This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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, NOVA School of Science and Technology, Portugal
[2]Google Research
jmag@fct.unl.pt, szpektor@google.com

Abstract

Generated video scenes for action-centric sequence descriptions, such as recipe instructions and do-it-yourself projects, often include non-linear patterns, where the next video may need to be visually consistent not with the immediately preceding video but with earlier ones. Current multi-scene video synthesis approaches fail to meet these consistency requirements. To address this, we propose a contrastive sequential video diffusion method that selects the most suitable previously generated scene to guide and condition the denoising process of the next scene. The result is a multi-scene video that is grounded in the scene descriptions and coherent w.r.t. the scenes that require visual consistency. Experiments with action-centered data from the real world demonstrate the practicality and improved consistency of our model compared to previous work. Code and examples available at https://github.com/novasearch/CoSeD

1. Introduction

When people perform tasks involving numerous intricate steps, complementing textual instructions with visual illustrations enhances the user experience [11, 24]. For this reason, various platforms and tools provide multi-scene videos to convey instructional content, such as recipe instructions and do-it-yourself (DIY) projects [15].

State-of-the-art video synthesis methods demonstrate remarkable performance in generating single-scene videos [3, 4, 12, 25, 28]. Yet, only a few works address multi-scene video generation [16, 17, 30]. These methods focus on domains where a single character is central to all scenes, achieving coherence by reusing and combining visual elements across frames. However, multi-scene instructional video synthesis raises a number of challenges. First, the input is a *strict sequence of actions*, for which it is necessary to *generate the full sequence of videos*. A model should not generate just the last scene, like GILL [14], and one cannot provide the topic

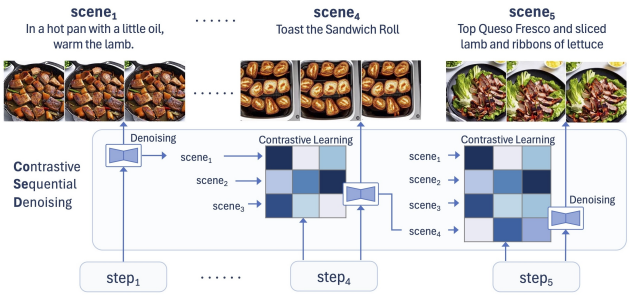

Figure 1. CoSeD is grounded on an input sequence of step actions to synthesize non-linear, multi-scene instructional videos.

and let the model generate a random sequence of text-video pairs similarly to VideoDrafter [17]. Second, similar to story generation [9, 18, 20, 22], the generative model needs to *determine which previous step in the sequence to use as the basis for grounding each new scene*. Third, while existing methods are focused on video generation where a single (typically human) character is the center of all scenes [17], instructional videos typically incorporate multiple objects instead of central characters. Hence, we argue that multi-scene instructional video synthesis requires an approach that is sequence-grounded by design, (see Figure 1).

To this end, we propose CoSeD (**Co**ntrastive **Se**quential **D**iffusion learning), a novel approach to instructional video generation. Our method generates candidate images for each step based on the textual description and latent information from previous steps. We then use a contrastive selection approach to choose the best image by evaluating it against prior step descriptions and images. Finally, we use these images to produce a video for each step, ensuring an accurate and consistent representation of the entire task sequence.

CoSeD was able to generate coherent video sequences across diverse instructional content, maintaining high fidelity and relevance in aligning language with vision. CoSeD showed a 20% improvement in human evaluations compared to existing multi-scene methods and was preferred

68% of the time in side-by-side comparisons. Additionally, the compact size of our model enables efficient training and fine-tuning. To summarize our contributions:

- To the best of our knowledge, we are the first to address multi-scene instructional video generation.
- We introduce contrastive diffusion learning over latents sampled from previous generations.
- We contribute to a better understanding of the role of seeds and the conditioning of the reverse diffusion process in prior latent representations.

2. Related Work

Various approaches have been explored to address coherence in image generation. AR-LDM [20] introduces a history-aware autoregressive latent diffusion model that incorporates information from previous steps into the diffusion model's cross-attention mechanism to guide generation. However, achieving the reported results requires intensive training of the entire pipeline for each dataset, which includes 650 million parameters. Make-a-Story [22] incorporates the complete history of intermediate image representations (latent vectors), which may introduce noise and potentially lead to content generation based on less relevant past information. GILL [14] fuses frozen text-only large language models (LLMs) with pre-trained image encoder and decoder models through a mapping network, enabling multimodal capabilities like image retrieval and generation. However, this can break coherence if retrieved images do not align with the context. SEED-LLaMA [10] integrates a visual tokenizer with a multimodal LLM to process text and images, excelling in multi-turn generation, but struggles with maintaining narrative coherence in story generation tasks.

