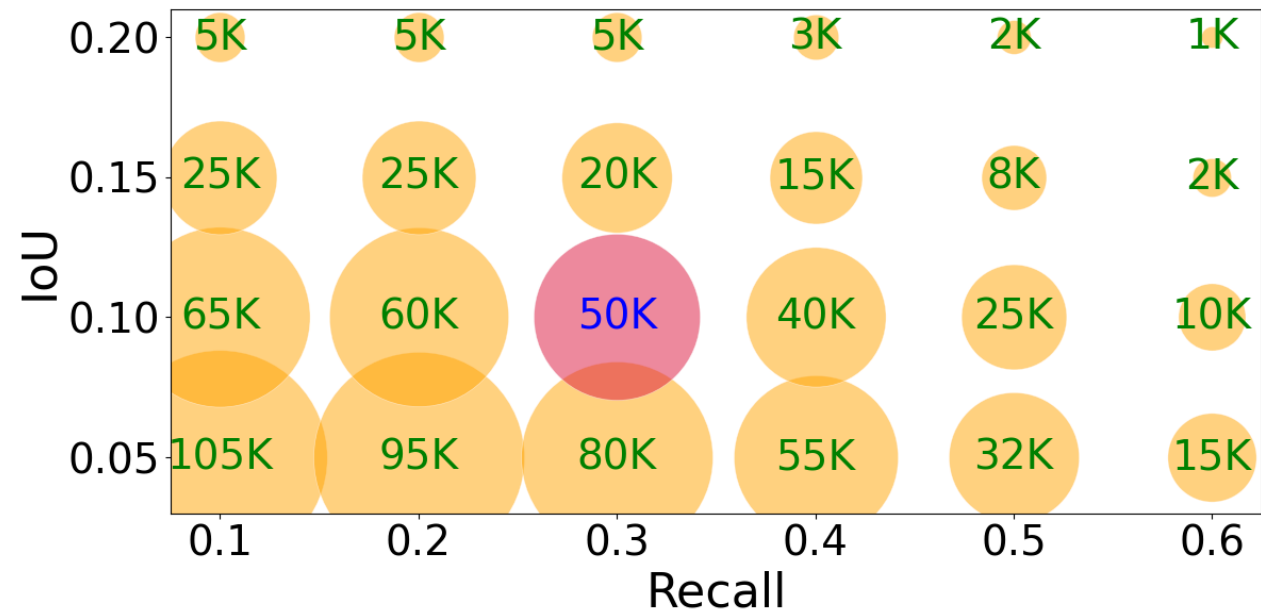
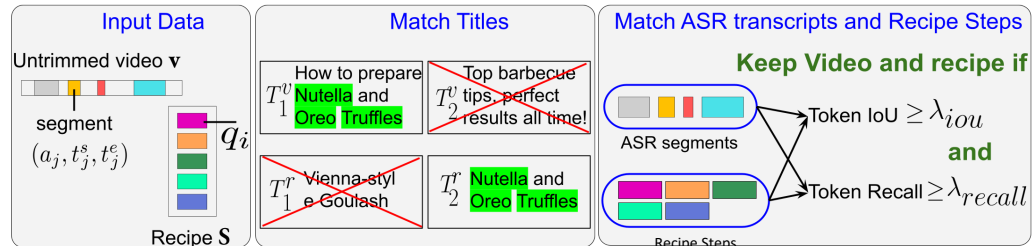
## Match ASR transcripts and Recipe Steps

**Keep Video and recipe if**





# Sieve Raw Datasets



# Sieve & Swap – ASR Segments

## ASR Segments

**[0.0 - 6.2]** Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

● - - - -

**[108.5 - 116.5]** So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling some of the cookie crumbs.

**[128.5 - 136.5]** And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

**[143.5 - 151.5]** [ .. ] I placed the heavy cream into microwave for a couple of seconds. So this could melt the chocolate away and it has to look [ .. ]

● - - - -

**[205.5 - 218.5]** [ .. ] If you did, please give it a thumbs up, comment down below if you liked it and remember to subscribe to my channel ...

