Procedural Video Generation for Instructional Videos

Advised By: Prof. Kristen Grauman

Presented By: Agrim Jain, Shabari Nair, Krishanu Saini

Motivation

- How-to cooking videos require **action and context consistency** across ingredients and utensils.
- **Learning from text alone is insufficient**; cooking is inherently procedural and visual.
- Existing video generation models produce **short clips**, struggling with long, consistent video generation.
- Current approaches often **retrieve clips or generate individual images**, lacking temporal and procedural consistency. (Stitch-a-recipe/Stitch-a-demo, Illustrated Instructions)
- Need for **consistent long-form video generation** for recipe instructions.

Main Idea

- Supervised Fine-tuning of Phantom-WAN using Diffusion Pipe
- Using VLM (here GPT 4o) to enhance text prompts online
- Using VLM prompt engineering on top of the fine-tuned model.
- Using a curated metric as a reward function for an RL based model to gain feedback on the generated video (motivated from Identity-GRPO)

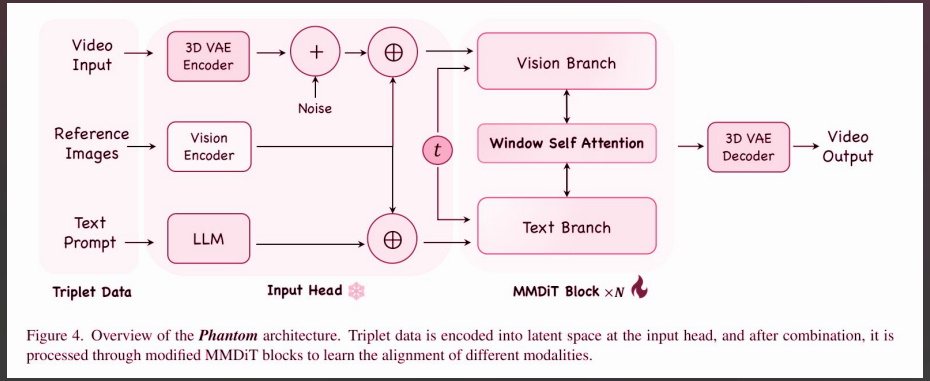


Figure 4. Overview of the ***Phantom*** architecture. Triplet data is encoded into latent space at the input head, and after combination, it is processed through modified MMDiT blocks to learn the alignment of different modalities.

What is New?

- Currently, no end-to-end video generation models for instructional cooking videos.
- The models in this domain are primarily retrieval based (eg: Stitch-a-recipe).
- Phantom-WAN is pre-trained primarily for subject consistency and often does not take background (persistent objects) into consideration.
- Identity-GRPO is trained for consistency on human subjects and has a reward function defined accordingly.

Metrics

Identity-GPRO

Human evaluation dataset.

Consistency Based Reward Model

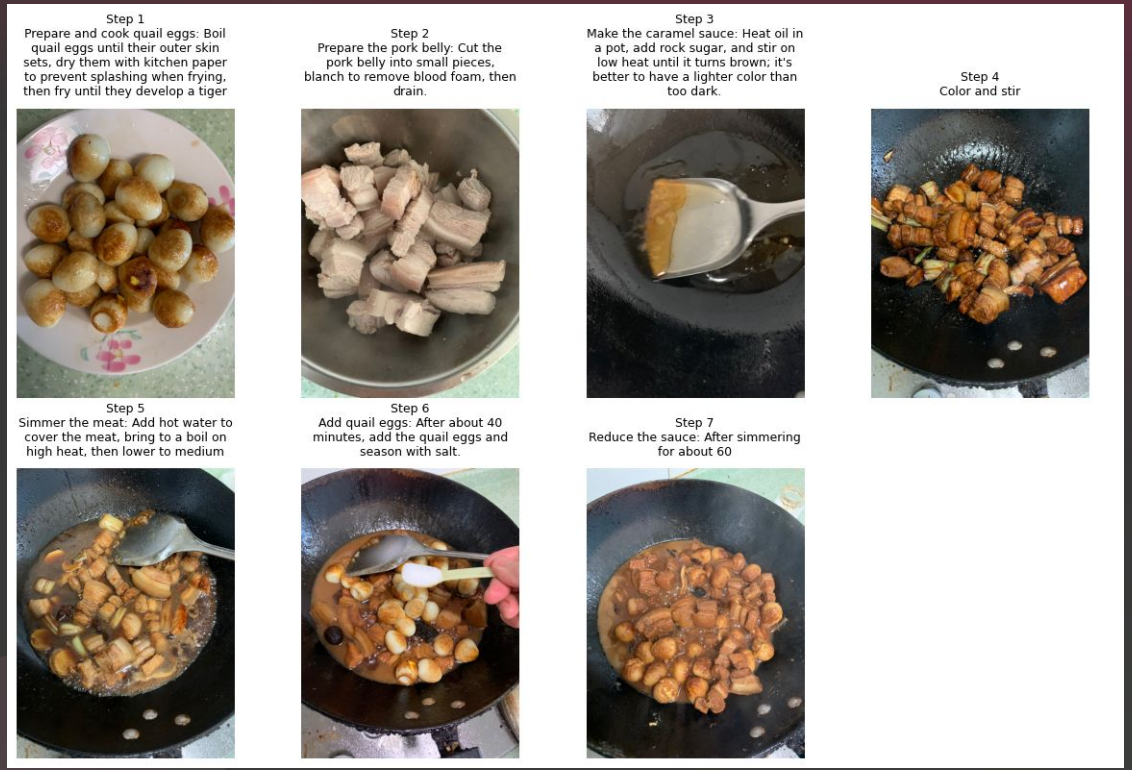
Alignment metric that mimics human scoring.