To generate long single-scene videos Blattmann [4] and Yin [29] generate sparse key frames and interpolate intermediary frames recursively to enhance the frame rate. Extending this idea, Stable Video Diffusion [3] creates a large dataset of annotated video clips by filtering out those with low motion or excessive text, resulting in the generation of higher quality videos. In contrast, Lumiere [2] generates the entire video in a single pass, eliminating the need for sparse key frames and interpolation.

Improvements have also been made in generating coherent multi-scene videos. Video Drafter [17] employs brute-force LLM prompting to create distinct scenes and detailed descriptions for each element. Then it generates image templates that are combined with scene descriptions to produce the final video. Similarly, VideoDirectorGPT [16] employs a two-stage process in which GPT-4 [19] expands text prompts into detailed descriptions and ensures visual continuity by generating textual descriptions and entity layouts. Mora [30] introduces a multiagent framework, breaking tasks into subtasks like refining prompts [26], and generating images to create small video segments, which are then assembled [7] into a coherent final video, achieving performance comparable to closed-source models such as SORA [6]. However, it relies on using only the last frame of each video segment to start the generation of the next one, which can be problematic if the current step is not directly related to the previous one. StoryDiffusion [31] maintains coherence across frames using a self-attention mechanism and a module for smooth transitions, ensuring that videos faithfully depict the input prompt. Lastly, TALC [1] enhances text-to-video (T2V) models by improving the temporal alignment between video scenes and text segments, improving visual fidelity and narrative coherence.

Building on these advancements, we develop a method that addresses the challenge of maintaining scene coherence while respecting text descriptions, while keeping the model compact to facilitate fine-tuning across multiple domains.

3. Problem Setting

This section outlines our methodology for generating scene sequences that align with each instruction while preserving continuity with preceding scenes. By enhancing Latent Diffusion Models, we can effectively learn the interdependencies across scenes and guarantee a cohesive visual progression, even when step relations are non-linear.

3.1. Sequential Scene Dependency

Given a set of tasks, $\mathcal{D} = \{T_1, T_2, ...\}$, where each task T_j comprises a sequence of step-by-step text instructions $T_j = \{s_{j_1}, \ldots, s_{j_n}\}$, our goal is to generate the sequence of scenes $V_j = \{v_{j_1}, \ldots, v_{j_n}\}$ that are best aligned with the corresponding step instruction and all previous visual scenes. The result is a multi-scene video that depicts the steps of the task consistently across all scenes. For simplicity, we will omit the task index j from our notation.

In our setting, we depart from the linear dependency assumption used in previous works [30] and acknowledge the possibility of a more complex and non-linear sequential structure [8]. To address this assumption, the model needs to consider not only the current step description and visual scene pair (s_n, v_n), but also the pairs from the previous steps and visual scenes, $\{(s_1, v_1), \ldots, (s_{n-1}, v_{n-1})\}$. This ensures coherence in the visual elements generated, maintaining consistency, and reflecting the progression of the task, even when individual steps are vague or missing details.

3.2. Sequential Multi-Scene Diffusion

Latent Diffusion Models are designed to synthesize one single image or video at a time. Our goal is to move beyond this limitation and propose a sequential-diffusion model that learns how semantic and visual dependencies should exist in