# Sieve - & - Swap

## ASR Segments

**[0.0 - 6.2]** Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

**[108.5 - 116.5]** So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling some of the cookie crumbs.

**[128.5 - 136.5]** And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

**[143.5 - 151.5]** [ .. ] I placed the heavy cream into microwave for a couple of seconds. So this could melt the chocolate away and it has to look [ .. ]

**[205.5 - 218.5]** [ .. ] If you did, please give it a thumbs up, comment down below if you liked it and remember to subscribe to my channel ...

sieve

embedding space

sieve

Overview

Approach

Results

Conclusion

# Sieve - & - Swap

## ASR Segments

[0.0 - 6.2] Hi everyone! Welcome back to another of my videos. Today I'm very excited to share with you all of this incredible recipe.

[108.5 - 116.5] So basically what I'm just doing is dipping the balls into a chocolate and then just adding or sprinkling some of the cookie crumbs.

[128.5 - 136.5] And there you have it. Just remember to store any leftover truffles in a tightly covered container in the refrigerator.

[143.5 - 151.5] [ .. ] I placed the heavy cream into microwave for a couple of seconds. So this could melt the chocolate away and it has to look [ .. ]

[205.5 - 218.5] [ .. ] If you did, please give it a thumbs up, comment down below if you liked it and remember to subscribe to my channel ...

sieve

swap

swap

swap

sieve

[108.5 - 116.5] melt your chocolate and dip the cookie balls into the chocolate , tapping off the excess chocolate.

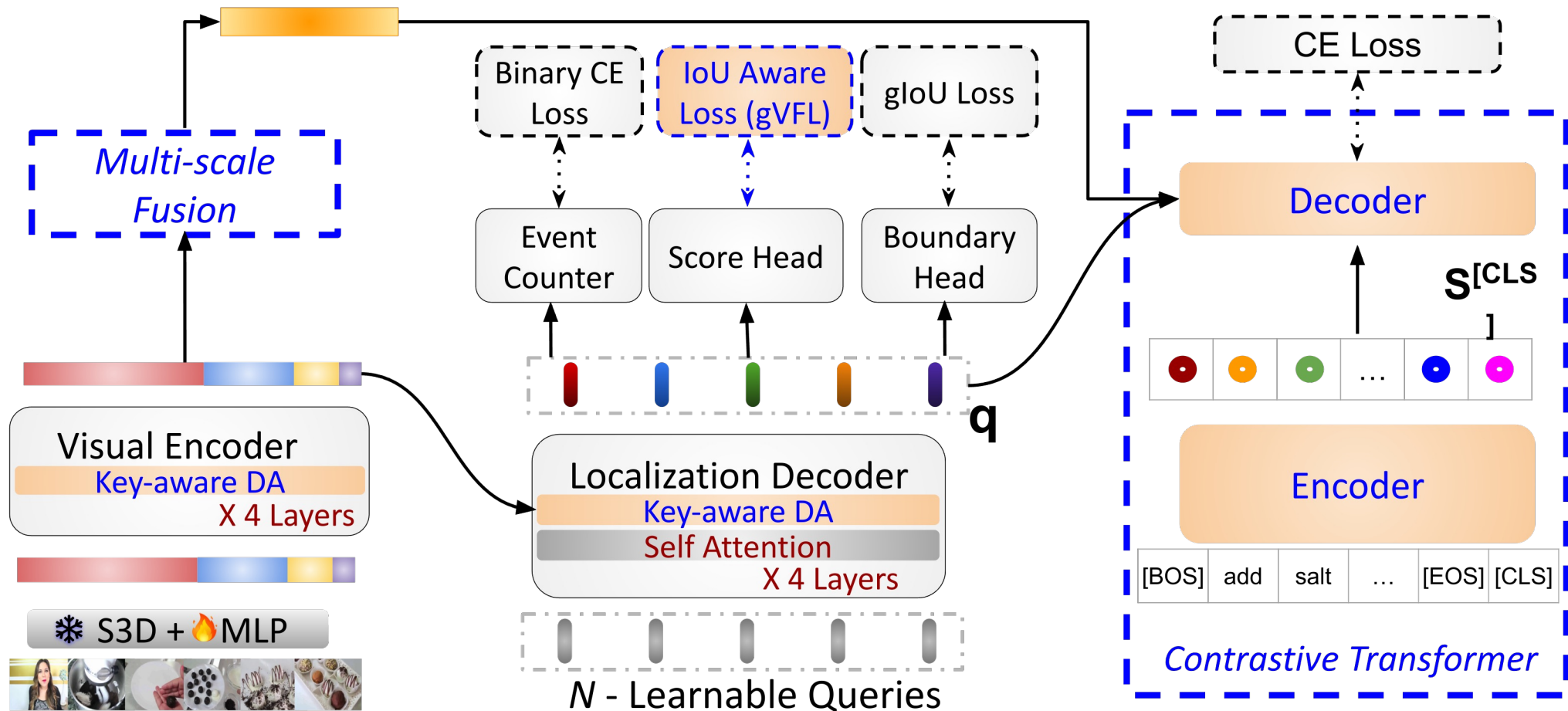
[128.5 - 136.5] store any leftover truffles in tightly covered container in refrigerator.