State Dict
Carrot: Fresh carrots are being cut

State Dict
Carrot: Cut carrots are being Boiled

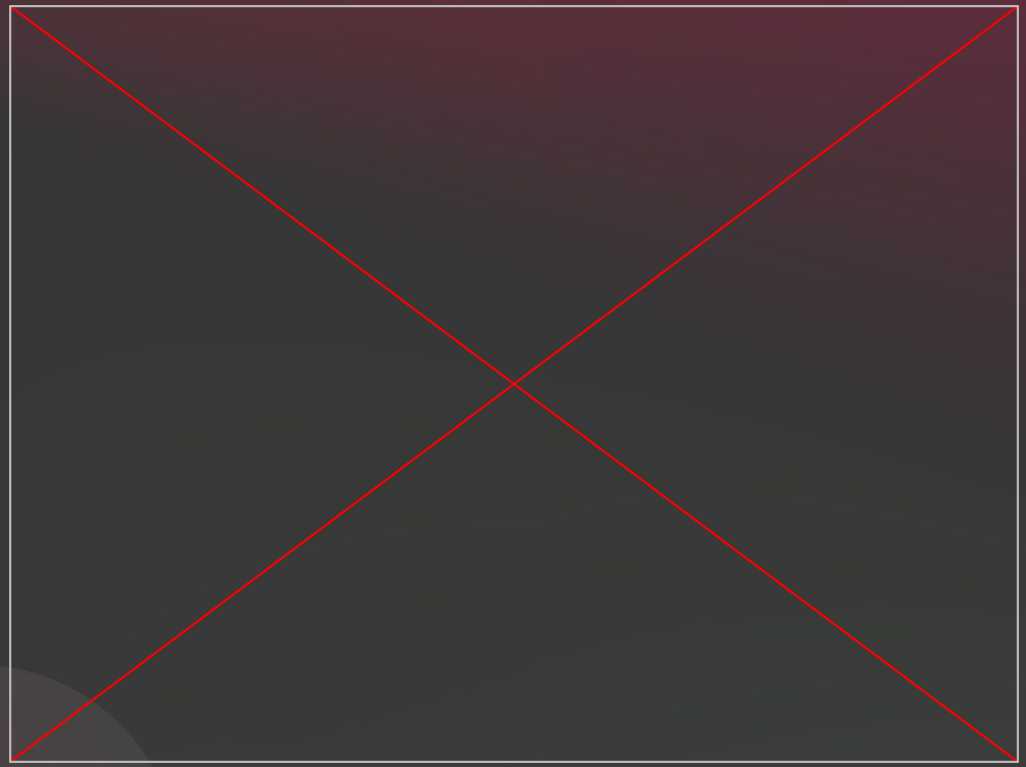
{
"global_id": 0,
"object_name": "carrot",
"is_consistent": true,
"reasoning": "Consistent across all clips. Clip 1: First clip - initial state."Clip 2: Second clip shows cut carrots being boiled which is consistent.
} ...

Recipe Goal: Braised Pork with Quail Eggs



Ground Truth

Baseline



Consistency Based Reward Model
40% Lot of new objects are generated.

FineTuned Version

Consistency Based Reward Model
75%

Finetuned + Memory Module

Consistency Based Reward Model
77%

Finetuned + Memory Module

Consistency Based Reward Model

77%

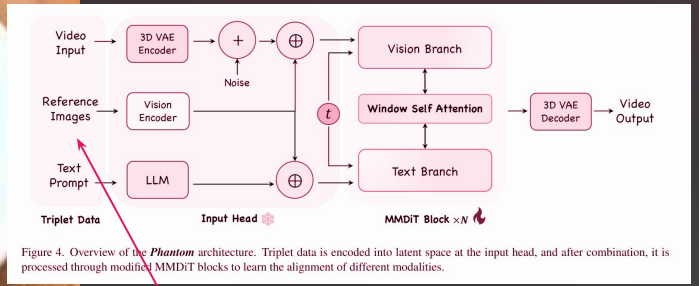


Figure 4. Overview of the ***Phantom*** architecture. Triplet data is encoded into latent space at the input head, and after combination, it is processed through modified MMDiT blocks to learn the alignment of different modalities.

Object bank

Results

	shot-consistency metric (TransNet-V2)	Dino-L2 Distance	Consistency Based Reward Model
Baseline	0.86	0.18	66
Finetuned	0.46	0.12	75
Finetuned + Memory	0.44	0.13	77

Future Work

- Larger scale training.
- The RL based model is yet to be trained. Currently constrained by compute and time constraints
- Using our methods in an online detouring scenario, along the lines of ‘VidDetours’
- Prompt engineering the inputs to VLMs for more deterministic and accurate behaviour.

THANK YOU!