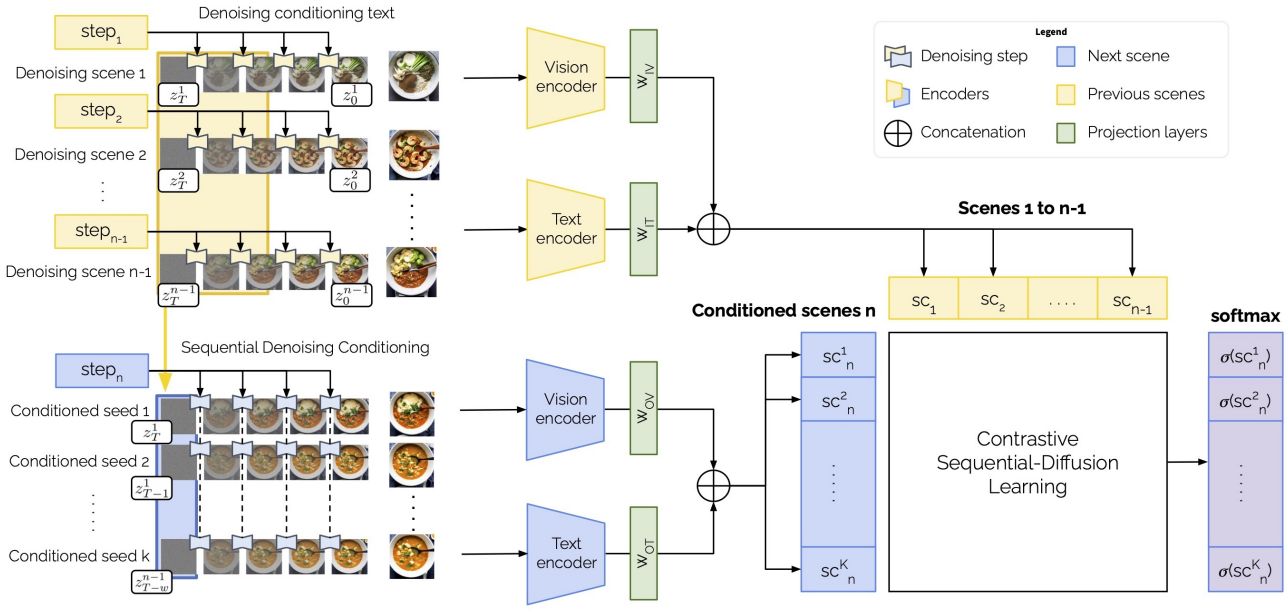

Figure 2. The proposed contrastive denoising diffusion learning architecture. The contrastive learning component captures the temporal relationships between conditioned scenes and preceding scenes, ensuring coherent transitions throughout the video.

a sequence of multiple scenes. Using the Latent Diffusion Models formulation proposed by [23], the independent denoising process for each isolated step s_n of a sequence is the direct application of the model,

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_t^n, s_n, \epsilon, t}\Big[\|\epsilon - \epsilon_\theta(z_t^n, t, \tau_\theta(s_n))\|_2^2\Big] \quad (1)$$

where z_t^n corresponds to the denoising iteration t of the visual scene v_n, hence $z_0^n = E(v_n)$. Formally, we wish to learn a sequence model that iteratively estimates the v_n scene that maximizes the likelihood given the entire sequence of all previous $n-1$ steps. Formally, we have,

$$\sum_{n=2}^{N} p(v_n | s_n, (s_{n-1}, v_{n-1}), \dots (s_1, v_1)), \quad (2)$$

where N is the total number of steps in a given sequence. We propose to ensure sequential consistency through the text conditioning encoder $\tau_\theta(s_i)$ and the visual denoising seed z_T^n. The proposed contrastive sequential diffusion learning, Figure 2, incorporates these two methods and will be discussed in the next section.

4. CoSeD: Contrastive Sequential Diffusion

Our approach aims to find the most accurate image for generating a video that depicts a step of a task. We begin by using a text decoder model to better align the step description with a visual caption/prompt (Section 4.1). Then we generate candidate images based on the information from previous images and the visual caption of the current step (Section 4.2). Next, we use a contrastive selection method to choose the most suitable candidate image for video generation (Section 4.3). This involves encoding both the step description and the visual scene (Section 4.3). Finally, we use the sequential information of the tasks to train our model to accurately select the most coherent image (Section 4.3).

4.1. Sequential Language Conditioning

Following the work of Bordalo et al. [5], we use an LLM to transform the sequence of text descriptions of each step into visual captions. This has been shown to produce text-to-image prompts that are visually richer, leading to better results [5, 13]. Hence, we train a decoder model φ to convert the entire context into one self-contained description,

$$\varphi(s_n | \{s_{n-1}, \dots s_1\}), \quad (3)$$

whose output is used to condition the denoising process on the entire sequence of actions, leading to the loss function

$$\mathcal{L}_{CoSeD} = \mathbb{E}_{z_t^n, s_n, \epsilon, t}\Big[\|\epsilon - \epsilon_\theta(z_t^n, t, c_n)\|_2^2\Big], \quad (4)$$

where $c_n = \tau_\theta(\varphi(s_i | s_{<i}))$ is the conditioning embedding vector that is passed to the cross-attention of the U-Net ϵ_θ.

4.2. Sequential Denoising Conditioning

Previously, we discussed the sequential dependency assumption $(s_n | s_{n-1}, \dots s_1)$ for the input of the denoising process. However, achieving sequential dependency within the denoising process itself poses a challenge. Although

aligning the input description with the desired output enhances the final result, it does not inherently enforce the generation of visually consistent images. This can lead to accurate depictions of steps but without visual coherence. Therefore, guiding the denoising process is essential to ensure visual coherence across sequential outputs.