[143.5 - 151.5] place chocolate in a heatproof bowl , and pour the heavy cream over the chocolate

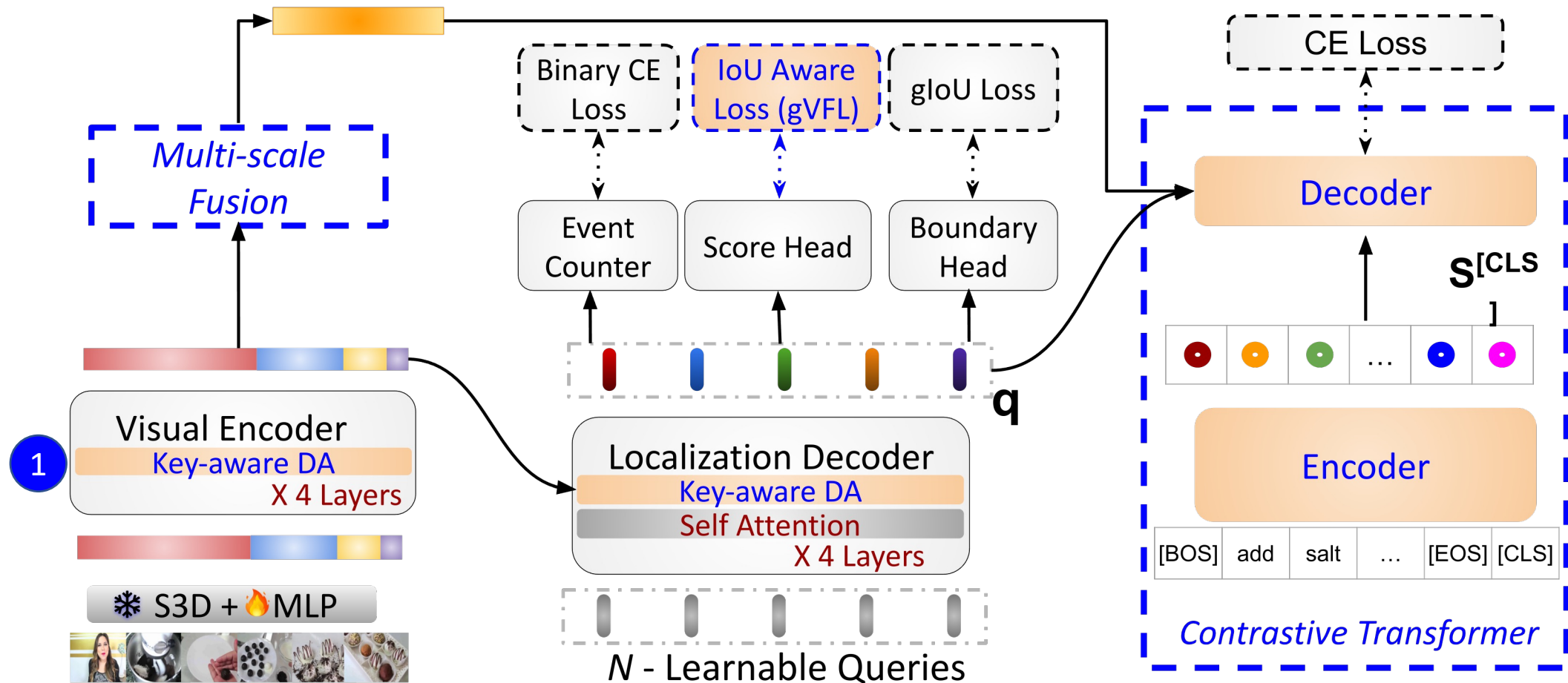
