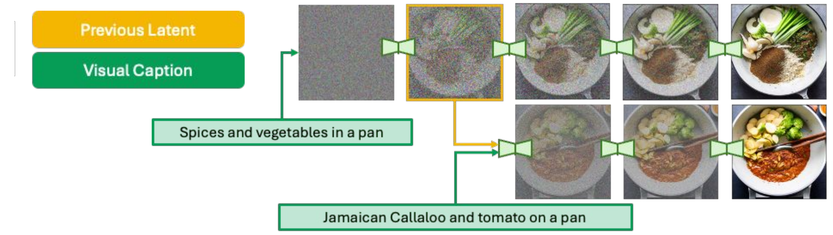
Contrastive Sequential-Diffusion Learning: Non-linear and Multi-Scene Instructional Video Synthesis

Vasco Ramos[1], Yonatan Bitton[2], Michal Yarom[2], Idan Szpektor[2], Joao Magalhaes[1]
[1]NOVA LINCS [2]Google Research

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

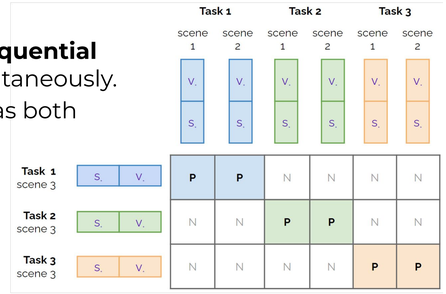
Coherent Generation

$$\mathcal{L}_{CoSeD} = \mathbb{E}_{z_T^n, s_n, \epsilon, t=T}\left[\|\epsilon - \epsilon_\theta(z_T^n, c_n)\|_2^2\right]$$

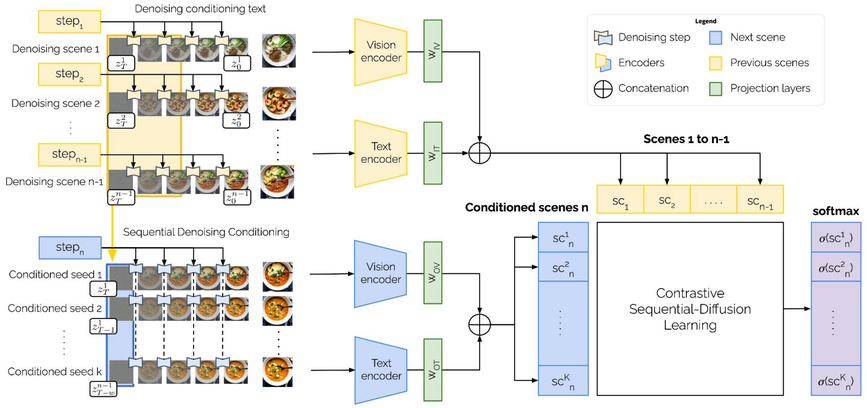


Training

- Learn relationships between **sequential steps** across various tasks simultaneously.
- Ground truth information acts as both **positive and negative pairs**.
- Enable the model to leverage **information** from **multiple tasks** for improved learning.



Architecture



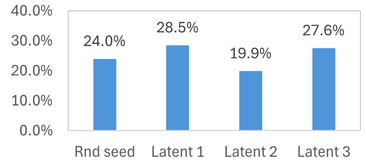
Specificalized in **cooking** recipes and **DIY** task videos.

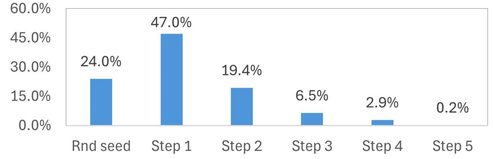
Contrastive selection ensures more **coherent sequence generation**.

Improved Seed Selection Method for **better Visual** and **Semantic Coherence.**

Achieves up to **90%** preference in **human evaluation** compared to baselines.

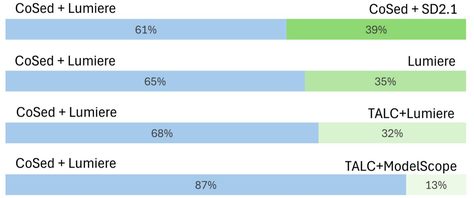
Non-Linear Behaviour





Human Evaluation

Methods	Video Length	Semantic Consist.	Sequence Consist.
CoSeD + Lumiere	20.8 s	**85.0**	**74.2**
CoSeD + SVD	14.9 s	78.3	69.2
TALC + ModelScope	7.4 s	38.3	50.8
TALC + Lumiere	5.0 s	30.0	50.8
SD + SVD	14.9 s	80.0	66.3
Lumiere	20.8 s	81.7	72.9



Qualitative Results

1. Heat the Coconut oil, then add the Onion,Garlic and Scallion.
2. Add the Shrimp, and cook.
3. Turn the heat up and add the Jamaican Callaloo and Tomato.
4. After 10 minutes, taste for salt.

This work was partially supported by a Google Research Gift and by the FCT project NOVA LINCS Ref. (UIDB/04516/2020).