We propose a contrastive method to select the image that best represents s_n in terms of its description and the preceding scenes. This selective approach strikes a balance between conditioning in all pairs [20], which provides comprehensive information but can be slow and difficult to train, and conditioning on only a single preceding step [5], which offers a faster but less detailed approximation. Our method effectively captures the non-linear nature of the steps in the task that we aim to model. Formally, the denoising iterations follow equation 4, except for the starting iteration T of the reverse diffusion process,

$$\mathcal{L}_{CoSeD} = \mathbb{E}_{z_T^n, s_n, \epsilon, t=T}\left[\|\epsilon - \epsilon_\theta(z_T^n, c_n)\|_2^2\right], \quad (5)$$

where instead of initializing the latent variable z_T^n with a random sample from $z_T \sim \mathcal{N}(\mu, \sigma^2)$, we propose to sample denoised latents from prior steps $s_{<n}$ in the sequence. Formally, for each step n, we consider the set of latents produced in previous denoising iterations of earlier steps, i.e.

$$\{z_T^i, z_{T-1}^i, \ldots, z_{T-w}^i\}^{i\in(1,\ldots,n-1)}, \quad (6)$$

where i indexes all steps from 1 to $(n-1)$ and w is the window size over the first denoising latents of each step. Finally, all candidate visual scene v_n^i are generated with the set of latents. By conditioning the generation of step n on latents from all previous steps in this complex non-linear way, the coherence of the generated sequence is improved.

This method allows for the selection of the most suitable latent representations for each step, ensuring coherence and continuity throughout the entire generation process.

4.3. Multi-Scene Contrastive Selection

Text and Vision Scene Embeddings. To effectively handle the text and visual modalities of a scene $sc_n = (s_n, v_n)$ in a sequence, we encode both modalities using CLIP [21], see Figure 2. Subsequently, the output of each encoder, is linearly projected to reduce its dimension to half the original size. For the projection, we use four distinct weight matrices: W_{IT} and W_{OT} for the text embedding, and W_{IV} and W_{OV} for the visual embedding. W_{IT} and W_{IV} project the embeddings of past scenes, while W_{OT} and W_{OV} project the embeddings of the current scene. The projected embeddings are then concatenated into one single vector.

The resulting embedding projections sc_n^i of all candidate scenes (s_n, v_n^i) and all past scenes $sc_{<n}$ allow us to represent all scenes within a unified embedding space.

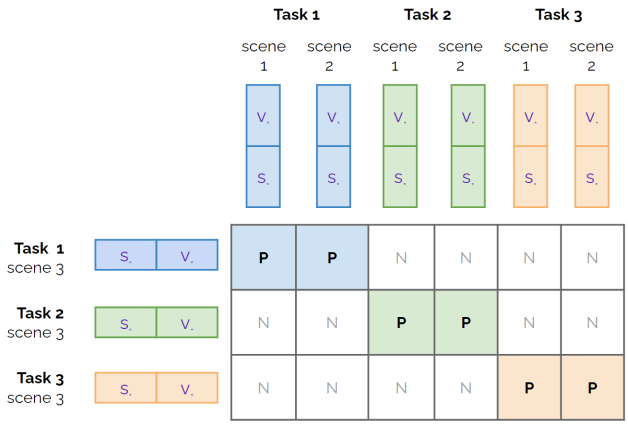

Figure 3. Multi-scene V&L contrastive learning uses multiple sequences. This multi sequence information serves as both positive and negative pairs helping the model to learn the best next scene according to the ground-truth scenes.

Contrastive Selection. We prioritize the visual representation that achieves the best overall consistency throughout the sequence. This selection is achieved by comparing the conditioned scenes with the previous scenes, ensuring that the final output maintains visual coherence. To achieve this, we first represent a scene $sc_n = (s_n, v_n)$ as the concatenation of its text and visual embeddings, and then calculate the dot product between each conditioned scene and the previous scenes, Figure 3. This allows us to measure the similarity between the conditioned scene in step n, denoted as sc_n, and all preceding scenes $sc_{<n}$. We define this similarity as

$$\sum_{k=1}^{n-1} sc_n \cdot sc_k. \quad (7)$$

Next, we apply the softmax function to these similarity scores to convert them into probabilities, making it easier to compare how each conditioned scene is related to the previous scenes. Finally, we select the conditioned scene sc_n^i with the highest probability of generating the video for the next step. This corresponds to computing the

$$\arg\max_{v_n^i} \sigma_{\text{CoSeD}}(sc_n^i, sc_{<n}), \quad (8)$$

where σ_{CoSeD} is the softmax function applied over all conditioned scenes, and v_n^i is the image associated with the selected conditioned scene $sc_n^i = (s_n, v_n^i)$.

By following this process, we ensure that each step in the video is generated based on the conditioned scene that has the strongest visual and contextual relationship to the previous steps, optimizing the flow of the overall sequence.