embedding space

**Retrieved recipe steps**

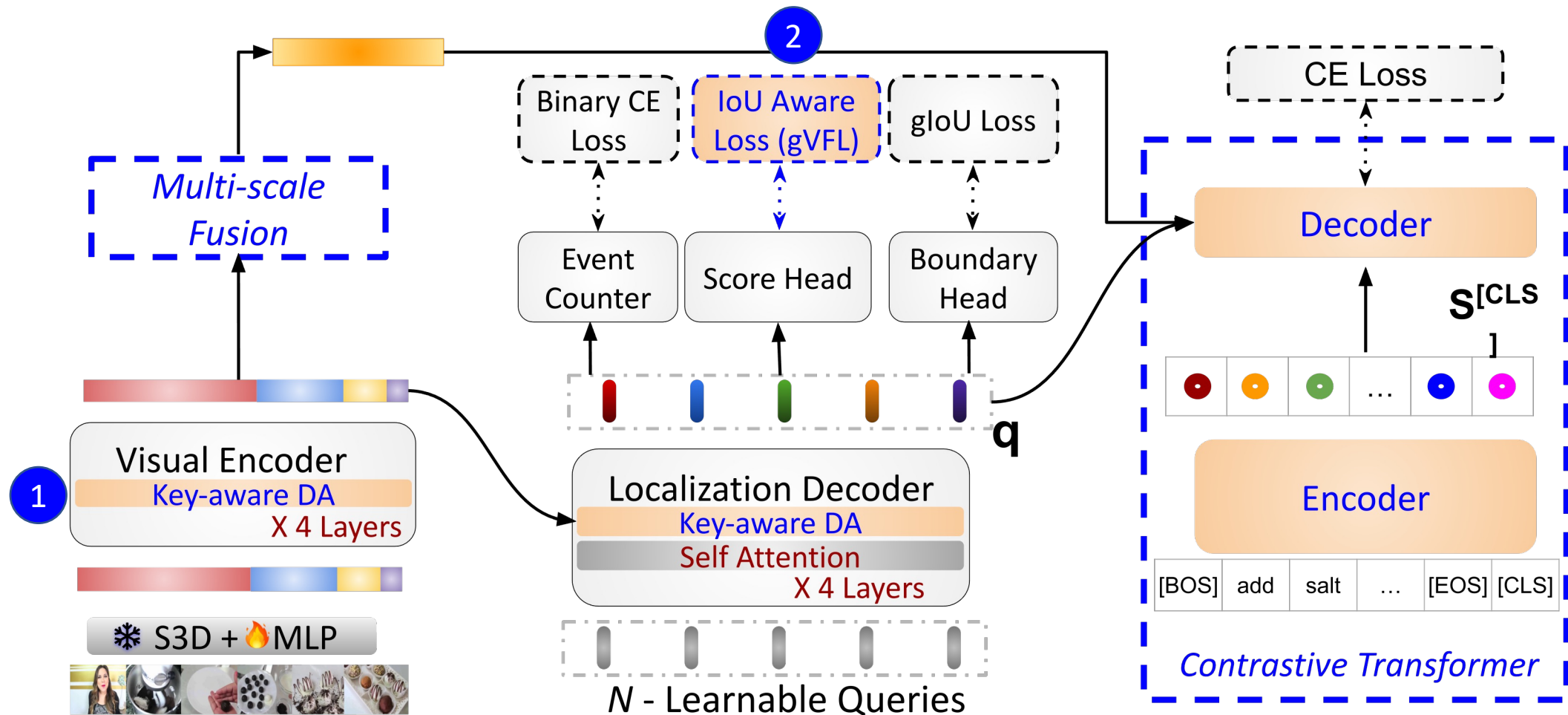
# Procedure Transformer (ProcX)