Contrastive Training. During the contrastive selection training phase, we fine-tune a 600K-parameter model to learn

the relationships between sequential steps across multiple tasks simultaneously. The model processes a set of N steps (both descriptions and images) from a pool of M tasks. For each task, the model is given a step to be processed in the 'next scenes', while all preceding steps of that task, located in the 'past scenes', serve as context. This setup allows the model to effectively leverage sequential dependencies and learn how each future step relates to its corresponding past steps within the same task, thus improving coherence (see Figure 3). Formally, we adopted the cross-entropy loss function,

$$\arg\min_{w_*} \sum_{t}^{M} \sum_{k=1}^{N} l_{t,k} \log \sigma_{CoSeD}(v_n^j, s_n) \tag{9}$$

to guide the learning process, by comparing the model's predictions $\sigma_{CoSeD}(\cdot)$ with one-hot encoded ground truth labels $l_{t,k}$ for each task t and step k. These ground truth labels indicate whether a specific step belonged to a task represented in the context. More details can be found in the appendix file.

5. Experimental Setting

In this section, we describe the experimental setup used to evaluate the performance of CoSeD in generating multi-scene video and image sequences for manual tasks. We provide details on the dataset used, the backbone models, and the baselines chosen for comparison. The aim is to demonstrate CoSeD's ability to generalize across different models and generate coherent task-oriented outputs.

Dataset. We used a dataset [5] consisting of publicly available manual tasks in recipes and DIY domains. Each manual task has a title, a description, a list of ingredients, resources, and tools, and a sequence of step-by-step instructions, which may or may not be illustrated. Details about the dataset can be found in the appendix file.

Video Diffusion Backbone Models. Since CoSeD is independent of the video generation method, we experimented with both Stable Video Diffusion [3] and Lumiere [2] models for multi-scene video generation. Stable Video Diffusion was selected for its public availability, while Lumiere was chosen for its enhanced capability to represent complex motion effectively.

Baselines. To evaluate the effectiveness of CoSeD in generating coherent image and video sequences for real-world manual tasks, we compared its performance against existing approaches: TALC [1] with ModelScope [27] and with Lumiere [2], SD 2.1 [23] with Stable Video Diffusion [3], stand-alone Lumiere [2], and for image sequences we tested Gill [14] and Seed-LLama [10]. During the evaluation, we prompted all models to generate a complete task.

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	<u>81.7</u>	<u>72.9</u>

Table 1. Manual evaluation of multi-scene video generation models based on two key criteria: **Semantic Consistency**, which measures the alignment of generated content with the described task steps, and **Sequence Consistency**, which assesses the visual coherence, text alignment, and overall quality of the video.

6. Results and Discussion

This section presents the evaluation results and discussion of our model's performance, with both human and automatic evaluations. We first examine the results of the human evaluation, followed by the automatic evaluation metrics. Finally, we discuss qualitative results and present ablation studies.

6.1. Human evaluation

The human evaluation was conducted with a focus on two primary criteria. Annotators were asked to assess **Semantic Consistency**, which measures how well the video matches the instructional text, and **Sequence Consistency**, which involves rating the video on text alignment, visual consistency and video quality. See the appendix file for details.

Multi-Scene Consistency Assessment. The results in Table 1 show that CoSeD combined with Lumiere achieves the highest Sequence Consistency at 74.2% and leads in Semantic Consistency with 85.0%, highlighting its effectiveness. This shows that the combination of CoSeD and Lumiere is particularly effective for multi-scene generation tasks.

In contrast, methods such as TALC + ModelScope and TALC + Lumiere show significantly lower Semantic Consistency scores (38.3% and 30.0%, respectively) and a Sequence Consistency of 50.8%. Although SD + SVD and Lumiere alone perform better, they still do not match the performance of CoSeD + Lumiere, underscoring the advantages of our approach in achieving better coherence and text adherence in the generated videos.

Videos Length. We also report the length of the videos generated in Table 1. CoSeD-based methods (CoSeD + SVD and CoSeD + Lumiere) achieve an average video length of around 15 and 21 seconds, respectively. This is substantially

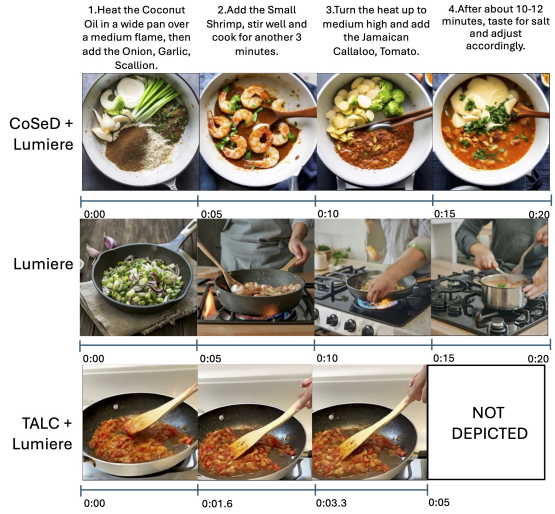

Figure 4. Example of an illustration for the recipe domain.

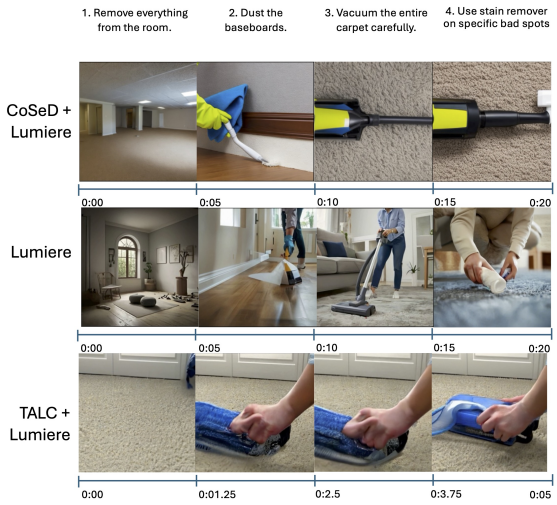

Figure 5. Example of an illustration for the DIY domain.

Figure 6. Example of an illustration for the recipe domain.

Figure 7. Example of an illustration for the DIY domain.

longer than TALC-based methods (TALC + ModelScope and TALC + Lumiere) which generate shorter videos, around 7 and 5 seconds on average.

A visual inspection of the generated videos (Figure 4) clearly indicates that CoSeD successfully depicts all steps in the task, whereas TALC, despite being a multi-scene model, cannot achieve it. Even when TALC successfully depicts all steps, each scene typically lasts no more than 1.5 seconds, considering the average task length of 4.9 scenes (as detailed in the appendix file). In contrast, our model consistently achieves at least 3 seconds per scene, effectively providing double the duration for each step.

Side-by-Side Evaluation. To assess models prioritizing coherence against our best model, we conduct a side-by-side evaluation. Annotators choose which videos better represent task steps, directly comparing each model's coherence across scenes. See the appendix file for details. This evaluation focuses on our best model, CoSeD + Lumiere, against coherence-focused models such as TALC + ModelScope, TALC + Lumiere, CoSeD + SVD, and the second-best model from Table 1, Lumiere.

According to the side-by-side evaluation results, Figure 8, CoSeD + Lumiere consistently outperforms all other models, with annotators repeatedly selecting it over competitors. For example, CoSeD + Lumiere achieves a selection rate 87% compared to just 13% for TALC + ModelScope, demonstrating its superior ability to maintain coherence. It is also chosen 68% of the time over TALC + Lumiere, which has a selection rate of 32%, reflecting its better task consistency.

Compared to other CoSeD variations, CoSeD + Lumiere maintains its advantage. It is selected 61% of the time over CoSeD + SVD and outperforms Lumiere with a selection rate of 65% versus 35%. These results highlight its exceptional

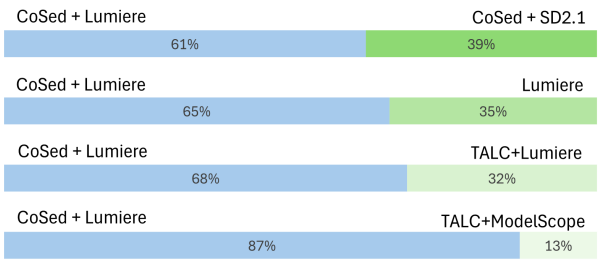

Figure 8. Annotators choice of coherent video generation models in a side-by-side comparison.

coherence in multi-scene tasks.

This evaluation clearly demonstrates that annotators consistently prefer CoSeD + Lumiere over other models, highlighting its superior ability to maintain coherence across scenes and establishing it as the most effective model for managing multi-scene tasks.

CoSeD vs Groundtruth. To evaluate the absolute quality of the generated video sequences, human annotators rated each sequence on a scale of 1 to 5, comparing them to ground-truth sequences. Deviations like hallucinated visual artifacts or inconsistent actions affect perceived quality. As shown in Table 2, our method achieves more than 64% of the ground truth score, with ground truth sequences scoring just 0.5 points below the maximum.

Method	Average Rating
CoSeD +Lumiere	2.9 ± 0.99
Ground-truth	4.5 ± 0.55

Table 2. Human annotation for the comparison of the proposed method with ground-truth scenes.