# Procedure Transformer (ProcX)

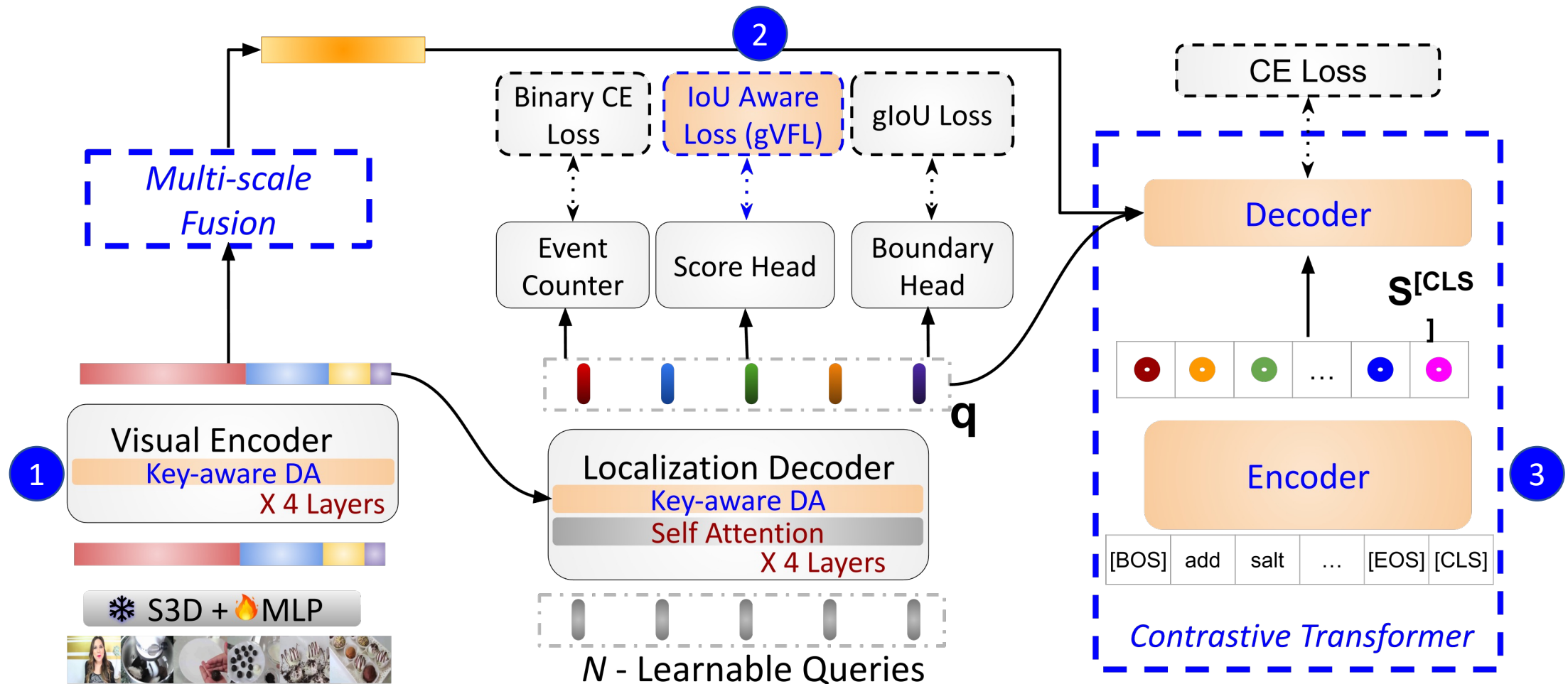


# Procedure Transformer (ProcX)

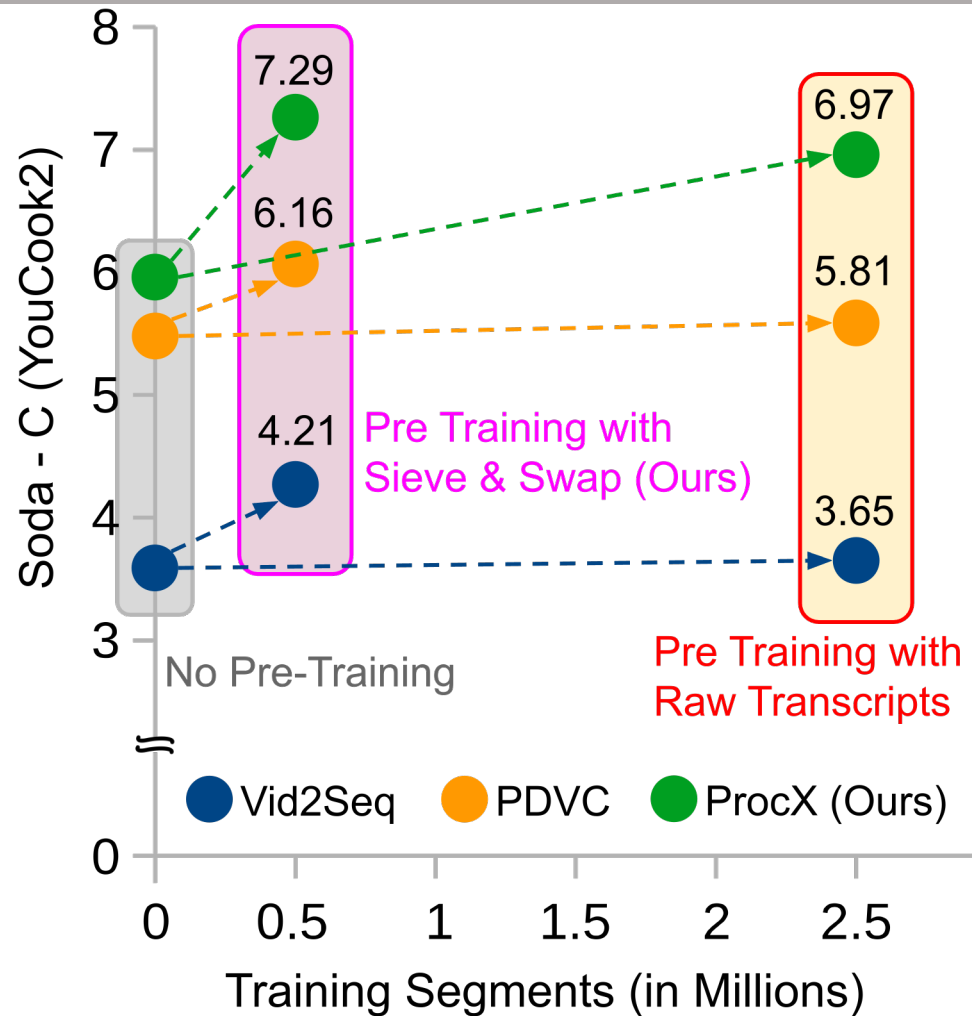




# Procedure Transformer (ProcX)



# State-of-the-Art Comparison



Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. CVPR 2023  
PDVC: End-to-end dense video captioning with parallel decoding. ICCV 2021