6.2. Automatic evaluation

We employ CLIP [21] to evaluate the sequence similarity of each task ($V \mapsto V$) and to assess the adherence of the generated image to the given textual descriptions ($T \mapsto V$). This provides a comprehensive evaluation of the alignment of the generated images with the intended textual descriptions and their visual coherence throughout the sequence.

CoSeD Performance. CoSeD achieves a sequence similarity score ($V \mapsto V$) of 84.8 and a description adherence score ($T \mapsto V$) of approximately 27.1 (see Table 3), outperforming the image and video baselines in textual adherence. These results highlight CoSeD's ability to generate sequences that are both visually coherent and semantically aligned with the text, demonstrating its significant potential for practical applications.

	Method	$V \mapsto V$	$T \mapsto V$
Image	CoSeD	84.8	**27.1**
	Seed-Llama	88.0	16.5
	GILL	**88.2**	22.3
Video	CoSeD + SVD	84.8	**27.1**
	CoSeD + Lumiere	84.8	**27.1**
	TALC + ModelScope	81.8	12.4
	TALC + Lumiere	**90.7**	15.3
	SD + SVD	82.1	26.4
	Lumiere	83.4	14.9

Table 3. Automatic evaluation in terms of CLIP visual similarity ($V \mapsto V$) and CLIP semantic similarity ($T \mapsto V$).

Comparison with Baseline Models. In a comparison to other sequence-generating models (see Table 3), CoSeD with both Video Stable Diffusion and Lumiere consistently achieves the highest description adherence ($T \mapsto V$) without discarding sequence similarity ($V \mapsto V$).

When comparing CoSeD with the best model in sequence similarity, our model lags only 5.9% while it gains 11.8% in description adherence. Although TALC + Lumiere excels in maintaining high sequence similarity, CoSeD demonstrates superior adherence to descriptions without compromising sequence similarity. This strong description adherence score highlights the effectiveness of our model in aligning generated content with text descriptions, which is crucial for tasks such as accurately converting textual descriptions into videos. Compared to vanilla video-only models, our approach surpasses both metrics, leading to improved results.

6.3. Qualitative Analysis

Figure 4 and 6 provide a closer look at how different methods influence the quality of generated videos and image sequences for a certain recipe. Our model excels in maintaining a consistent background, keeping the same pan, and ensuring that the ingredients evolve seamlessly from raw to final recipe. This clear depiction of the sequence provides viewers with a visually stable and easy-to-follow video.

For the out-of-scope DIY tasks shown in Figures 5 and 7, our model effectively handles broader actions, such as showing a room without furniture or a cleaning process, with good text adherence. However, it struggles with depicting complex tools such as vacuum cleaners. Other methods often generate very similar images, failing to accurately represent the task or omitting steps entirely, as demonstrated by TALC.

6.4. Ablation Studies

This section analysis CoSeD contrastive selection of latents and steps along with its role in ensuring alignment with (possibly non-linear) sequences of instructions.

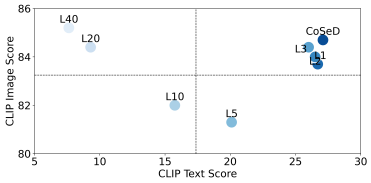

Figure 9. Impact of denoising latents in the performance of CoSeD's visual and semantic similarity.

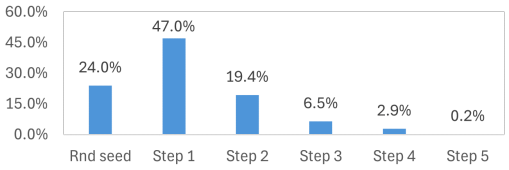

Figure 10. The average number of times that CoSeD selected each task step for generating the next scene.

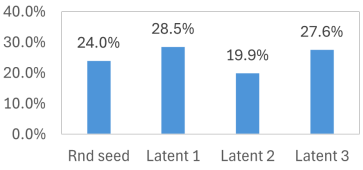

Figure 11. The average number of times that CoSeD selected a given latent to generate the next scene.

Denoising Latents. Understanding how latent variables affect sequence coherence and adherence is key to refining our model. Figure 9 shows a correlation between latent settings and the model's ability to produce coherent, textually aligned sequences. The analysis in Figure 9 identifies four key latent configuration areas. The ideal zone balances high sequence similarity (CLIP Image Score) and description adherence (CLIP Text Score). Latent 5 shows moderate text adherence, but lacks visual coherence. Latents 20 and 40 produce coherent images but deviate from task steps, leading to weak text adherence. Latent 10 performs poorly overall.

Figure 12. Impact of later latents on sequence generation coherence and text adherence.