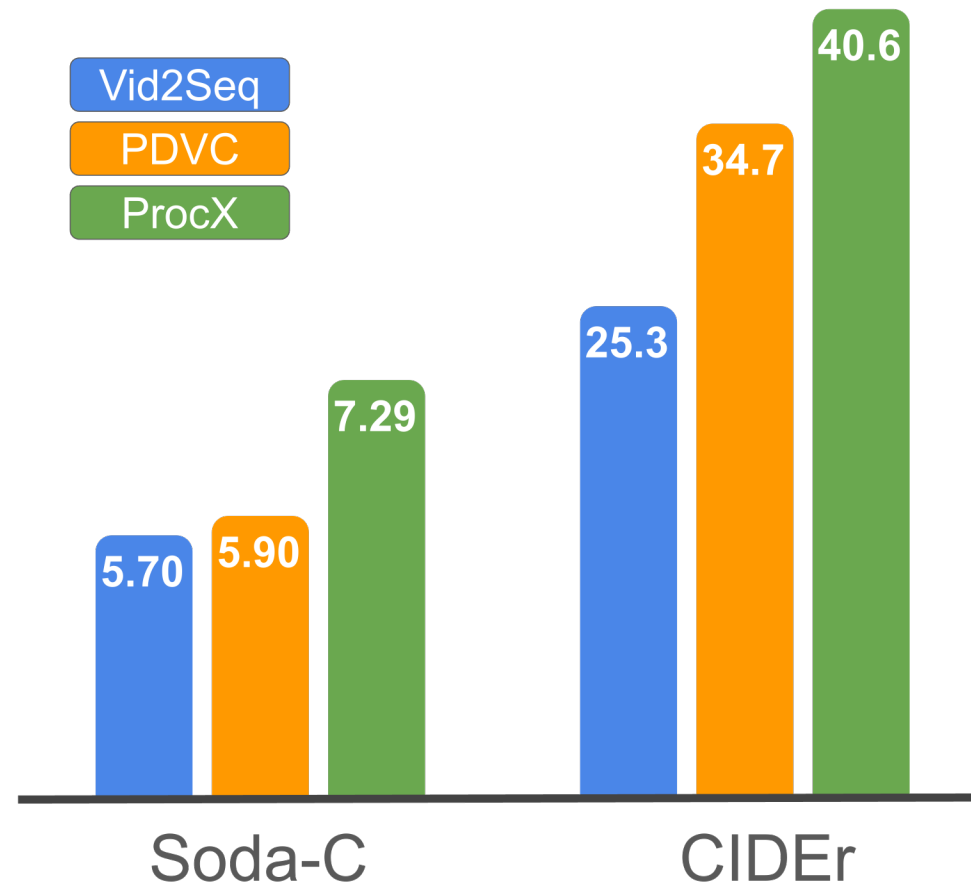
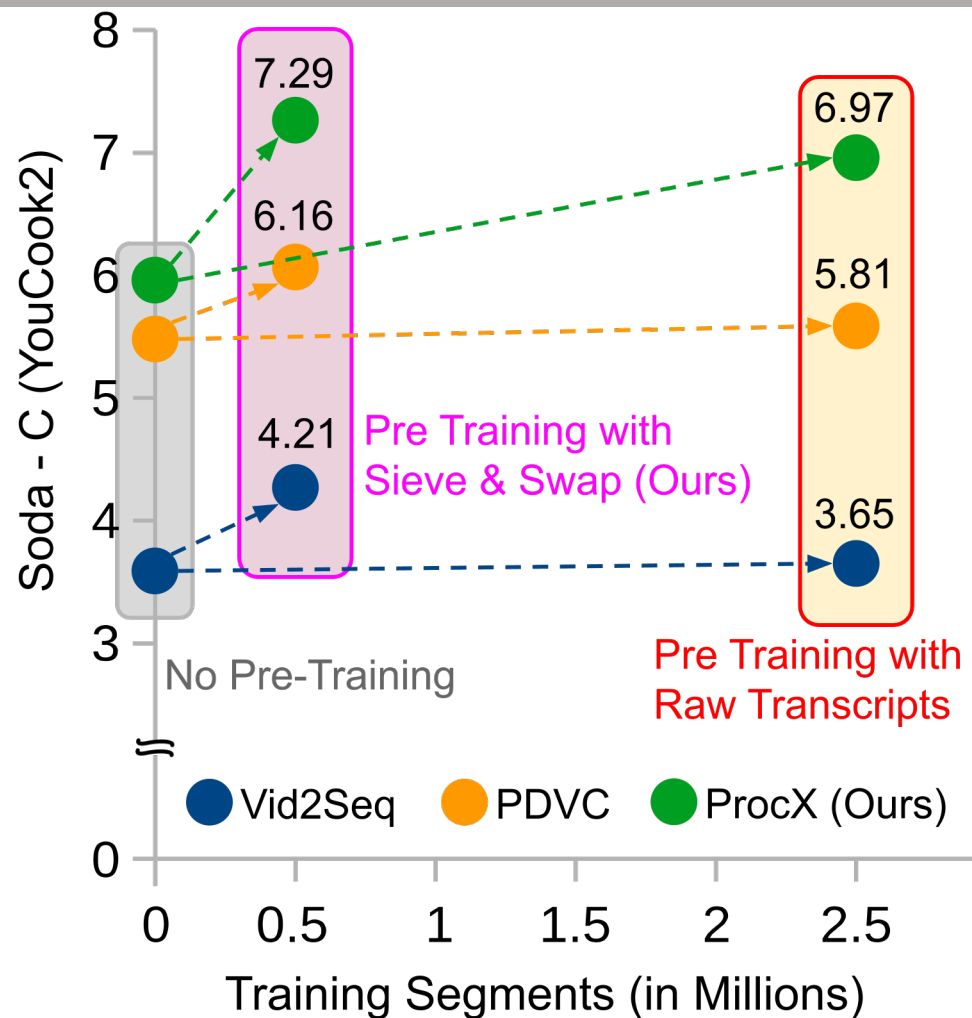
Overview

Approach

Results

Conclusion

# State-of-the-Art Comparison



Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. CVPR 2023  
PDVC: End-to-end dense video captioning with parallel decoding. ICCV 2021

Overview

Approach

Results

Conclusion

# Conclusion

- Simple Embedding similarity-based approach can remove the irrelevant ASR text.
- Bridging language stylistic difference plays a significant role.
- Reduce the noise in pre-training dataset and still achieve competitive results.