Figure 12 provides examples from these key latent areas: Latent 5 generates images that adhere to the prompt but lack coherence, Latent 10 starts with good coherence and text alignment but loses coherence over time, and Latent 20 produces overly similar images with low text adherence.

Ultimately, CoSeD (top right mark in Figure 9) achieves superior performance by leveraging early denoising latents, using contrastive selection to enhance results compared to using individual latents.

Non-linear Video Scene Generation. A key feature of CoSeD is its ability to evaluate denoising iterations across previous steps. As shown in Figure 10, CoSeD exploits this non-linearity by selecting latents from various steps, rather than focusing solely on the immediately preceding one.

Similarly, Figure 11 indicates that the model does not always select the same latent. This variability suggests that different latents contribute with different information to the final generation, which is why CoSeD chooses the most suitable latent rather than opting for the most recent one.

7. Conclusion

Generating multi-scene instructional videos for complex tasks, such as DIY projects and recipes, all while maintaining sequence coherence and accurate scene representation, is not a trivial feat. The proposed method addresses these challenges with key contributions. First, we employ a decoder model to generate visual prompts from the sequence of instructions to ground the generation in a common context, thus ensuring better scene accuracy. Second, by conditioning the diffusion process on images from previous scenes, the method maintains coherence across scenes. Third, and more importantly, the CoSeD's contrastive selection of the most consistent image enables the assessment of all previous steps. The result is the selection of the image that is most related to the overall sequence rather than just the preceding step. Additionally, the contrastive nature of CoSeD enables the generation of non-linear sequences of video scenes, an exclusive feature of CoSeD.

In addition, the compact design of the model helps to achieve efficient training and easy domain-specific fine-tuning, while its flexibility supports ensembles of diffusion models for optimal performance.

Evaluations confirm that our method effectively maintains scene coherence and accurately represents textual descriptions, as demonstrated by both manual and automatic evaluations. Importantly, side-by-side human annotation shows that annotators prefer our model over 65% of the time, highlighting the effectiveness of our sequence-grounded approach. This preference underscores the value of our method in producing coherent and high-quality instructional videos.

Acknowledgements

We thank anonymous reviewers for their valuable comments and suggestions. This work was partially supported by a Google Research Gift and by the FCT project NOVA LINCS Ref. (UIDB/04516/2020).

References

[1] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation, 2024. 2, 5

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 2, 5

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. 1, 2, 5

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22563–22575. IEEE, 2023. 1, 2

[5] João Bordalo, Vasco Ramos, Rodrigo Valerio, Diogo Glória-Silva, Yonatan Bitton, Michal Yarom, Idan Szpektor, and João Magalhães. Generating coherent sequences of visual illustrations for real-world manual tasks. *CoRR*, abs/2405.10122, 2024. 3, 4, 5

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[7] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 2

[8] Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. Aligning actions across recipe graphs. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6930–6942, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 2

[9] Zhangyin Feng, Yuchen Ren, Xinmiao Yu, Xiaocheng Feng, Duyu Tang, Shuming Shi, and Bing Qin. Improved visual story generation with adaptive context modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4939–4955, Toronto, Canada, July 2023. Association for Computational Linguistics. 1

[10] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama SEE and draw with SEED tokenizer. *CoRR*, abs/2310.01218, 2023. 2, 5

[11] Patrizia Grifoni. *Multimodal human computer interaction and pervasive services*. IGI Global, 2009. 1

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *CoRR*, abs/2210.02303, 2022. 1

[13] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: equip diffusion models with LLM for enhanced semantic alignment. *CoRR*, abs/2403.05135, 2024. 3

[14] Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. Generating images with multimodal language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 2, 5

[15] Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. A recipe for creating multimodal aligned datasets for sequential tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4871–4884, Online, July 2020. Association for Computational Linguistics. 1

[16] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *CoRR*, abs/2309.15091, 2023. 1, 2

[17] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with LLM. *CoRR*, abs/2401.01256, 2024. 1, 2

[18] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, page 70–87, Berlin, Heidelberg, 2022. Springer-Verlag. 1

[19] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2

[20] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. *CoRR*, abs/2211.10950, 2022. 1, 2, 4

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4, 7

[22] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2493–2502. IEEE, 2023. 1, 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 3, 5

[24] Frank Serafini. *Reading the visual: An introduction to teaching multimodal literacy*. Teachers College Press, 2014. 1

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 2

[27] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 5

[28] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LAVIE: high-quality video generation with cascaded latent diffusion models. *CoRR*, abs/2309.15103, 2023. 1

[29] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: diffusion over diffusion for extremely long video generation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1309–1320. Association for Computational Linguistics, 2023. 2

[30] Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework, 2024. 1, 2

[31] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation, 2024. 2